X CONGRESO ISKO CAPÍTULO ESPAÑOL (Ferrol, 30 junio - 1 julio 2011)

NORMALIZACIÓN DEL VOCABULARIO DE INDIZACIÓN EN BASES DE DATOS MULTIDISCIPLINARES DE CIENCIAS SOCIALES Y HUMANAS: ANÁLISIS COMPARATIVO DE LA EXPERIENCIA DE ISOC (ESPAÑA) Y CLASE (MÉXICO) DESDE LA DÉCADA DE 1970 HASTA 2010

Luis Rodríguez Yunta
CSIC, Centro de Ciencias Humanas y Sociales, Madrid, España, luis.ryunta@cchs.csic.es
José Octavio Alonso-Gamboa
UNAM, Dirección General de Bibliotecas, Ciudad de México, oalonso@unam.mx

RESUMEN:
Análisis comparativo entre las bases de datos ISOC, producida por el Consejo Superior de Investigaciones Científicas, y CLASE, elaborada en la Universidad Nacional Autónoma de México, las dos especializadas en revistas de Ciencias Sociales y Humanas. Se resaltan las similitudes y diferencias entre ambos productos, con especial atención a los criterios utilizados en el análisis documental de contenido y la normalización de los lenguajes de indización. Se concluye que entre ambos productos acumulan una amplia experiencia y que sus vocabularios de indización pueden servir de modelo en estas disciplinas, por lo que se sugiere cooperar en la construcción de tablas de equivalencias entre los términos acumulados en sus índices y tesauros.

PALABRAS CLAVES:
Bases de datos documentales, normalización, lenguajes de indización, control del vocabulario, Ciencias Sociales y Humanas, España, México, América Latina

TITLE:
Standardization of indexing vocabulary for multidisciplinary databases specialised in Social Sciences and Humanities: Comparative analysis from the experience of ISOC (Spain) and CLASE (Mexico) from 1970's to 2010

ABSTRACT:
Comparative analysis was carried out between two databases specialising in the Social Sciences and Humanities: ISOC produced by Consejo Superior de Investigaciones Científicas (Spain) and CLASE produced by Universidad Nacional Autonoma de Mexico. Similarities and differences between the two databases are remarked upon with emphasis on criteria for the standardization of content analysis and indexing languages. Conclusions point out that both products have accumulated a wealth of experience and can serve as models for databases in these disciplines. It is suggested that collaborative work can be done in the development of equivalents between the terms accumulated in their respective indexes and thesauri.

KEYWORDS:
Bibliographic databases, Standardization, Indexing languages, Controlled vocabulary, Social Sciences, Humanities, Spain, Mexico, Latin America

1. INTRODUCCIÓN: OBJETIVOS Y METODOLOGÍA

Las bases de datos ISOC (Consejo Superior de Investigaciones Científicas) y CLASE (Universidad Nacional Autónoma de México) son dos productos ya tradicionales de documentación científica Aunque surgieron en la etapa inicial de desarrollo del mercado de la información electrónica (Rodríguez Yunta, 2009), se han ido adaptando a las nuevas demandas de la sociedad y actualmente siguen siendo recursos bibliográficos importantes tanto por su cobertura documental como por su nivel de puntos de acceso para la recuperación de información. Ambos productos responden al modelo de bases de datos bibliográficas referenciales con análisis documental de contenido. Comparten además la condición de ser sistemas de carácter multidisciplinar, que abarcan diferentes disciplinas de Ciencias Humanas y Sociales, por lo que han tenido que enfrentarse a problemas muy similares en el tratamiento del lenguaje científico y las necesidades de recuperación de información en estas áreas del conocimiento. Aunque provenientes de diversos orígenes geográficos, las revistas cubiertas en ISOC y CLASE presentan también similitudes que se reflejan en el proceso de indización por materias que ambos sistemas realizan. Puede decirse que, junto a los productos paralelos para ciencia y tecnología ICYT e IME en España y PERIÓDICA en México, son productos representativos del desarrollo histórico de la documentación científica en España y América Latina desde la década de 1970.

Por todo ello, resulta relevante realizar un análisis comparativo de sus características básicas, con especial atención a los criterios utilizados en el análisis documental de contenido y la normalización de los lenguajes de indización.

La metodología del análisis tiene en cuenta diferentes aspectos:

- Establecimiento de diferencias y semejanzas en la definición de características y procesos técnicos de ambos productos: los aspectos formales, la evolución de estos productos documentales, con especial hincapié en el establecimiento de criterios para la indización. Como fuentes se utilizan los manuales de normas de uso interno en los que se define la estructura de ambos productos y se establecen criterios para el análisis de contenido y la construcción de términos de indización.
- Comparación entre los vocabularios de indización empleados en ambos productos: políticas de normalización y consecuencias para las estrategias más eficaces en la recuperación de información.

Finalmente, se pretende extraer conclusiones sobre la hipótesis de una posible cooperación en una única interfaz de recuperación entre ambos productos documentales.

2. COMPARACIÓN GENERAL ENTRE LOS PRODUCTOS DOCUMENTALES ISOC Y CLASE: CARACTERÍSTICAS BÁSICAS Y CRITERIOS DE INDIZACIÓN

La base de datos ISOC se originó en el Instituto de Información en Ciencias Humanas y Sociales, centro perteneciente al Consejo Superior de Investigaciones Científicas (CSIC), creado en 1975. Su objetivo principal fue identificar y facilitar las búsquedas bibliográficas de los artículos de revistas científicas españolas en estas disciplinas. Su producción dependió del Centro de Información y Documentación Científica (CINDOC) en el periodo 1991-2007, pasando posteriormente a integrarse dentro del Centro de Ciencias Humanas y Sociales (CCHS), siempre dentro del mismo organismo público, el CSIC. Su modelo se ha basado en el análisis documental de contenido realizado por documentalistas especializados entre los que se realiza un reparto de tareas por disciplinas. El producto es distribuido por el propio organismo mediante dos interfaces de usuario, una de ellas de carácter gratuito (http://bddoc.csic.es:8080) que sólo da acceso a los datos de sumario de las publicaciones, y otra por suscripción (http://bddoc.csic.es:8085) que permite

interrogar y visualizar los registros completos. El equipo de trabajo de la Unidad ISOC está compuesto en 2010 por 20 personas entre titulados superiores, personal de apoyo y contratos externos. El crecimiento anual del producto se sitúa en dicho año en más de 31.000 registros bibliográficos, superando los 619.000 acumulados desde 1975.

