Modelos de indización temática utilizados en repositorios digitales de acceso abierto de argentina.

Models used for subject indexing in open access digital repositories of Argentina.

Nadina Yedid.
Universidad de Buenos Aires. Facultad de Filosofía y Letras. Departamento de Bibliotecología.
Dirección postal: Rivera Indarte 438 1° C (1406)
Correo electrónico: nadineyedid@hotmail.com

Resumen: El presente trabajo indaga respecto de las formas de indización utilizadas en los repositorios digitales de acceso abierto en Argentina. El principal objetivo es dilucidar si en estos nuevos recursos, desarrollados en ámbitos exclusivamente digitales, se mantienen las formas tradicionales de indización, o se utilizan nuevas formas de indización asociadas a los entornos digitales. Para ello se establecen cuatro categorías de análisis, siendo ellas: la indización "tradicional" mediante el uso de vocabularios controlados; la indización automática mediante software especializado; la indización mediante palabras claves, ya sea asignadas por personal interno del repositorio, autores, o usuarios; y la no indización (asociada a la recuperación por cadena de caracteres). Se consultan veintiún repositorios digitales de acceso abierto, y se analizan los resultados obtenidos. Las respuestas obtenidas permiten concluir que en este tipo de recursos se mantienen los modelos tradicionales de indización, utilizándose como principal modelo, la indización mediante vocabularios controlados realizada por profesionales.

Summary: This paper investigates on ways of indexing used in open access digital repositories of Argentina. The main objective is to elucidate whether these new resources developed exclusively in digital environments, maintain traditional forms of indexing, or uses new forms of indexing associated with the digital environments. In order to find this out, we establish four categories of analysis, being them: "tradicional" indexing using controlled vocabularios; automatic indexing using specialized software; indexing by keywords, either assigned by repository's indexers, authors, or users; and no indexing (associated with the recovery by string of characters). Twenty-one open access digital repositories are studied, and the results are analyzed. The responses obtained allow to conclude that these resources maintain traditional indexing models, using as the main model, the indexing by controlled vocabularios carried on by professionals.

Palabras claves: Indización, Vocabularios controlados, Indización automática, Etiquetado, Repositorios digitales.
Key words: Indexing, Controlled vocabularios, Automatic indexing, tagging, Digital repositories.

Introducción

En las últimas décadas ha tenido lugar un fenómeno conocido con el nombre de "crisis de las publicaciones científicas", asociado con la escalada en el costo económico de las mismas y la creciente dificultad de las unidades de información para poder adquirirlas. Esta situación, sumada a las nuevas posibilidades de comunicación que ofrece Internet, ha favorecido el desarrollo de un movimiento conocido con el nombre de *Open Access*, o acceso abierto.

Este movimiento plantea la necesidad de generar espacios en los que los científicos puedan hacer circular sus investigaciones y comunicaciones de forma gratuita, para así garantir la difusión de las mismas, y favorecer el desarrollo del conocimiento científico. Con miras a lograr este objetivo, el movimiento *Open Access* plantea dos estrategias posibles: por un lado, la publicación de artículos en revistas de acceso abierto (ruta dorada) y, por otro lado, el archivo de los artículos en repositorios digitales de acceso abierto (ruta verde).

Los repositorios digitales se configuran entonces como recursos cuyo principal objetivo es facilitar el acceso al conocimiento científico. Es de esperar, por lo tanto, que los mismos implementen las herramientas que permitan la más eficaz y eficiente recuperación de información. Dado que la recuperación temática es una de las formas más significativas de acceso a la información, cabe preguntarse qué modelos de indización utilizan los repositorios digitales de acceso abierto de Argentina para representar los contenidos temáticos de los artículos que albergan. Conocer los modelos de indización utilizados, permitirá analizar si estos constituyen, o no, las herramientas más eficaces para la recuperación temática de información en este tipo de recursos, caracterizados por funcionar exclusivamente en entornos digitales.

El presente trabajo se propone conocer las características relativas a la indización temática de documentos en los repositorios digitales de acceso abierto de origen argentino; determinar si se realiza algún tipo de indización temática de los documentos que contienen; e identificar el/los modelo/s de indización más utilizado/s en este tipo de repertorio.

Hipótesis

Los repositorios digitales de acceso abierto de Argentina promueven el reemplazo de la indización temática de los documentos por el almacenamiento de los textos completos, para ofrecer técnicas de recuperación de información basadas en las búsquedas por cadena secuencial de caracteres, dejando de lado otros modelos de indización que podrían mejorar la recuperación de la información para asegurar la transferencia del conocimiento científico.

Marco teórico

Para la presente investigación se toma la definición de repositorios digitales de acceso abierto propuesta por Melero (2005), que define a los mismos como …"archivos donde se almacenan recursos digitales (textuales, de imagen o sonido) [que] surgen de la llamada comunidad *e-print*". Este tipo de archivos puede contener pre-publicaciones de artículos (versiones anteriores a la revisión del referato), post-publicaciones de artículos (versiones ya aprobadas por el referato), como así también tesis, disertaciones, materiales de enseñanza, archivos de datos, registros institucionales, y otros materiales relacionados con la actividad académica (Suber, s.f.). Los repositorios digitales suelen clasificarse en repositorios temáticos o repositorios institucionales. Los primeros son aquellos en los que se reúnen trabajos de investigadores de diferentes instituciones, pero que se enmarcan dentro de la misma disciplina o área del conocimiento (De Volder, 2008). Por su parte, los repositorios institucionales reúnen la producción de una institución (por ejemplo, de una Universidad), para preservar y brindar acceso a los trabajos resultados de la actividad científica o académica llevada adelante en la propia institución (Sánchez García & Melero, 2007). Ambos tipos de repositorios serán considerados dentro de esta ponencia.

Para comprender las formas de indización utilizadas en este tipo de recursos, se establecen tres categorías de análisis: La primera responde a la indización mediante el uso de algún tipo de vocabulario controlado, ya sea un tesauro, una lista de encabezamientos de materia, o una ontología, realizada por indizadores humanos. La segunda responde a la llamada indización automática, en lenguaje natural. La tercera responde a la indización mediante etiquetas (no controladas), ya sea asignadas por el autor, o bien por la comunidad de usuarios, como en el caso de las folksonomías. A los fines de comprobar o refutar la hipótesis establecida, se utilizará además una cuarta categoría de análisis que responde al almacenamiento

de los textos completos para permitir la recuperación de la información contenida en los mismos a partir de búsquedas por cadenas secuenciales de caracteres.

