UNIVERSIDAD NACIONAL DE LA PLATA
FACULTAD DE INFORMÁTICA
SECRETARÍA DE POSGRADO

TESIS DOCTORAL

Una metodología de evaluación de repositorios digitales para asegurar la preservación en el tiempo y el acceso a los contenidos

ING. MARISA RAQUEL DE GIUSTI

Directora de tesis: DRA. SILVIA GORDILLO

Tesis presentada para obtener el grado de Doctor en Ciencias Informáticas

La Plata, 2014

De Giusti, Marisa Raquel
Una metodología de evaluación de repositorios digitales para asegurar la preservación en el tiempo y el acceso a los contenidos. - 1a ed. - La Plata : Universidad Nacional de La Plata, 2014.
E-Book.

ISBN 978-950-34-1210-7

1. Informática. 2. Educación Superior. I. Título
CDD 005.3

Fecha de catalogación: 12/05/2015

Dedicatoria

Dedicado a mi familia que siempre me ha acompañado con su confianza y con un recuerdo grandote para mis papás, ya ausentes, pero presentes en el recuerdo, quienes siempre nos guiaron para ser perseverantes en el esfuerzo y a enfrentar la vida con fe y optimismo.

Agradecimientos

A mi directora, Silvia, por aceptar este rol y apoyarme para que lograra cumplir el trabajo.

A mis compañeros de trabajo y amigos: Gonzalo Villarreal y Ariel Lira, por su orientación cotidiana, sus consejos y sus enseñanzas en informática; a Agustín Terruzzi, Paula Lacunza y Facundo Adorno, por las herramientas adicionales de software vinculadas al caso de estudio de la tesis; a Lucas Folegotto, por su ayuda paciente con la gráfica; a Gonzalo nuevamente por la cuidadosa lectura y los comentarios; a Carlos Nusch, hoy día parte de mi familia, por la dedicada revisión y la normalización de la tesis, y a Analía Pinto, amiga y editora, por sus meticulosas correcciones sintácticas y por dejarla tan bonita!

Por extensión, a todo el grupo de trabajo de PREBI-SEDICI, con los que comparto, con orgullo, buena parte del cada día y a la Facultad de Ingeniería de la UNLP, por ofrecerme el espacio físico donde he desarrollado siempre mis actividades y donde permanezco desde que comencé mis estudios como ingeniero.

Al personal de la oficina de Posgrado de la Facultad de Informática, por su colaboración constante y la excelente atención y predisposición de las personas que allí trabajan.

A la Universidad Nacional de la Plata y a la Comisión de Investigaciones Científicas de la provincia de Buenos Aires por brindarme el apoyo económico que me permitió realizar este doctorado.

Advertencia: todas las citas en inglés utilizadas en este trabajo han sido traducidas por la autora y en todos los casos figura a pie de página la versión en su idioma original. Tal proceder es considerado por la Facultad de Informática de la UNLP como la modalidad más adecuada para facilitar la lectura de la tesis.

Sé que pueden quemar libros, arrasar bibliotecas,
prohibir lenguas, desterrar creencias,
borrar pasados, dibujar presentes, ordenar futuros,
torturar y ejecutar personas...
Pero también sé que aún
no han descubierto cómo matar
el cuerpo intangible y luminoso de una idea,
de un sueño o de una esperanza.

EDGARDO CIVALLERO Y SARA PLAZA

ÍNDICE

ÍNDICE DE FIGURAS

RESUMEN

Un repositorio institucional es un depósito de documentos digitales, cuyo propósito es gestionar, organizar, almacenar, preservar y difundir en acceso abierto la producción resultante de la actividades de una organización. La variedad de materiales que se alojará en un repositorio institucional dependerá de la política de contenidos que determine la propia institución; los contenidos, en principio, podrían mantenerse a perpetuidad y el repositorio ser implementado de modo tal de asegurarlo, pero este punto también dependerá de la política de preservación que la institución determine.

El objetivo central de esta tesis es, entonces, proponer una metodología de evaluación para repositorios institucionales. Con esto se busca mejorar la calidad de los repositorios, así como la estandarización, ayudar a la interoperabilidad y obtener una mayor visibilidad de las producciones que una institución, en este caso educativa, guarda en un repositorio.

Entre los objetivos específicos está el de asegurar la preservación de los contenidos del repositorio, de modo que siempre sea posible acceder a ellos y que éstos resulten legibles tanto para usuarios humanos como máquinas. Para lograrlo es necesario conocer el campo de actividad de estos repositorios, enmarcado por la Iniciativa de Acceso Abierto que definió sus alcances y funciones, y elaborar una correcta definición de ellos, para responder a las preguntas fundamentales: ¿qué es un repositorio? y ¿qué estructura y funciones lo caracterizan mejor?