La estructura de la base de datos cuenta con 30 campos visibles para los usuarios, de los que nueve son obligatorios. Se incluyen campos específicos para la recuperación por materias, descriptores, identificadores, topónimos, periodo histórico, siglos, legislación y jurisprudencia; así como resumen y otros elementos que facilitan el análisis de la producción científica española, afiliación institucional de los autores, clasificación, idioma, año de publicación,… (ABEJÓN *et al*., 2009).

Por su parte, la base de datos CLASE nació en 1975 en la Universidad Nacional Autónoma de México (UNAM). De 1975 a 1997 la producción de la base de datos fue responsabilidad del Centro de Información Científica y Humanística (CICH) hasta su incorporación en 1997 a la Dirección General de Bibliotecas (DGB). Su objetivo es compilar, analizar y difundir los contenidos publicados en una selección de revistas académicas latinoamericanas especializadas en las diferentes disciplinas de las Ciencias Sociales y Humanas. El análisis documental es realizado por siete especialistas en indización; adicionalmente se cuenta con un programa de servicio social que acepta pasantes de licenciatura para apoyar el análisis de contenido en CLASE. Durante 2010 se añadieron más de 21.000 registros a la base de datos, superando los 317.000 acumulados desde 1975. La estructura del registro cuenta con 21 campos diferentes, 12 de ellos obligatorios y 13 visibles para consulta gratuita en http://clase.unam.mx.

Las disciplinas abarcadas en ambos productos son similares, pues se circunscriben a las Ciencias Sociales y las Humanidades. En la base ISOC, la clasificación permite distribuir varios subproductos de consulta independiente que se ofrecen como opción complementaria de la interrogación del conjunto completo:

- Antropología social y cultural;
- Arqueología y Prehistoria;
- Bellas Artes;
- Biblioteconomía y Documentación;
- Ciencias Jurídicas;
- Ciencias Económicas;
- Ciencias de la Educación;
- Filosofía;
- Geografía, Urbanismo y Arquitectura;
- Historia;
- Lingüística y Literatura;
- Psicología;
- Sociología y Ciencias Políticas;
- Estudios sobre América Latina.

En el caso de CLASE las revistas analizadas deben corresponder a alguna de las siguientes disciplinas: Administración, Antropología, Arte, Bibliotecología, Ciencia de la Información, Ciencias de la Comunicación, Contaduría, Demografía, Derecho, Economía, Educación, Filosofía, Geografía, Historia, Lingüística, Literatura, Política, Psicología, Relaciones Internacionales, Religión y Sociología. También se analizan revistas de carácter multidisciplinario cuyos contenidos sean relevantes para las Ciencias Sociales y Humanas.

El número de revistas analizadas en ISOC es de casi 2.600 títulos editados en España desde 1975, aunque muchos de ellos ya no están vigentes y a partir de 2006 se aplican criterios más restrictivos de selección. Por ello, en la actualidad se trabaja con un conjunto de aproximadamente 1.250 publicaciones. El número de registros acumulados en diciembre de 2010 supera los 619.000 documentos, de los cuales 301.000 cuentan con datos de afiliación institucional, 199.000 con

resumen de autor y 91.000 con enlace al texto completo. La mayor parte son artículos de revista y apenas un 5% se corresponden con otras modalidades documentales, comunicaciones, artículos de monografía o documentos de trabajo. En CLASE se han analizado históricamente poco más de 1.700 títulos provenientes de 20 países de América Latina y el Caribe (REYNA, 2010) y hasta 2010 se habían ingresado más de 317.000 registros, 12.000 resúmenes de autor y 35.000 enlaces a artículos a texto completo. Las revistas son sometidas a un proceso de selección por parte de un comité interno.

En ISOC se ha desarrollado un sistema propio de clasificación que además de permitir la división del producto en las diferentes sub-bases enumeradas anteriormente, cuenta con otros dos niveles de profundidad en cada disciplina, con objeto de definir el tema central tratado en el documento. El sistema de indización utiliza la diferenciación de campos entre descriptores, identificadores, topónimos, legislación, jurisprudencia, periodo histórico, décadas y siglos. Respecto a los criterios utilizados en la cumplimentación de los registros, la Unidad de Producción de las bases de datos ISOC cuenta con un Manual de normas (CINDOC, 2000), que actualmente se encuentra en proceso de actualización. Se trata de un documento de uso interno que recoge las decisiones tomadas por el equipo de trabajo con el objetivo de buscar la homogeneidad en la metodología y el tratamiento documental que aplican las diversas áreas temáticas. Se especifican diferentes casos y ejemplos que deben seguirse en la grabación de datos o cumplimentación de los formatos diseñados para la alimentación de la Base de Datos ISOC. Respecto al control de vocabulario se apuesta por la utilización de lenguajes controlados:

> Las bases de datos ISOC se elaboran mediante la indización con vocabulario controlado, por lo que todas las áreas mantienen sus respectivos léxicos o tesauros de acuerdo con sus necesidades específicas. En el momento de la indización pueden utilizarse en este campo tanto el vocabulario controlado del área como cualquier otro término admitido en los tesauros o léxicos de las restantes materias. Cuando los tesauros estén acabados será un campo que se podrá validar automáticamente y utilizar en las búsquedas en la base de datos.
> Deben utilizarse siempre descriptores suficientemente específicos. El número de descriptores asignados a un registro y su nivel de especificidad han de estar en relación con la profundidad del contenido del documento analizado. El descriptor más específico que refleje el tema central ha de ser seleccionado siempre, pudiendo añadirse aquellos términos de indización más genéricos que se juzguen convenientes. (CINDOC, 2000: 52-53).