Un vocabulario controlado consiste básicamente en una lista de términos autorizados, que incluye algún tipo de estructura semántica. Esta estructura permite, en primer lugar, controlar los sinónimos y diferenciar los homógrafos. Además, facilita la reunión o vinculación de términos que presentan una relación conceptual entre sí (Lancaster, 1996). De acuerdo con Taylor (2003) los vocabularios controlados pueden ser de tres tipos: tesauros, listas de encabezamiento de materias, y ontologías. Tanto los tesauros como las listas de encabezamientos de materias proponen una lista de términos autorizados, con referencias hacia las formas no autorizadas, y demarcan las relaciones existentes entre los términos. Las principales diferencias entre estos dos tipos de vocabularios se relacionan principalmente con la forma en que se establecen las relaciones entre los términos. Los tesauros responden al tipo de vocabulario postcoordinado, en donde los descriptores se asignan de forma discreta, y luego se coordinan al momento de realizar la búsqueda mediante el uso de operadores booleanos. Por su parte, los términos que conforman las listas de encabezamiento de materias, llamados epígrafes, suelen incluir frases precoordinadas. (Taylor, 2003). Además, los tesauros resultan mucho más específicos al determinar las relaciones existentes entre los diferentes términos descriptores. Las ontologías, por su parte, permiten describir y representar un área de conocimiento. Se utilizan para compartir información en un dominio particular, tanto entre personas como, principalmente, entre personas y computadoras. Son similares a los tesauros y a las listas de encabezamiento de materias en el hecho de que reúnen todas las formas variantes para expresar un concepto, y en establecer relaciones entre los mismos. La principal diferencia es que en las ontologías cualquiera de los términos que se ingrese en el momento de la búsqueda, permitirá recuperar la información para todos los términos equivalentes (Taylor, 2003).

El uso de un vocabulario controlado en la indización comporta grandes beneficios, tanto en el momento de la indización, como en el momento de la recuperación. Esto se debe precisamente al control de la sinonimia, la polisemia y la homonimia, y al establecimiento de relaciones entre los términos. Sin embargo, su uso trae aparejadas también algunas desventajas, sobre todo en lo que respecta a los aspectos económicos. Claramente, la utilización de un vocabulario controlado en la indización requiere más tiempo, indizadores más calificados y, por lo tanto, resulta más costosa que los modelos de indización que no utilizan vocabularios controlados. Esta es una de las principales desventajas de los mismos, y ha

llevado a que muchos sistemas de información utilicen el lenguaje natural en la indización, ya sea mediante la asignación de etiquetas no controladas, o mediante la indización automática.

Dentro del contexto de los sistemas de recuperación de información, la expresión “lenguaje natural” suele utilizarse para referirse a las palabras que aparecen en los propios documentos, es decir, que fueron utilizadas por el autor para expresar sus ideas, en oposición al vocabulario controlado. (Lancaster, 1996)

Posiblemente, la principal ventaja de un sistema de recuperación mediante lenguaje natural, en oposición al vocabulario controlado, reside justamente en que se trata de un vocabulario ilimitado, lo que permite una gran especificidad en la recuperación de información. Los vocabularios controlados se basan en la reunión de conceptos próximos bajo un mismo término descriptor como parte de su estrategia para evitar la diseminación de la información. Así mismo, es posible que un vocabulario controlado no alcance todos los niveles de especificidad que pueden tener los documentos, sino que exija indizarlos mediante un término más general que los abarque. En cualquiera de estos dos casos, el lenguaje natural ofrece la ventaja de proveer mayor especificidad, y ser capaz de distinguir matices de significado (Lancaster, 1995). Otro de los factores que deben ser considerados es la novedad de los términos. Nuevas voces y términos aparecen en los documentos mucho antes de que puedan ser reflejadas en un vocabulario controlado. Por este motivo, también para las búsquedas de temas que resulten novedosos el lenguaje natural se presenta como la mejor opción para la recuperación (Lancaster, 1996).

Según Taylor (2006), los indizadores, como cualquier ser humano, son falibles, inconsistentes, subjetivos, operan a un ritmo acorde a sus capacidades intelectuales y son el componente más costoso del proceso de indización de documentos. Es sabido que muchas veces existe poca coherencia o consistencia en la indización llevada a cabo por indizadores humanos, ya que la capacidad de indización se basa tanto en las habilidades intelectuales de la persona, como en su formación profesional, su experiencia y el grado de conocimientos en la disciplina que abarca el documento (Lancaster, 1996). Además, un indizador humano puede dejar de lado temas que aparecen en el documento, por no considerarlos importantes para el usuario de ese sistema de información, o simplemente por una cuestión de distracción. Las personas, en su trabajo, se encuentran altamente influenciadas por factores externos e internos, como pueden ser las condiciones de trabajo, la fatiga, o el humor. Todo esto puede afectar la forma en que un documento es indizado y, por tanto, afectará también su recuperación. Por otro lado, tal como se expuso en apartados anteriores, la indización hecha por seres humanos toma tiempo, lo que no se adapta a la constantemente creciente cantidad de documentos que se incorporan diariamente a una base de datos, sobre todo desde la

explosión del ambiente digital. Y, por sobre todas las cosas, la indización hecha por humanos resulta costosa en términos económicos, ya que se debe pagar gran cantidad de horas a personal capacitado para que realice esta tarea, lo que no se condice con el decreciente presupuesto que observan hoy la mayoría de los centros de información.

Por todos estos motivos, en la actualidad muchos sistemas de información han optado por utilizar métodos de indización que requieran poca o ninguna intervención humana, como sucede con la indización automática. Tal como lo indica su nombre, la indización automática consiste en el proceso de indización, por extracción o asignación de términos, pero hecha por una computadora, sin intervención humana (Obaseki, 2010). De acuerdo con Lancaster (1996) existen actualmente dos formas de llevar a cabo una búsqueda en texto libre, la primera es a partir de la llamada "indización automática" que implica la generación de un índice conocido con el nombre de "archivo invertido", en el que figuran todas las palabras claves existentes en la colección, y se indica en qué documento aparece cada palabra. La segunda forma se basa en la incorporación de los textos completos a la base de datos, para luego realizar sobre los mismos búsquedas secuenciales, palabra por palabra, o caracter por caracter. Lancaster llama a este modelo el método caudaloso, o de "no indización", y se corresponde con la cuarta categoría establecida en esta investigación.

La indización automática es un modelo de indización muy utilizado y se realiza principalmente mediante la extracción de palabras del propio texto. En su forma más sencilla, la extracción se realiza a través del "método de frecuencia absoluta": el programa funciona mediante un algoritmo que compara las palabras encontradas en los textos contra una lista de palabras prohibidas o stop list. La stop list se compone de palabras no significativas, que no serían útiles como términos de indización, tales como artículos, preposiciones, conjunciones y otras. Todas las palabras que no figuran en la stop list son ordenadas de acuerdo a su frecuencia de aparición en el texto, es decir, se ubican más arriba en la lista aquellas palabras que más frecuentemente aparecen en el documento. Las palabras que más frecuentemente aparezcan en el texto, serán seleccionadas para ser los "términos de indización". Algunos programas realizan un proceso intermedio, llamado *stemming*, en el que reducen las palabras a su raíz antes de incluirlas en el índice. Además de este método, existe también el método de frecuencia relativa, que compara la frecuencia de aparición de una palabra en un texto, con su frecuencia de aparición en toda la base de datos, para determinar qué palabras tendrán mayor poder de discriminación (Lancaster, 1996). En la actualidad existen también gran cantidad de programas que trabajan con algoritmos de ponderación de términos,

para permitir una indización basada no sólo en la frecuencia, sino también en la posición de las palabras en el texto. Además, existe toda una corriente de investigación que trabaja en el desarrollo del "procesamiento del lenguaje natural". Estas investigaciones se orientan a encontrar algoritmos que permitan realizar la indización automática de los documentos sobre las bases de un análisis morfológico, sintáctico y hasta semántico de las palabras contenidas en los textos (Taylor, 2003 y 2006; Coyle, 2008).

Otro método de indización automática es aquel que en lugar de extraer los términos del propio texto, indiza mediante la asignación de términos de un vocabulario controlado. En estos casos, el programa suele contar con un diccionario en donde cada palabra autorizada presenta un "perfil" de palabras o frases que pueden ocurrir en un texto. Cada vez que el sistema identifica alguna de esas palabras o frases, asigna el término correspondiente. (Lancaster, 1996). Estos programas, evidentemente, resultan de muy difícil aplicación y están aún en desarrollo, por lo que no serán tenidos en cuenta dentro de esta categoría de análisis.

En base a lo expuesto en este apartado, se establecen entonces dos categorías de análisis diferenciadas. La primera estará conformada por todos los modelos de indización que realicen el proceso de forma automática y por extracción de palabras del texto, más allá del método elegido para llevar adelante la extracción (estadístico, sintáctico, morfológico, ec.). La segunda estará conformada por aquellos sistemas que recurran al llamado "método caudaloso" o de NO indización, es decir, que realicen búsquedas por cadena de caracteres en todo el documento.

Además de la ya tradicional indización automática de documentos, han surgido nuevas formas de describir el contenido temático de un objeto digital a partir del auge de la web 2.0. La filosofía 2.0, basada principalmente en el uso de la inteligencia colectiva, la participación del usuario y la colaboración, llevó al desarrollo de una nueva forma de descripción temática conocida con el nombre de *tagging* o etiquetado. El etiquetado no es más que la asignación de palabras claves (no controladas) a un documento u objeto digital (como puede ser un sitio web, una fotografía, un video, etc.). En el entorno de la web 2.0, estas palabras claves reciben el nombre de etiquetas, y la colección de etiquetas utilizada dentro de una plataforma es llamada "folksonomía" (Weller, 2007). Una de las principales ventajas de las folksonomías, es que permiten generar una descripción de los contenidos basada en una "indización emergente". Es decir, que existe un consenso entre los usuarios respecto de cuáles son las etiquetas que mejor permiten representar los contenidos. Estas etiquetas serán las que aparezcan como las etiquetas más populares dentro de la nube. Tal como explica Woolwine (2011), distintos estudios han demostrado que las

folksonomías siguen una distribución de escala libre, en donde unas pocas etiquetas son altamente utilizadas, mientras que una gran cantidad de etiquetas reciben muy poco uso. Por este motivo, las folksonomías mejoran cuanto más personas participan, incrementando la impronta de la inteligencia colectiva. De acuerdo con Spiteri (2007), este tipo de indización aparece como un movimiento opuesto a los esquemas tradicionales de indización, jerárquicos y autoritarios, que reflejan un punto de vista externo que no siempre coincide con la visión de los usuarios.

Moreiro González (citado por Soler Monreal & Gil Leiva, 2010) da cuenta de otros beneficios que pueden aportar las folksonomías: para comenzar, implican un muy bajo costo en comparación con otros modelos de indización (como el uso de vocabularios controlados). Además requieren muy poco esfuerzo, ya que se trata de sistemas de indización muy simples y de uso sencillo, en donde el trabajo de descripción lo hacen los propios usuarios (en contraposición con los sistemas que requieren de indizadores profesionales). Por otro lado, la presentación en forma de nubes de etiquetas resulta visualmente atractiva y la navegación mediante las etiquetas relacionadas permite el hallazgo fortuito (*serendipity*). Finalmente, las folksonomías permiten la coexistencia de diversos puntos de vista, y reflejan la frescura y dinamicidad de la lengua (Soler Monreal & Gil Leiva, 2010). Este último punto constituye una de las principales ventajas de las folksonomías, ya que estas capturan el lenguaje en uso activo de una comunidad, facilitando la inclusión rápida de nuevos conceptos que puedan surgir en el área (Spiteri, 2007). Sin embargo, tal como explica Weller (2007), su principal fortaleza es también su principal debilidad, ya que la completa libertad en la elección de palabras para generar etiquetas resulta en un vocabulario sin ningún tipo de control, lo que expone a todo el modelo a los mismo problemas que presenta cualquier modelo de indización mediante lenguaje natural. Entre los principales problemas, Spiteri (2007) destaca el alto nivel de ambigüedad, la ausencia total de control de la sinonimia y la polisemia, y la variación de nivel básico. Esto último implica que un mismo objeto digital puede ser descrito por una gran cantidad de etiquetas relacionadas que implican un mayor o menor nivel de especificidad. Todos estos aspectos resultan en la generación redundante de etiquetas diferentes para representar los mismos conceptos. Y esto sin contar las etiquetas que se duplican por existencia de errores ortográficos o diferencias en la forma de escritura de las palabras. Por otro lado, Rolla (2009) destaca el hecho de que en muchas ocasiones, las etiquetas se construyen con términos de naturaleza personal, que sólo resultan de utilidad para quien creó la etiqueta o su grupo de pertenencia. Finalmente, Moreiro González también apunta el hecho de que en las folksonomías sólo existen relaciones de asociación, por

co-aparición de etiquetas, no pudiendo establecerse ningún otro tipo de relación entre los términos que conforman las etiquetas (Soler Monreal & Gil Leiva, 2010).