En el proceso de trabajo aplicado por la Unidad ISOC, los registros pueden ser grabados directamente por los documentalistas con todos los campos necesarios, o bien de forma gradual, ser introducidos en una fase previa por personal de apoyo, sólo con los datos catalográficos, afiliación institucional, resumen y clasificación genérica. De este modo, los documentalistas pueden añadir posteriormente una clasificación más específica y los términos de indización adecuados. Con ello, se mejora el control del proceso de trabajo y la actualización de los datos de sumarios que pueden consultarse en la interfaz de acceso gratuito, aunque el inconveniente es la inconsistencia de una parte de los registros que no cuentan con descriptores.

La indización de los documentos en CLASE se rige también por un manual de indización, (ALONSO *et. al*, 2011), que contiene las reglas generales aplicables a todas las disciplinas que cubren las revistas analizadas en esta base de datos. Hasta mediados de 2010 se contemplaron en CLASE tres niveles de indización para el análisis de los contenidos:

1. *Disciplina,* que representa el nivel superior jerárquico compuesto por las grandes disciplinas del conocimiento;
2. *Subdisciplina*, que resulta ser una clasificación derivada de la disciplina mayor y que por tratarse de una lista preestablecida facilita el control de los términos a seleccionar y
3. *Palabras clave*, que representan el nivel más específico aplicable para describir los contenidos de un documento.

A partir de la segunda mitad de 2010 las subdisciplinas se incorporaron a las palabras clave, con carácter de uso obligatorio, quedando formalmente sólo dos niveles: disciplina y palabras clave. Las palabras clave en CLASE se asignan consultando tanto las reglas contempladas en el manual de uso interno, así como los descriptores contenidos en diferentes tesauros y glosarios construidos en lengua castellana para otros servicios de información. En esencia se ha impulsado en CLASE el uso del lenguaje libre para mitigar los constreñimientos de un vocabulario totalmente controlado y contender con la constante actualización terminológica de las Ciencias Sociales y Humanas. En CLASE los registros son grabados completos y se hacen disponibles para consulta en tiempo real.

Tanto CLASE como ISOC cuentan con varios campos con listas asociadas o criterios de normalización, a partir de las cuales los analistas deben elegir el término que asignarán al documento en el análisis. Las listas de algunos campos pueden ser enriquecidas con nuevas ocurrencias, otras no. Los campos con listas se describen sucintamente a continuación:

- Título de la revista: Corresponde a los nombres de las revistas analizadas. El listado refleja cambios de títulos y para un mejor control de los datos los analistas no eligen de la lista un título propiamente dicho sino una clave numérica que el sistema traduce en el título completo normalizado y su respectivo ISSN.
- Autores: Representan a los autores personales de cada documento. La forma de asentarlos es entrando por los apellidos. Los manuales de uso interno contemplan diferentes reglas según el origen nacional de los nombres. El listado se revisa para aplicar normalización a los nombres ya que tanto las revistas como los propios autores no siempre los asientan de la misma manera.
- Disciplinas / Clasificación: CLASE cuenta con 29 grandes áreas del conocimiento que son las mismas que se usan para clasificar a las revistas. Este listado permite el acceso por materia o categoría temática y no debe ser modificado por los analistas siendo obligatorio seleccionar de una hasta tres disciplinas para cada documento. En ISOC se aplica una tabla codificada de clasificación. Se utilizan códigos numéricos de seis dígitos, que permiten agrupar los documentos en 17 clases principales y establecer tres niveles de profundidad.
- Términos de indización. En CLASE se utilizan palabras clave, que representan el contenido más específico del documento y están subdivididas en dos grupos: 1) las subdisciplinas y 2) las palabras clave propiamente dichas. Las subdisciplinas son una subclasificación de las disciplinas; se trata de un listado compuesto por 364 términos que al igual que las disciplinas no debe ser modificado durante el trabajo de análisis. Las palabras clave por su parte se utilizan para representar los términos más específicos; además de las palabras clave temáticas, se incluyen nombres de personas, nombres geográficos, nombres de instituciones u organismos, siglas y acrónimos, así como períodos de tiempo. El listado puede ser enriquecido con nuevos términos, pero el sistema alerta cuando un nuevo término va a ser registrado en la lista dando oportunidad al analista de reflexionar sobre su integración o no a la lista controlada. En ISOC se emplean varios campos para el análisis de contenido y diferentes herramientas de normalización. En el campo de descriptores se cuenta con tesauros o léxicos de indización, especializados por disciplinas. La lista depurada de valores acumulados en este campo se aplica como tabla de validación que facilita el control de la corrección en la grabación de datos. También en topónimos se ha desarrollado un tesauro para las entidades político-administrativas y listados de apoyo para otros lugares geográficos. Los campos de identificadores, legislación y jurisprudencia se regulan a través de normas establecidas en el manual de uso interno.
- Instituciones de adscripción: Se trata de los lugares de trabajo de cada autor. Esta información se consigna en CLASE desde 1975 por lo que resulta ser muy rica y variada con más de 3,000 instituciones diferentes ubicadas en más de 100 países. En los primeros años solamente se indizaba la institución del primer autor y a partir de 1986 todas las adscripciones diferentes. Los nombres de las instituciones aparecen completos

y según el país de origen se escriben en español, portugués, inglés o francés; las instituciones ubicadas en países con idioma diferente a los antes listados, se representan en inglés. Los elementos codificables son cuatro: la institución de primer nivel, el nivel inmediato inferior, la ciudad y el país. Desde 2009 la relación autor-lugar de trabajo se establece mediante un número, como suele presentarse en los artículos. Por su parte en ISOC la afiliación institucional de los autores se incorpora desde 1987, aplicándose desde un principio a todos los autores de cada documento. Su construcción no está codificada, pero si se especifica en el manual de normas con una estructura muy similar de niveles y el empleo de abreviaturas.