Por último, cabe destacar que si bien la forma tradicional de las folksonomías implica la asignación de etiquetas por parte de los usuarios de los recursos, también existe la posibilidad de que sean los propios autores de los documentos quienes se ocupen de realizar el etiquetado. A los fines de esta investigación, los dos tipos de etiquetados serán contemplados dentro de esta categoría, tanto el modelo tradicional de folksonomía, en la que los usuarios de los recursos son quienes asignan las etiquetas, como el modelo según el cual los propios autores de los documentos son los encargados de etiquetarlos.

Aspectos metodológicos

En la presente investigación, de carácter exploratorio, se empleará el método cuantitativo, como forma de configurar una investigación de tipo descriptiva. Se toma como universo de la presente investigación los repositorios digitales de acceso abierto de origen argentino, ya sea que se traten de repositorios temáticos o repositorios digitales. La muestra se encuentra conformada en base a los repositorios registrados en los directorios internacionales ROAR y Open Doar. Además se incluyen otros repositorios, que resultan relevantes por su envergadura (por ejemplo, repositorios correspondientes a universidades nacionales, o que albergan gran cantidad de documentos relativos a una temática particular). Se identifican un total de veintisiete repositorios, de los cuales se consigue relevar con respuestas positivas a veintiuno (identificados en el Anexo) que conforman la muestra final.

La técnica utilizada en la presente investigación es la encuesta, redactada a los fines específicos de esta investigación. Se utiliza como instrumento de recolección de datos, un cuestionario de diecinueve preguntas, tanto abiertas como cerradas, diseñado ad-hoc para la presente investigación. Los porcentajes obtenidos son redondeados a números enteros para evitar complejidades en la lectura y comprensión de los datos. Para el análisis estadístico de las diferentes variables se toma como población el total de las respuestas positivas (obtenidas), ignorando las no-respuestas.

Resultados

Indización temática de los documentos incorporados al repositorio

De los veintiún casos encuestados, sólo diecinueve declaran realizar algún tipo de indización temática de los documentos incorporados. Sin embargo, en base al análisis de las encuestas completadas por los dos repositorios que indican no realizar ningún tipo de indización temática, puede inferirse que sí realizan algún tipo de indización según los modelos establecidos para esta investigación. En este sentido, uno de los repositorios realiza indización mediante vocabulario controlado y etiquetas no controladas, y el otro repositorio realiza indización automática como así también indización mediante etiquetas no controladas. Teniendo esto en consideración, se puede establecer a partir del análisis de las respuestas obtenidas, que el 100% de los repositorios encuestados realizan algún tipo de indización temática de los documentos al momento de su incorporación a la base.

Indización mediante vocabulario controlado

Son diecisiete los casos que declaran utilizar algún tipo de vocabulario controlado en la indización temática de los documentos. Deberá considerarse además un caso que, si bien contestó que no utiliza vocabulario controlado, en base al resto de las respuestas proporcionadas en la encuesta, se infiere que sí utiliza un vocabulario en la indización de los documentos. Por su parte, dos repositorios responden negativamente la pregunta, indicando que no utilizan este tipo de lenguajes en la indización de documentos, mientras que un tercer repositorio no responde la pregunta, por lo que es desestimado de la muestra para el análisis de esta sección. De esta forma, de la muestra total de veinte repositorios que respondieron la pregunta, aquellos que utilizan un vocabulario controlado totalizan el 90% de la muestra (18 repositorios), mientras que sólo el 10% (2 repositorios) no utiliza este modelo de indización.

Los dieciocho repositorios que emplean vocabularios controlados indican que los términos del vocabulario son asignados por personas: ninguno utiliza un software que permita la asignación automática de términos pertenecientes a un vocabulario controlado. En el 60% de los casos esta indización es realizada por bibliotecarios, mientras que en el 40% restante la misma es llevada a cabo por personal

interno que se especializa en otra área del conocimiento (muchas veces asociada a la temática principal del repositorio).

Respecto de la cantidad de documentos indizados mediante vocabulario controlado, el 60% indiza la totalidad de los documentos incorporados al repositorio. En mucha menor cantidad, sólo tres repositorios (16%) indizan entre el 75 y el 100% de los documentos que incorporan a sus bases. Aún en menor cantidad, dos repositorios (12%) indizan entre el 50 y el 75% de los documentos. Y finalmente, sólo un repositorio (6%) indiza entre el 25 y el 50% de los documentos, y otro repositorio (6%) entre el 0 y el 25 % de los documentos ingresados. Se puede decir entonces que casi la totalidad de los repositorios digitales indizan por lo menos más del 50% de los documentos incorporados a sus bases.

Figura 1

En cuanto al tipo de vocabulario controlado empleado, sólo tres repositorios declaran utilizar más de un tipo de vocabulario controlado. Mientras que el tesauro es utilizado por la mayoría de los repositorios (72%), el 33% utiliza una lista de encabezamientos de materia; siendo poco significativo el uso de ontologías (tan sólo un caso y lo hace conjuntamente con el uso de una lista de encabezamientos de materia. Finalmente, un repositorio (6%) declara utilizar un vocabulario controlado propio, que no incluye en ninguna de las categorías establecidas en la encuesta (tesauro, lista de encabezamientos, ontología y taxonomía).

Figura 2

De la totalidad de los repositorios que indizan mediante vocabularios controlados, seis (33%) utilizan vocabularios desarrollados ad-hoc para esa institución en particular. Diez repositorios (55%) utilizan vocabularios consolidados, de uso común. Finalmente, los dos repositorios restantes utilizan tanto vocabularios ad-hoc como consolidados.

Figura 3

Indización automática de documentos

De los veintiún repositorios encuestados, ocho (42%) declaran realizar indización automática de los documentos incorporados, mientras que otros once (58%) declaran no realizar indización automática.