- Idioma del documento y de los resúmenes: Como su nombre indica, representa el o los idiomas en que está escrito un documento y sus resúmenes. Se trata de dos listados diferentes con un número limitado de ocurrencias, que puede ser enriquecido si un documento o sus resúmenes aparecen en algún idioma diferente a los preestablecidos en ambos listados.
- Tipo de documento: En CLASE se utiliza una lista controlada que incluye 26 clasificaciones diferentes conforme el documento que se analiza. En ISOC se divide en dos niveles, uno obligatorio con 6 valores y otro opcional, denominado modo de documento, con 32 variantes. En ambos casos, se trata de listados que no pueden ser enriquecidos; el analista debe elegir uno de los tipos listados y asignarlo al documento.
- Enfoque. Este campo aparece solo en CLASE, cuenta con un listado que representa el tratamiento (analítico, descriptivo, teórico,…) que el autor da a su documento y el analista debe elegir solamente de entre 13 opciones listadas.

En el proceso de análisis de contenido, puede decirse que en ambos productos se aplica la filosofía de la *indización por asignación*. Este método implica que en el registro se incorporan los términos más adecuados existentes en listas preestablecidos que representan las materias tratadas en el documento, aunque no necesariamente se corresponden con la forma elegida por el autor en el propio texto (LANCASTER, 1996: 14). Esta metodología es coherente con la filosofía de la *indización centrada en el usuario*, que también se refleja en algunos manuales teóricos (GIL, 2008: 62). Aunque no es posible asegurar la consistencia de este criterio en todos los registros, puede decirse que preferentemente se ha dado prioridad a reflejar aquellos conceptos que puedan coincidir con necesidades de información y posibles búsquedas de los usuarios.

En cuanto a la sintaxis de los términos de indización, tanto en CLASE como en ISOC se ha utilizado un lenguaje poscoordinado. Se aplica un grado limitado de precoordinación, siguiendo el principio de fraccionamiento de las entradas a su forma más simple que tenga sentido pleno sin ambigüedad. Con ello, difieren claramente del modelo de lenguaje precoordinado de los encabezamientos de materia utilizados en los catálogos generales de las bibliotecas de la UNAM y el CSIC. En ISOC existen tesauros o léxicos por disciplinas que reflejan cuando se emplea un unitérmino y cuando es necesario utilizar un grupo nominal para evitar la ambigüedad o mejorar la precisión. En CLASE, el uso del lenguaje libre determina en gran medida la construcción de los términos. Por ejemplo, cada vez es más frecuente encontrar términos compuestos en la literatura científica los cuales son ingresados en su forma directa a la lista de palabras clave; términos como “evaluación educativa”, “lucha de clases” o “mercado de trabajo”, si se coordinaran al momento de la recuperación perderían sentido para el usuario. No obstante, cuando un concepto puede representarse mediante varios términos por separado, sin pérdida de sentido, así se hace.

Para la construcción de los términos se prefieren los sustantivos o frases sustantivadas. Se recomienda respetar el orden natural de las palabras y estar atentos a la terminología más utilizada en las diferentes disciplinas. Se valora el uso de terminología científica y se tiene cuidado con la terminología más coloquial aunque ésta última puede ser incluida si su uso es generalizado y si el analista así lo considera. El manual de la base CLASE recomienda no usar adjetivos como palabras clave, salvo que el término así sea utilizado, como en el caso de “muy alta frecuencia”. También se desalienta el uso de los infinitivos y participios de los verbos como términos de indización.

También se recomienda no abusar del uso de gentilicios, para evitar el crecimiento desmedido de la lista de palabras clave y dar un mejor uso a los nombres geográficos.

En CLASE, un aspecto que ha impactado a la normalización, principalmente de las listas de palabras clave, autores e instituciones de adscripción, es la propia pervivencia de la base de datos. Con 35 años de existencia, CLASE ha experimentado diversas metodologías en la forma de organizar el acceso a sus contenidos lo que aún se refleja en los listados. De manera particular vemos esta situación en las palabras clave donde aún son muchos los términos con una sola ocurrencia que hacen que la lista sea muy grande y dispersa. En cuanto a la presentación de los términos, durante sus primeros 22 años (de 1975 a1997) todos los términos se ingresaban en mayúsculas y sin diacríticos; en 1998, cuando CLASE contaba ya con cerca de 150,000 registros almacenados y más de 80,000 palabras clave diferentes, se migró a otro sistema de ingreso de datos que permitía representar las palabras con mayúsculas, minúsculas y diacríticos, por lo que se decidió hacer una reconversión general. Esta tarea supuso muchas horas-hombre de trabajo durante los primeros años posteriores a 1998 y el trabajo se centró en los términos con mayores ocurrencias los cuales, a la fecha, están aceptablemente normalizados; sin embargo, permanece en la lista una amplia variedad de términos con ocurrencias pequeñas que aún faltan por revisar y normalizar.

En la Unidad de las bases de datos ISOC se ha apoyado el uso y construcción de lenguajes controlados, se han elaborado algunos tesauros disciplinares y un tesauro de topónimos. Sin embargo, al no haberse completado la elaboración de herramientas para todas las áreas, no ha sido posible explotar todas las ventajas de los lenguajes controlados en la recuperación de información. Se cuenta con tesauros publicados tan sólo en algunas disciplinas: Economía, Urbanismo, Psicología, Historia contemporánea, Derecho, Biblioteconomía y Documentación. En otras áreas se cuenta con tesauros no publicados y en Ciencias de la Educación se ha trabajado tradicionalmente de acuerdo con el léxico del Tesauro Europeo de Educación.

A consecuencia de trabajar con léxicos específicos por disciplinas, se han producido diferencias de criterio en la construcción del vocabulario de indización. Para reducir las inconsistencias se han formado en diferentes momentos equipos de trabajo y proyectos específicos dirigidos a homogeneizar los descriptores presentes en la base de datos, pero es un proceso lento y costoso que aún no ha finalizado. Al igual que en la experiencia mexicana, se aprovechó la conversión a minúsculas para normalizar entradas. La tendencia a la dispersión de variantes de un mismo concepto y a la incorporación continua de nuevos descriptores y erratas de grabación, se frenó considerablemente al disponer que este campo contara con una tabla de validación de entradas preestablecidas. Actualmente existe un campo de nuevos términos, no visible para el usuario pero cuyos términos si son recuperables en la búsqueda básica. De este modo es posible incluir descriptores candidatos en los registros, pero no se incorporan al campo de descriptores hasta que no sean añadidos en la tabla de validación. Este sistema de control sólo puede aplicarse en este campo, no en identificadores o topónimos, en los que la homogenización depende de visualización de listados y procesos manuales de corrección.

Por su parte CLASE no cuenta con tesauros propios. La naturaleza multidisciplinaria de la base de datos dificulta la selección de un solo tesauro como herramienta de trabajo pero se alienta a los analistas a consultar aquellos que consideren más adecuados según el área temática de análisis. La última versión del manual correspondiente a 2011 lista una serie de tesauros y glosarios como recursos de apoyo a la indización, entre ellos todos los que se usan en ISOC. Sin embargo, su uso estará siempre matizado por las diferencias terminológicas propias de cada país. Cuando existe un tesauro mexicano o latinoamericano se termina prefiriendo su uso en relación con uno español. CLASE no posee un tesauro propio debido a que la construcción de estas herramientas resulta costosa en tiempo, dinero y esfuerzo y requiere además de un grupo de trabajo con prácticamente un especialista por cada disciplina analizada y precisa de una actualización periódica.

3. COMPARACIÓN ENTRE LOS LÉXICOS DE INDIZACIÓN UTILIZADOS EN ISOC Y CLASE

3.1. CRITERIOS GENERALES UTILIZADOS EN LOS PROCESOS DE NORMALIZACIÓN DE LOS VOCABULARIOS DE INDIZACIÓN

A pesar de no contar con un tesauro multidisciplinar como herramienta única de normalización de su vocabulario de indización, tanto ISOC como CLASE aplican criterios básicos para mejorar la consistencia y homogeneizar en lo posible las entradas de materias. Puede decirse que cuentan con *vocabulario controlado*, "es decir, un conjunto limitado de términos que deben utilizarse para representar las materias de los documentos" (LANCASTER, 2002:19). De las posibles modalidades de este tipo de vocabulario según Lancaster, que puede ser una lista de encabezamientos de materias, un esquema de clasificación, un tesauro o simplemente una lista autorizada de frases o palabras clave, ambos productos se encuadran en este último caso.

- *Uso de términos genéricos / específicos*. En ISOC se emplean habitualmente ambos niveles dentro del campo de *descriptores*, aunque determinados conceptos genéricos se reflejan preferentemente a través de la *clasificación*. En CLASE se utilizan ambos en los campos de *disciplina* y *palabras clave*. Las clasificaciones de disciplinas y subdisciplinas corresponden más bien a términos genéricos. Las disciplinas de CLASE guardan equivalencia con las grandes áreas temáticas de ISOC y bien podrían servir para parcelar la base de datos en diferentes subproductos especializados por temas. La mayor especificidad posible se aprecia en las palabras clave, las cuales deben ser indizadas siempre de acuerdo con las reglas generales documentadas en el manual.

- *Tipología de los términos de indización*: A diferencia de ISOC, en CLASE no se hacen distinciones de palabras clave según su tipología. En un mismo campo el usuario puede recuperar términos temáticos, identificadores, onomásticos o geográficos. Esta decisión se adoptó pensando que para el usuario es más fácil recuperar todos en un solo campo. Por el contrario, en ISOC se optó por la diferenciación en campos específicos para los conceptos (*descriptores*), nombres propios de personas e instituciones (*identificadores*) y nombres propios de lugares geográficos (*topónimos*), a fin de facilitar el control del vocabulario. Así, es teóricamente posible la aplicación de tesauros en los campos de *descriptores* y *topónimos*, sin la intromisión de listas abiertas en continuo crecimiento de las entradas sobre personas e instituciones. Aunque no ha sido posible aplicar un control eficaz, al carecer de tesauros en todas las disciplinas, sí se ha podido utilizar una lista de validación para la normalización en los *descriptores*. Sin embargo, no se utiliza una herramienta similar en el campo de *topónimos*, ya que el tesauro se limita a las entidades político-administrativas y en este campo se permite incorporar además otras denominaciones, por ejemplo comarcas y accidentes geográficos.

- *Forma de asentar los nombres de personas*: Estos nombres se asientan como términos de indización cuando el contenido de los documentos así lo requiere. En CLASE se incorporan como *palabras clave*, mientras que en ISOC se consideran *identificadores*. En ambos casos, se ha preferido la forma bibliográfica (entrando por los apellidos) a la forma directa que se usa en otros sistemas. Sin embargo, es necesario normalizar la forma en muchos casos, a través de las normas de uso interno, como en el caso de nombres extranjeros, seudónimos, autores clásicos, monarcas, santos, entre otros.

- *Uso de singular y plural*: En CLASE su uso está regido por el manual y sigue los criterios establecidos en "*Guidelines for the establishment and development of monolingual thesauri*" de 1986 con algunas adecuaciones reseñadas en el "Manual de Indización: Teoría y

práctica" de Isidoro Gil Leiva (2008). Sin embargo, la consistencia en el uso de las formas singular y plural depende fuertemente del cuidado que el analista ponga al momento de consultar la lista. Un analista novato o poco cuidadoso bien puede "inaugurar" la variante de un término ya aceptado, especialmente si éste no se encuentra cercano al término que busca. En el caso de ISOC su manual refleja igualmente la guía básica y ejemplos de elección entre forma singular o plural. A pesar de ello, en el desarrollo histórico del producto, también se producían frecuentes inconsistencias en los índices de descriptores, algunas de ellas provocadas por los propios tesauros utilizados. La homogeneidad mejoró notablemente cuando se aplicó la tabla de validación de los valores admitidos en este campo, que permitió reflejar una única forma y resolver los ejemplos dudosos. No obstante, en algunos casos se admite la doble entrada con significados diferentes, por ejemplo: "árabe" para el idioma y "árabes" como gentilicio.