Los ocho repositorios que llevan a cabo este tipo de indización indican que el proceso se realiza a partir de la extracción de palabras en el documento. Si se consideran además las respuestas obtenidas en la sección anterior, en la que todos los repositorios que trabajan con vocabularios controlados indican que la asignación de los mismos es realizada por personas, el análisis conjunto de ambas variables permite establecer que en ninguno de los repositorios encuestados se lleva a cavo el proceso de asignación automática de términos de vocabularios controlados por medio de software sin intervención humana. En todos los casos las palabras son extraídas del título y del resumen de los documentos. Además, en cinco de los ocho repositorios también se extraen palabras del cuerpo del texto (o texto completo). Dos de los ocho repositorios indican otras fuentes de extracción de palabras, como ser el autor (indicado por los dos repositorios), el título de la revista, el evento científico y las palabras claves. Sin embargo, excepto en lo que se refiere a las palabras claves, los demás campos no parecen corresponder a elementos relacionados con la indización temática, sino con otro tipo de datos identificatorios de los documentos.

Figura 4

En cuanto al algoritmo a partir del cual se extraen las palabras que configurarán los términos de indización, tres de los ocho repositorios declaran desconocer cómo funciona dicho algoritmo. En los restantes cinco repositorios, todos indican entre las bases de funcionamiento del algoritmo de extracción la frecuencia de aparición de las palabras, ya sea simple (cuatro repositorios) o ponderada (un repositorio), es decir, la frecuencia de aparición de una palabra en el texto medida en relación a la frecuencia de aparición de dicha palabra en la totalidad de la base de datos (un repositorio). Además, tres de los cinco repositorios agregan un segundo método de extracción, que es el lugar donde se encuentran las palabras (por ejemplo, palabra del título). A su vez, otro de estos cinco repositorios declara emplear un método de extracción relacionado con el análisis sintáctico de las palabras. Finalmente, uno de estos repositorios indica un tercer método de extracción utilizado: el análisis semántico de las palabras. En este caso, dado que, como es sabido, los desarrollos para la generación automática de metadatos a partir de la comprensión semántica de las palabras, se encuentran aún en estado muy incipiente, resulta poco válida la respuesta obtenida. Se tiene presente por lo tanto, que la respuesta puede tratarse de un error conceptual por parte del encargado de responder la encuesta, o falta de conocimiento del software de indización, o simplemente una mala comprensión o llenado de la encuesta.

Teniendo en cuenta todas las respuestas obtenidas, y considerando la salvedad indicada en el párrafo anterior, se pueden establecer las siguientes aproximaciones:

Figura 5

No indización (o recuperación por cadena de caracteres)

De los veintiún repositorios encuestados, en doce (60%) el sistema almacena los textos completos para permitir la recuperación por cadena secuencial de caracteres. Otros ocho (40%) indicaron no permitir este tipo de recuperación, mientras que un último repositorio no dio respuesta a la pregunta. De los 12 casos que sí cuentan efectivamente con este tipo de recuperación, sólo cinco realizan además indización automática de los documentos, mientras que en los restantes siete repositorios este es el único tipo de recuperación facilitada por medios automáticos.

Figura 6

Indización mediante etiquetas no controladas

El 84% de los repositorios indicaron utilizar etiquetas no controladas. En todos los casos se permite la asignación de etiquetas por parte del personal interno del repositorio. Entre este 84%, existe un 37% en los cuales la asignación es realizada exclusivamente por personal interno, mientras que el 47% permite además que los autores de los trabajos agreguen sus propias etiquetas no controladas. Finalmente, 16% de los repositorios reservan la asignación de etiquetas no controladas para uso exclusivo de los autores que remiten sus trabajos al repositorio. Ninguno de ellos permite la asignación de etiquetas por parte de aquellos que sólo consultan los documentos disponibles en el repositorio.

Figura 7

Paradójicamente, en los repositorios que permiten la asignación de etiquetas por parte de los autores, es en donde se observa el menor control sobre las etiquetas asignadas. En este sentido, en los tres repositorios en los que la asignación de etiquetas recae exclusivamente sobre los autores, sólo en uno se realiza algún tipo de control sobre las etiquetas asignadas. En los repositorios en los que la tarea es compartida entre autores y personal interno, las tareas de control aparecen apenas en la mitad de los casos, existiendo cinco repositorios que realizan algún tipo de control, y cuatro que no llevan adelante

ningún control. Finalmente, el mayor control se observa en aquellos repositorios donde el propio personal realiza la asignación de etiquetas: seis repositorios llevan a cabo un control de las etiquetas asignadas, y sólo uno permite la asignación sin ningún tipo de control. La investigación no permitió relevar específicamente los aspectos sobre los cuales se realiza el control: errores conceptuales, errores de tipeo, etc.

Respecto de la visualización de la nube de etiquetas (que podría funcionar como una ayuda para la normalización en la asignación de etiquetas) sólo uno de los tres repositorios en los que la asignación de etiquetas es realizada exclusivamente por los autores, muestra la nube de etiquetas. Así mismo, sólo tres de los nueve repositorios en los que la asignación de etiquetas es compartida por los autores y el personal interno muestran la nube de etiquetas. Finalmente, dos de los siete repositorios en los que las etiquetas son asignadas exclusivamente por el personal interno, también permiten ver la nube de etiquetas. En todos los casos, el porcentaje de repositorios que muestra la nube de etiquetas ronda entre el 28 y el 33 %.

Análisis general de los resultados

Si se analizan los resultados en general, se puede observar que los modelos más utilizados son: el modelo de indización mediante vocabulario controlado (86%) y la indización mediante etiquetas no controladas (90%). Por su parte, el almacenamiento para la recuperación por cadena de caracteres es realizado por el 52%. Mientras que la indización automática es apenas realizada por el 38% de la población analizada.

Figura 8

Finalmente, resulta interesante conocer cómo se combinan dichos modelos de indización. Así, se puede establecer que casi la totalidad de los repositorios utilizan más de un método de indización temática y sólo uno utiliza un único método de indización: la asignación no controlada de etiquetas. La combinación observada en la mayoría de los casos es el uso de un vocabulario controlado en conjunto con la asignación de etiquetas no controladas (52%). De estos once, cinco agregan además la posibilidad de recuperación por cadena de caracteres. Con menor cantidad siguen los repositorios que combinan los tres métodos de indización observados, y además permiten la recuperación por cadena de caracteres, es decir, que responden a las cuatro categorías de análisis (19%). Ya con mucha menor cantidad, se encuentran los

repositorios que combinan los tres métodos de indización, pero que no permiten la recuperación por cadena de caracteres (9%).