- *Sinonimia, homonimia y polisemia*: En CLASE no se agrega a los términos un calificador o nota de alcance que permita resolver los casos de sinonimia, homonimia y polisemia. El contexto de estos términos se aclara solamente cuando se ve el registro completo pero se pierde en la lista alfabética disponible para el usuario. En el caso de los sinónimos el analista puede decidir, con el documento enfrente, que ninguno de los sinónimos es el preferido por lo que puede incluir varios en un mismo registro; esto produce dificultades en la recuperación de información pero al mismo tiempo permite un enriquecimiento de los registros. En ISOC sí se emplean calificadores, de forma ocasional en los *descriptores*, por ejemplo en la entrada "Enfermería (Disciplina)"; y de forma sistemática en los topónimos, Así por ejemplo se distingue "Barcelona (Provincia)" de "Barcelona", reservando la entrada simple para las ciudades; o "Guadalajara (MEX)" de "Guadalajara", reservando la forma simple para los topónimos españoles que no coincidan con nombre de países.

- *Extranjerismos*: Nuestra experiencia indica que su uso es mucho más restringido en las Ciencias Sociales y Humanas a diferencia de lo que ocurre en Ciencia, Medicina y Tecnología. Sin embargo, su empleo, especialmente los anglicismos, no está prohibido y menos aún si la literatura especializada opta por ellos como en los casos de "iPod" y "Wi-Fi", presentes en CLASE, o "Blogs" y "Western", utilizados en ISOC.

- *Siglas y acrónimos*. En CLASE se ha usado indistintamente tanto la forma desarrollada como la abreviada, causando problemas innecesarios en la recuperación. Sin embargo, la última versión del manual a publicarse a mediados de 2011 resuelve esta situación optando por la forma desarrollada pero acompañada por la abreviatura o sigla entre paréntesis, por ejemplo: *Partido Revolucionario Institucional (PRI)* con lo cual el usuario puede elegir entre una u otra forma y recuperar los registros correspondientes. En ISOC se ha optado por la forma desarrollada en la mayor parte de los casos ("Pequeña y mediana empresa" en lugar de "pyme", "Unión Europea" en vez de "UE"), utilizando la abreviatura solamente cuando es más conocida que su desarrollo ("FMI", "UNESCO").

3.2. COMPARACIÓN FORMAL DE LOS LÉXICOS DE INDIZACIÓN A PARTIR DE LOS DESCRIPTORES DE ALTA FRECUENCIA DE APARICIÓN EN ISOC

Para establecer la comparación entre los léxicos de indización empleados en ISOC y CLASE se ha tomado como punto de partida la lista de descriptores de la base de datos del CSIC con una alta frecuencia de uso. Se pretendía constatar si estas entradas tenían un empleo igualmente relevante en el producto de la UNAM y detectar ejemplos relevantes de diferencias de criterio en la indización de materias. De los 467 términos con más de 1.000 ocurrencias de uso en ISOC (lo que

puede considerarse un uso muy alto), tan sólo en 56 casos (12%) la entrada no estaba presente en CLASE. Por tanto, aparentemente existía una fuerte similitud formal en la construcción de los términos. Sin embargo, sólo en 140 (29%) les corresponde una frecuencia de uso igualmente elevada en ambas bases de datos, que asciende a 185 (39%) si se tienen en cuenta las entradas con más de 500 registros (véase figura 1). Por tanto, existe una mayoría de términos de uso generalizado en ISOC que no alcanzan un empleo similar en CLASE, un 12% no están presentes y un 49%, aunque efectivamente están utilizados, su frecuencia de registros es considerablemente menor.

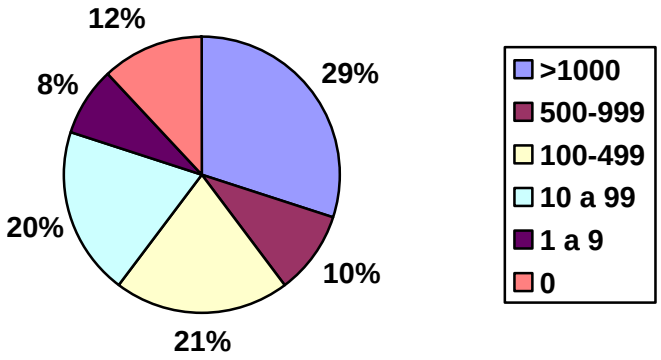

Figura 1. Ocurrencia de uso en CLASE de los descriptores con más de 1.000 registros en ISOC (marzo de 2011)

Se han analizado con mayor profundidad los casos en los que los términos de alta frecuencia en ISOC tienen un uso en CLASE muy bajo o inexistente. De la comparación se establecen diferentes tipos de causas:

- Preferencia por otro término diferente, que alcanza igualmente un alto uso, siendo ambos utilizados de forma habitual por los autores tanto españoles como latinoamericanos. Así por ejemplo, mientras en el CSIC se utiliza “Desarrollo sostenible” o “Enseñanza superior”, en la UNAM se ha optado por “Desarrollo sustentable” y “Educación superior”. En algunos de estos casos la forma es casi idéntica y se limita al uso de singular y plural (Deporte/s, Vivienda/s) o el acrónimo frente al nombre desarrollado (Pymes / Pequeña y mediana empresa). En otros se debe al empleo de forma simple frente al grupo nominal (Asentamientos/Asentamientos humanos). De forma similar existen ejemplos en los que varía la forma más generalizada entre cuasi-sinónimos: jóvenes y juventud aparecen en ambos productos, si bien el primer término es más utilizado en CLASE y el segundo es preferido en ISOC.
- En muy escasos ejemplos, los vocabularios reflejan diferencias en la terminología empleada por los propios investigadores de cada región. Mientras en España se prefiere “Yacimientos arqueológicos”, los autores latinoamericanos utilizan “Sitios arqueológicos”.
- Diferencias de uso que provienen del diferente peso de algunos temas en la investigación española frente a la latinoamericana. Conceptos como “Guerra civil española”, “Segunda República”, “Franquismo”, “Comunidades autónomas”, “Derecho comunitario”, “Mercado único europeo”, “Edad del Bronce”, “Edad del Hierro”, “Catalán” o “Vasco”, apenas están presentes en la literatura publicada en Latinoamérica. Por tanto, podrían tener una aceptación similar, pero su necesidad de empleo es radicalmente diferente.
- Diferencias que traducen divergencias en los criterios aplicados en el proceso de indización por materias, que obedecen a los hábitos de trabajo del equipo de analistas. Así mientras en ISOC el concepto “Numismática” figura en 1.063 registros, por

solamente 17 en CLASE, se debe a la práctica habitual de selección sistemática o no de este tipo de entradas al asignar las materias, más que a la existencia de documentos relacionados con esta disciplina.