Figura 9

Conclusiones

La investigación permite inferir, en base a las tendencias observadas, que en los repositorios digitales de acceso abierto de argentina se realiza por lo menos algún tipo de indización temática de los documentos para facilitar la recuperación temática de los mismos. Además, la investigación permite identificar las tendencias en cuanto a qué tipo de modelos de indización son los más utilizados en este tipo de recursos.

En contraposición a lo expuesto en la hipótesis de trabajo, la gran mayoría de los repositorios se vale de más de un modelo de indización para facilitar el acceso a la información. A pesar de tratarse de recursos altamente dinámicos, de crecimiento acelerado, la mayoría de los repositorios estudiados utiliza el modelo tradicional de indización mediante el empleo de un vocabulario controlado, que requiere gran cantidad de esfuerzo e inversión de tiempo por parte de indizadores profesionales, pero incorpora un valor agregado inestimable a los metadatos de recuperación. Resulta interesante destacar que los repositorios que indizan los documentos mediante el uso de un vocabulario controlado, indican que los términos del vocabulario son asignados por personas. Esto refuerza la idea expresada en el marco teórico, respecto del insipiente estado de desarrollo de softwares que permitan la asignación automática de términos provenientes de un vocabulario controlado (es decir, asignación de términos realizada exclusivamente por un software, sin participación humana).

Se observa también un amplio uso del modelo de indización mediante etiquetas no controladas. Sin embargo, en la gran mayoría de los casos, dichas etiquetas son asignadas por los propios colaboradores del repositorio, es decir, que se utilizan como un complemento a otras formas de indización, pero no como una forma de “socializar” la indización de los contenidos. Aquellos repositorios que hacen una mayor distribución del trabajo, permiten también que los autores asignen las etiquetas no controladas. Sin embargo, ningún repositorio de los observados permite la asignación de etiquetas por parte de los usuarios del recurso.

En mucha menor medida aparece el uso de indización automática. Posiblemente, la explicación de esta tendencia puede encontrarse en el hecho de que muchos recursos de información, como en este caso los repositorios, se muestran reticentes a la utilización de la indización automática debido a que la misma no ha alcanzado aún lo que se conoce con el nombre de “patrón oro” o “patrón dorado”. Es decir, que la indización obtenida por este medio no tiene aún la misma calidad que puede alcanzar un buen indizador humano. Posiblemente, la tendencia ilustre una cierta preocupación por mantener la calidad de la indización por sobre la rapidez y la economía en el proceso. Por otro lado, existe la posibilidad de que el bajo nivel de utilización de la indización automática se deba a otras cuestiones, como la reticencia al cambio, la falta de conocimientos o asesoramiento informático, o el incipiente estado de desarrollo de los softwares de indización. Futuras investigaciones podrán centrarse en aclarar estos aspectos cualitativos relacionados con la elección de un método por sobre otros.

En referencia al almacenamiento del texto completo, y en contra de lo establecido en la hipótesis de trabajo, apenas poco más de la mitad de los repositorios encuestados almacenan el texto completo que permita la recuperación por cadena de caracteres. Resulta interesante destacar que en todos los casos, dicha tarea es llevada adelante como complemento a las tareas de indización. Esta observación permite refutar la hipótesis presentada, ya que no se advierte que los repositorios reemplacen la indización de documentos por este tipo de recuperación. Por el contrario, es mucho mayor la cantidad de repositorios que realiza uno o más tipos de indización, que la cantidad de repositorios que permiten la recuperación por cadena de caracteres.

Finalmente, se puede establecer como conclusión general que, si bien se pudieron identificar los modelos de indización más utilizados en los repositorios, y conocer la forma en que son empleados, resulta muy difícil determinar si estos constituyen los modelos más eficaces y eficientes para facilitar la recuperación de la información en este tipo de recursos. Como se expuso en el marco teórico, todos los modelos plantean sus ventajas y sus desventajas, y resulta difícil aseverar cuáles de ellas se ven potenciadas en este tipo de recursos de información. Claramente, se evidencia una tendencia hacia la priorización de la calidad de los metadatos y al reconocimiento de la “voz autorizada” (tanto en el uso del vocabulario controlado, como en las restricciones en la asignación de etiquetas no controladas). Sin embargo, cabe preguntarse, ¿es esta la mejor postura a adoptar en el caso de recursos marcados por un alto dinamismo?. Por otro lado, en pocos recursos científicos se evidencia tanto como en los repositorios digitales, el alto grado de solapamiento que existe entre los generadores y los usuarios de los contenidos. En este sentido,

se puede decir que los usuarios de estos recursos son tanto quienes publican sus documentos como quienes los consultan y, posiblemente, sean las mismas personas físicas quienes desempeñan ambos roles. Claramente el auge de este tipo de recursos se encuentra enmarcado dentro de un ambiente digital perteneciente a la era de la web 2.0, y se encuentra influenciado por la misma. Sin embargo, los recursos parecen apegarse a un modelo más tradicional de indización de la información, pensado originalmente para otro tipo de recursos. ¿Es correcto el traslado de este modelo de indización a los nuevos modelos de circulación de la información? ¿Continúan las ventajas siendo más importantes que las desventajas? ¿Son tantas las desventajas que presentan el uso de las folksonomías y la indización automática respecto a los métodos tradicionales de indización, que ni siquiera son utilizadas complementariamente? Todas estas preguntas quedan sin respuesta, y seguramente, podrán ser contestadas con futuras investigaciones que se ocupen de indagar en estos aspectos.

Bibliografía citada

Coyle, K. (2008). Machina indexing. *The Journal of Academia Librarianship*, 34, (6), pp. 530-531.

De Volder, C. (2008). Los repositorios digitales de acceso abierto en la Argentina: situación actual. *Información, Cultura y Sociedad*, 19, pp. 79-98. [Recuperado junio 10, 2012, de http://www.scielo.org.ar/pdf/ics/n19/n19a05.pdf].

Lancaster, F. W. (1995). *El control del vocabulario en la recuperación de información*. Valencia: Universitat de València.

Lancaster, F. W. (1996). *Indización y resúmenes: teoría y práctica* (E. Barber, trad.). Buenos Aires: EB publicaciones.

Melero, R. (2005). Acceso abierto a las publicaciones científicas: definición, recursos, copyright e impacto. *El profesional de la Información*, 14, (4), pp. 255-266. [Recuperado junio 10, 2012, de http://www.elprofesionaldelainformacion.com/contenidos/2005/julio/3.pdf].