3.3. COMPARACIÓN FORMAL DE LOS LÉXICOS DE INDIZACIÓN A PARTIR DE EJEMPLOS DE FAMILIAS LINGÜÍSTICAS SELECCIONADAS

Para completar el análisis comparativo entre los vocabularios de indización utilizados en las bases de datos CLASE e ISOC, se comprobaron las listas de términos que comenzaran por algunas entradas significativas, como ejemplo de denominaciones de familias lingüísticas que hacen referencia a disciplinas y subdisciplinas. En concreto se seleccionaron las entradas que comenzaban por Antropología, Arqueología, Economía, Geografía y Psicología. En ISOC se analizaron los términos de estas familias en el campo de descriptores: la mayoría de ellos tenían una representación media o alta en número de registros de la base de datos (tabla 1). De las 234 entradas, tan sólo un 13% no alcanzaban una frecuencia de uso inferior a 6 documentos, proporción que puede indicar que se trata de términos no consolidados en la bibliografía científica o bien de entradas escasamente representativas para posibles búsquedas bibliográficas.

Tabla 1. Distribución por frecuencia de uso en la base ISOC de los términos de indización (descriptores) que comienzan por las entradas seleccionadas.

Términos que comienzan por…	**>500 regs.**	**6-499 regs.**	**<5 regs.**	**Total**
Antropología	3	24	3	30
Arqueología	2	26	7	35
Economía	6	60	10	76
Geografía	1	42	7	50
Psicología	2	38	3	43
Total	14	190	30	234
Porcentaje	6%	81%	13%	100%

En CLASE se localizaron las entradas de las mismas familias en el campo de palabras clave, excluyendo aquellos casos en los que se reflejaban nombres propios, para realizar la comparación sobre la misma base, la representación de conceptos, disciplinas y subdisciplinas. En esta base de datos se aprecia una mayor dispersión del léxico de indización, aportando 404 entradas diferentes en estas mismas familias, frente a las 234 de ISOC. Como consecuencia de esta dispersión, la mayor parte de las entradas (66%) no superan los 5 registros en la base de datos (tabla 2).

Tabla 2. Distribución por frecuencia de uso en la base CLASE de los términos de indización (conceptos en el listado de palabras clave) que comienzan por las entradas seleccionadas.

Términos que comienzan por…	**>500 regs.**	**6-499 regs.**	**<5 regs.**	**Total**
Antropología	3	15	48	66
Arqueología	1	6	11	18
Economía	13	35	70	118
Geografía	2	12	14	28
Psicología	6	45	123	174
Total	25	113	266	404
Porcentaje	6%	28%	66%	100%

Comparando las dos tablas, se observa que la mayor dispersión del vocabulario en CLASE no se presenta igual en todas las disciplinas seleccionadas. Destaca el ámbito de la Psicología como la familia con mayor diferencia en el número de términos, 174 en CLASE por 43 en ISOC. Por el contrario, en Arqueología y Geografía es la base ISOC la que presenta más entradas diferentes. En estas diferencias influyen la diferente proporción de documentos de estas disciplinas en cada producto y también los hábitos de trabajo de los responsables de la indización en cada caso. El factor humano es importante, pues en ambos sistemas, es el documentalista quien toma las decisiones en la asignación de términos de indización en un registro. Sin embargo, la suma de los ejemplos seleccionados, muestra que la filosofía de indización por palabras clave aplicada en CLASE conlleva una mayor dispersión del vocabulario utilizado frente al modelo de indización por descriptores utilizado en ISOC. De las entradas analizadas en CLASE, la mayor parte (68%) no estaban presentes en ISOC, mientras que por el contrario, la mayoría (56%) de los términos localizados en ISOC si eran coincidentes (tabla 3).

Tabla 3. Comparación del número de términos de indización (conceptos) que comienzan por las entradas Antropología, Arqueología, Economía, Geografía y Psicología, en las bases de datos CLASE e ISOC.

Base de datos	Entradas en los ejemplos analizados	Términos coincidentes	%	Términos únicos	%
CLASE	404	131	32%	273	68%
ISOC	234	131	56%	103	44%

4. CONCLUSIONES

ISOC y CLASE son productos con una gran similitud en su filosofía de trabajo y diferencias concretas en su modo de aplicación y desarrollo. En ambos productos se aplican criterios de normalización que afectan a varios campos de los registros, no solo a la indización por materias, y cuentan para ello con un manual de normas de uso interno. Su desarrollo histórico les permite aportar más de 35 años de experiencia en la acumulación, entre los dos productos, de alrededor de un millón de registros bibliográficos con análisis documental de contenido. Los léxicos utilizados en la indización entre ambos sistemas, pueden servir como modelo de lenguajes documentales para las Ciencias Sociales y Humanas. ISOC y CLASE han definido criterios de análisis documental, que afectan a la construcción de los términos de indización. Sus vocabularios de indización guardan frecuentes coincidencias en la elección de forma y algunas divergencias que son motivadas principalmente por decisiones de los analistas, en muy escasos ejemplos proceden del empleo de variantes terminológicas entre los autores de Latinoamérica y España. Sería muy útil registrar sistemáticamente estas divergencias documentales para la construcción de nuevas herramientas globales de recuperación de bibliografía científica en español para estas disciplinas.

La principal diferencia entre estos productos radica en la apuesta por un modelo de indización por palabras clave en CLASE, frente al empleo en ISOC de descriptores y campos diferenciados para los nombres propios, geográficos, periodos históricos y normativa jurídica. Se trata de dos opciones aparentemente muy diferentes, que sin embargo mantienen un alto grado de convergencia en el seguimiento de normas básicas sobre la construcción de los términos empleados en la indización: limitación de la precoordinación a los conceptos que solo pueden expresarse sin ambigüedad en su forma compuesta, uso normalizado de singular y plural, forma bibliográfica en los nombres de persona o la limitación de extranjerismos, siglas y acrónimos.