Obaseki, T. I. (2010). Automated indexing: the key to information retrieval in the 21st century. *Library philosophy and practice.* [Recuperado junio 10, 2012, de http://digitalcommons.unl. edu/cgi/viewcontent.cgi?article=1347&context=libphilprac].

Rolla, P. J. (2009). User tags versus subject headings: can user-supplied data improve subject access to library collections?. *Library Resources & Technical Services*, 53, (3), pp. 174-184. [Recuperado junio 10, 2012, de http://www.ala.org/alcts/sites/ala.org.alcts/files/content/ resources/lrts/archive/53n3.pdf].

Sánchez García de las Bayotas, D. S. & Melero, R. (2007). *La denominación y el contenido de los repositorios institucionales en acceso abierto: base teórica de la "ruta verde"*. [Recuperado junio 10, 2012, de http://digital.csic.es/bitstream/10261/1487/1/OA2rm.pdf].

Soler Monreal, C. & Gil Leiva, I. (2010). Posibilidades y límites de los tesauros frente a otros sistemas de organización del conocimiento: folksonomías, taxonomías y ontologías. *Revista Interamericana de Bibliotecología*, 33, (2), pp. 361-377. [Recuperado junio 10, 2012, de http://www.scielo.unal.edu.co/scielo.php?script=sci_arttext&pid=S0120-09762010000200004&lng=es&nrm=].

Spiteri, L. F. (2007). The structure and form of folksonomy tags: the road to the public library catalogue. En Rodríguez Bravo, B. & Alvite Diez, M. L. (Eds.), La *interdisciplinariedad y la transdisciplinariedad en la organización del conocimiento científico* (pp. 459-468). León: Universidad de Leópn. [Recuperado junio 10, 2012, de http://scholar.google.com.ar/ scholar_url?hl=es&q=http://dialnet.unirioja.es/servlet/ dcfichero_articulo%3Fcodigo%3D2534223&sa=X&scisig=AAGBfm1T8GV2Kcf44mYPKbckWoY7Fd1A&oi=scholarr&ei=Rib7T4CKG4qY8g TLoozMBg&ved=0CFAQgAMoADAA].

Suber, P. (s.f.). *Open Access Overview*. [Recuperado junio 10, 2012, de http://www.earlham.edu/~peters/fos/overview.htm].

Taylor, A. G. (2006). *Introduction to cataloging and classification*. Englewood, Colorado: Libraries Unlimited.

Taylor. A. G. (2003). *Organization of information*. Englewood, Colorado: Libraries Unlimited.

Weller, K. (2007). Folksonomies and ontologies: two new players in indexing and knowledge representation. En *Online Information 2007 Proceedings*. [Recuperado junio 10, 2012, de http://www.phil-fak.uni-duesseldorf.de/fileadmin/Redaktion/Institute/Informationswissenschaft/ weller/1197280560weller009p.pdf].

Woolwine, D. (2011). Folksonomies, social tagging and scholarly articles. *The Canadian Journal of Information and Library Science*, 35, (1), pp. 77-92.

Anexo: Repositorios relevados

Figura 10

Notas: Se desea agradecer por su apoyo y colaboración a: Elsa Barber, Carolina Gregui, Gabriela De Pedro, Maria Rosa Mostaccio, Dominique Babini, Diego Ferreyra, Fernando López, y a todos los responsables y colaboradores de los repositorios encuestados, sin cuya participación esta investigación no hubiera sido posible.

Figuras

Figura 1: Porcentaje de documentos indizados

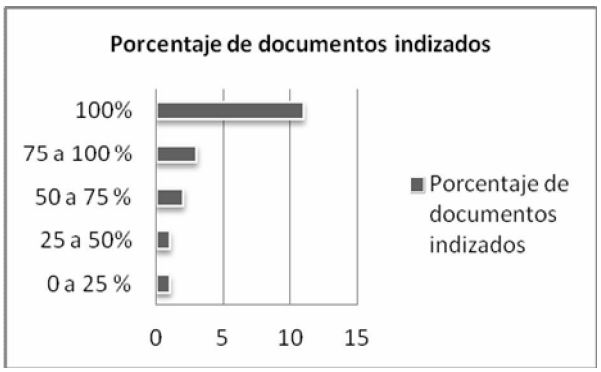

Figura 2: Tipo de vocabulario utilizado

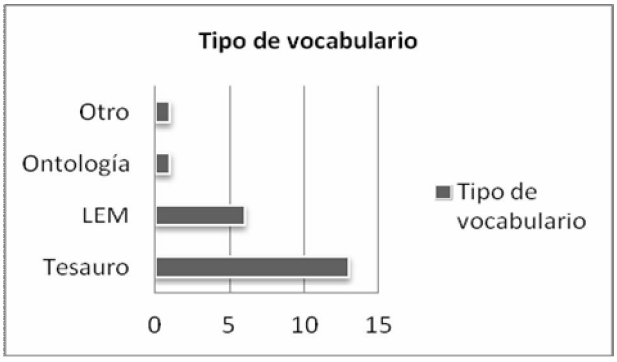

Figura 3: Vocabularios consolidados vs. desarrollos ad-hoc

Figura 4: Lugar de extracción de las palabras del texto

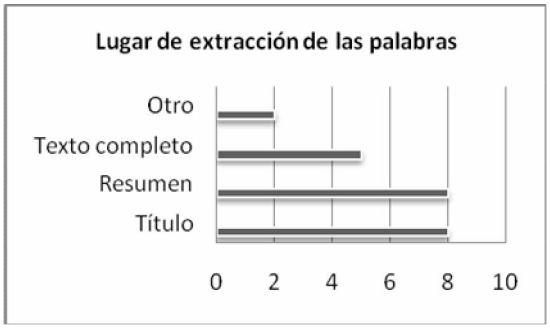

Figura 5: Algoritmo de extracción de palabras utilizado

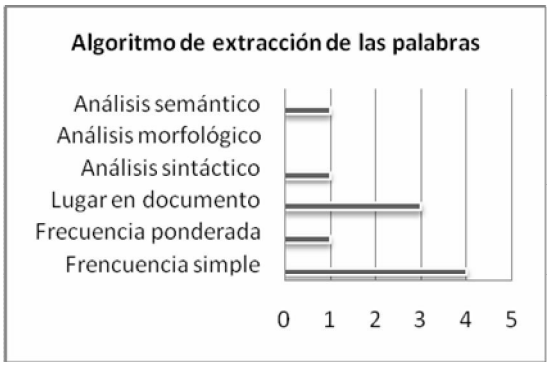

Figura 6: Almacenamiento para recuperación por cadena de caracteres e indización automática