Entre las fortalezas del lenguaje natural usado en CLASE se encuentra el hecho de incluir términos muy actualizados para la descripción de documentos, términos que bien podrían no estar aceptados en un tesauro especializado. Esto permite además una mayor exhaustividad en la

descripción de los contenidos, así como libertad para utilizar los términos especializados proporcionados por los propios autores. Por esta misma razón, en el método de trabajo de la base ISOC es esencial el empleo de un campo de uso interno para reflejar los términos candidatos que deben ser sometidos a revisión antes de incorporarse a la tabla de descriptores normalizados.

Otro aspecto a ser valorado es que el lenguaje natural siempre estará más cercano al de los usuarios que consultan CLASE, muchos de ellos alumnos de licenciatura que buscan artículos en español para sus quehaceres universitarios. En términos generales podría decirse que las palabras clave utilizadas en CLASE se asemejan a las que usan algunos recursos de información en Internet muy populares que se basan en una filosofía de flexibilidad y sentido común, como la Wikipedia.

La mayor debilidad es obvia: no se tiene un control total del vocabulario. Esto da lugar a que muchos sinónimos aparezcan en las listas, sin posibilidad hasta ahora de jerarquizar su uso. También un mismo término puede aparecer en su forma singular o plural a pesar de que en el manual referido se documentan ampliamente ejemplos de su uso. Pero se ha detectado que el manual interno no siempre es consultado al momento de realizar el análisis, con lo que se pierde una valiosa referencia para la toma de decisión. Entre las posibles causas de esta situación se encuentra la presión que los analistas tienen por aumentar la producción de registros, por lo que evitan distracciones al momento de realizar su trabajo, entre ellas la consulta puntual del manual disponible en forma impresa. Para paliar esta situación se ha sugerido que la versión 2011 esté disponible en línea con el mayor grado de interactividad para el analista.

La dispersión de entradas en el vocabulario presente en los índices de materias de CLASE es claramente superior a la que ofrece ISOC. La consecuencia es que gran parte de los términos no alcanzan una representatividad suficiente para asegurar su utilidad en la recuperación bibliográfica. El uso de tesauros en ISOC ha permitido un mayor nivel de control de vocabulario. Sin embargo, también en este producto se produjeron numerosas inconsistencias en la forma de los descriptores, que han comenzado a resolverse sobre todo a partir de la aplicación de una tabla de validación en el sistema de grabación de la base de datos.

Las características de los lenguajes de indización afectan a la recuperación de información por materias en estas disciplinas. Especialmente en los ámbitos de Ciencias Humanas y Sociales, es necesario contar con herramientas que faciliten el control de las ambigüedades y las variaciones de la terminología científica. La experiencia de la indización en ambos sistemas muestra que los factores humanos también contribuyen a marcar las diferencias en el análisis de contenido: a menudo son los hábitos del indizador los que marcan la preferencia por determinadas entradas que alcanzan un uso muy elevado en un producto y apenas tienen presencia en otro. En los sistemas referenciales, a menudo las búsquedas sólo pueden ser eficaces si se adaptan a los criterios empleados en los índices de materias. Las inconsistencias en su construcción limitan las posibilidades de recuperación. Sin embargo, los productos documentales basados en palabras clave aportadas por los autores están alcanzando un auge creciente (plataformas de revistas electrónicas, archivos abiertos y repositorios institucionales). Los sistemas de recuperación en estas bases de datos podrían enriquecerse si aplicaran herramientas de control terminológico adaptadas a corpus documentales heterogéneos.

ISOC y CLASE son productos complementarios: no existe solapamiento en su selección de publicaciones fuente y comparten características esenciales en la estructura de la información. Para una posible cooperación entres sus sistemas de recuperación, se deberían tener en cuenta las diferencias existentes entre sus léxicos de indización. La construcción de una tabla de equivalencia entre formas sinónimas o cuasisinónimas, empleadas en la documentación de ambas regiones, sería una herramienta de gran utilidad, también para su aplicación en otros sistemas de información.

BIBLIOGRAFÍA CITADA

ABEJÓN-PEÑA, Teresa; MALDONADO-MARTÍNEZ, Ángeles; RODRÍGUEZ-YUNTA, Luis; RUBIO-LINIERS, María-Cruz. “La base de datos ISOC como sistema de información y fuente para el análisis de las ciencias humanas y sociales en España”. *El profesional de la información*, 2009, vol. 18, n. 5, pp.521-528.
[http://eprints.rclis.org/handle/10760/15056, consultado el 14-03-2011]

ALONSO GAMBOA, José Octavio; ARANA MENDOZA, Celia y SÁNCHEZ PEREYRA, Antonio. *Manual de indización para las bases de datos CLASE y PERIÓDICA*, México, D.F.: UNAM, Dirección General de Bibliotecas, 2011, 110 p. ISBN en trámite.

CINDOC. *Manual de Normas para el Análisis Documental en la Base de Datos ISOC*. Madrid: 2000. (Documento de uso interno, inédito).

GIL LEIVA, Isidoro. *Manual de indización: Teoría y práctica*. Gijón: TREA, 2008, 429 p. ISBN 978-84-9704-367-0.

LANCASTER, F.W. *El control de vocabulario en la recuperación de información*. Valencia: Universidad, 2002, 286 p. ISBN 84-370-5444-3.

LANCASTER, F.W. *Indización y resúmenes: teoría y práctica*. Buenos Aires: EB Publicaciones, 1996, 337 p. ISBN 987-95809-2-3.

REYNA ESPINOSA, Felipe Rafael. “Trascendencia de la Bibliografía Latinoamericana de la UNAM”, *Biblioteca Universitaria*, 2010, vol. 13, no. 2, pp. 164-174

RODRÍGUEZ-YUNTA, Luis. “Las bases de datos documentales del CSIC en el desarrollo histórico del mercado de la información en España (desde sus antecedentes hasta 2008)”. En: *La documentación como servicio público. Estudios en homenaje a Adelaida Román*. Madrid: CSIC, 2009, pp.133-174.
[http://eprints.rclis.org/handle/10760/14820, consultado el 14-03-2011]