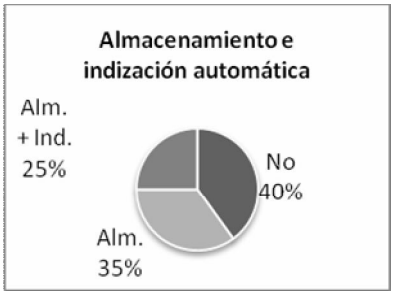

Figura 7: asignación de etiquetas no controladas

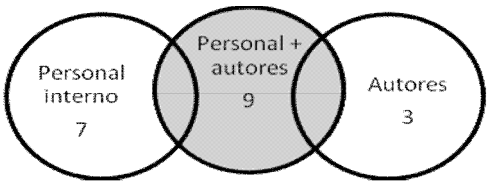

Figura 8: Modelos utilizados

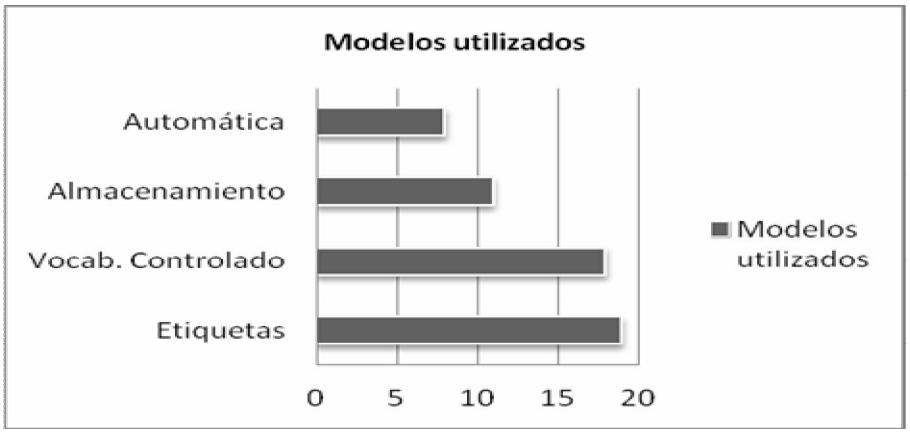

Figura 9: Modelos utilizados

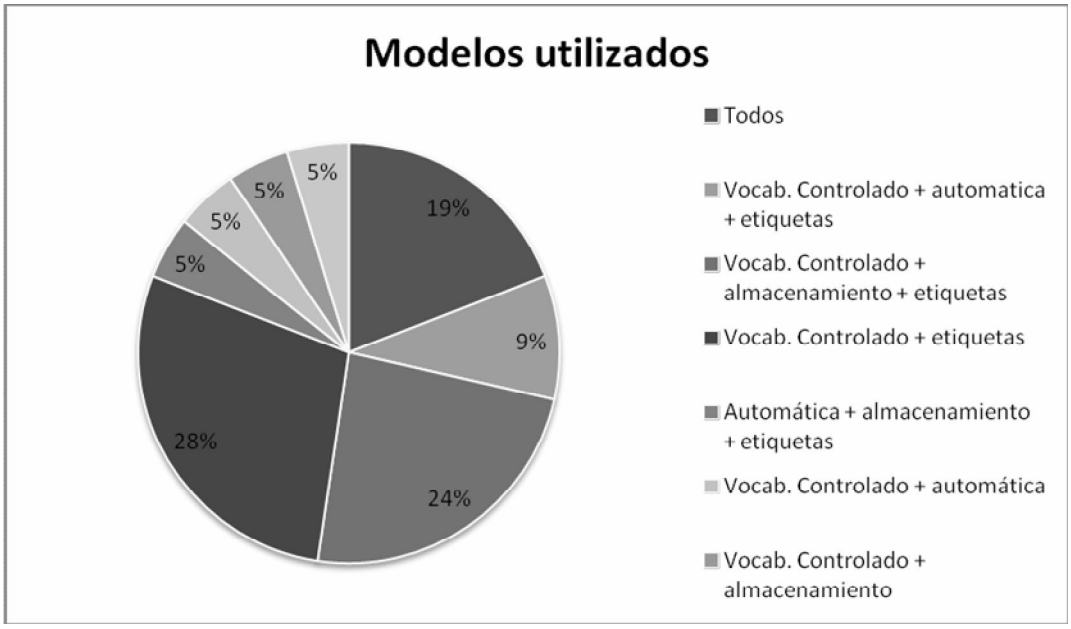

Figura 10: Anexo

Institución	Nombre
Centro Argentino de Información Científica y Tecnológica (CAICyT), CONICET	Scielo Argentina

Centro Atómico Bariloche e Instituto Balseiro	RICABIB
Consejo Latinoamericano de Ciencias Sociales (CLACSO)	Red de Bibliotecas Virtuales
E-Lis	E-Lis Argentina
Facultad de Agronomía, Universidad de Buenos Aires	FAUBA
Facultad de Ciencias Económicas y sociales, Universidad Nacional de Mar del Plata	Nülan
Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires	Biblioteca Digital
Facultad de Humanidades y Ciencias de la Educación, Universidad Nacional de la Plata	Memoria Académica
Facultad de Ciencias Naturales y Museo, Universidad Nacional de La Plata	Naturalis
Facultad de Psicología, Universidad Nacional de Mar del Plata	Rpsico
Instituto Nacional de Investigación y Desarrollo Pesquero	Osean Docs
Instituto Nacional de Tecnología Agropecuaria	Pro Huerta
Ministerio de Educación de la Nación	Repositorio Institucional
Universidad Católica Argentina (UCA)	Biblioteca Digital
Universidad de Ciencias Empresariales y Sociales (UCES)	Biblioteca Digital
Universidad Nacional de Córdoba	Repositorio Digital Universitario
Universidad Nacional de La Plata	SeDiCi
Universidad Nacional de Rosario	Repositorio Hipermedial
Universidad Nacional de Salta	Repositorio Digital
Universidad Nacional del Litoral	Biblioteca Virtual
Universidad Nacional Del Sur	Biblioteca Digital Académica
Universidad San Andres	Repositorio Digital