Con miras a responder tales interrogantes, se relevaron los modelos que a lo largo del tiempo han servido para representar un repositorio digital, elaborando una mirada crítica en cuanto a la utilidad de cada uno de ellos, y observando cuánto de las estructuras y funciones propuestas permanecen en los repositorios actuales. Se analizaron sus similitudes y diferencias, para identificar el modelo que mejor se ajustaba y para determinar la necesidad de contar con más de un modelo que representase el repositorio. Una vez elegido el modelo, se determinaron cuáles serían los parámetros de evaluación que interesaban a los objetivos planteados.

Como objeto de estudio y experimentación se seleccionó el repositorio institucional

central de la Universidad Nacional de La Plata —el Servicio de Difusión de la Creación Intelectual (SEDICI)—, anticipando que las conclusiones extraídas, en cuanto a líneas de acción para cumplir con los objetivos previstos, podrían ser extensibles a otros repositorios institucionales. Se realizó el relevamiento del estado de los objetos digitales del repositorio, para luego determinar las acciones a realizar, proponer cambios y delinear un plan a largo plazo vinculado al planeamiento de la preservación de los contenidos de modo de asegurar que los contenidos siempre estén disponibles y en una condición tal que permita la legibilidad por parte de los usuarios. Luego de cada análisis, se extrajeron algunas conclusiones y reflexiones breves. La tesis culmina con la exposición de las conclusiones generales y con los trabajos proyectados para el futuro, vinculados principalmente a la selección y migración a formatos más apropiados para la preservación, a la generación de nuevas tareas de validación de metadatos asociados a los contenidos y a la realización de un plan integral de preservación.

Palabras clave: *evaluación; repositorios institucionales; preservación; visibilidad; acceso abierto; metadatos.*

Capítulo 1 | El movimiento de Acceso Abierto: historia, motivaciones y estrategias

«Invitamos a gobiernos, universidades, bibliotecas, editores, publicistas, fundaciones, sociedades académicas, asociaciones profesionales, estudiosos y científicos que comparten nuestros puntos de vista, a que se sumen a la tarea de eliminar los obstáculos al acceso abierto, y a construir un futuro en el que, en todo el mundo, la investigación y la educación puedan desarrollarse con total libertad.»

DECLARACIÓN DE BUDAPEST, HUNGRÍA. 14 de febrero de 2004.

Síntesis: Este capítulo esboza una cronología del acceso abierto a través de sus principales hitos y vías de acción: las revistas de acceso abierto, los repositorios digitales de acceso abierto y las características de sus implementaciones. Este aspecto es muy relevante, puesto que permite distinguir qué diferencia al objeto de estudio "repositorio institucional", de cualquier otro tipo de base de datos o depósito de documentos digitales. Lo que aquí se expone sirve para advertir de la complejidad que entraña el manejo de contenidos digitales en el entorno de los Repositorios Institucionales (RI), debido al compromiso de mantener accesibles las obras en el tiempo frente a los constantes cambios tecnológicos. El capítulo culmina con un comentario acerca del rol de los repositorios en la investigación sobre el impacto de las obras y su compromiso para permitir el acceso y facilitar su uso; los cambios que el movimiento ha conseguido se extienden desde los circuitos tradicionales de publicación científica hasta los derechos de autor.

El movimiento de Acceso Abierto (AA)

Peter Suber, uno de los principales referentes del Acceso Abierto (AA), señala el año 1966 como el inicio del movimiento (Suber, 2009), en virtud de dos hechos fundacionales: el lanzamiento de ERIC (Educational Resources Information Center) por el Departamento de Educación de Estados Unidos y el lanzamiento de Medline (disponible en la web recién en 1997) por la Biblioteca Nacional de Medicina de ese mismo país. Como puede observarse por las fechas mencionadas, los inicios del movimiento resultan muy anteriores al devenir de Internet y de las TIC que,

naturalmente, lo han potenciado de manera significativa.

Hasta la década de 1990, la historia recoge numerosos hitos, de los cuales pueden citarse, sobre todo por su relevancia, el lanzamiento (en agosto de 1991) de ArXiv, definido entonces como un sistema de distribución automática de artículos de investigación, sin las operaciones editoriales asociadas a la revisión por pares. ArXiv cubre hoy día los campos de la física, las matemáticas, la biología, las finanzas y las ciencias de la computación. Los artículos son depositados por los autores antes de remitirlos a las revistas especializadas para la revisión por pares, aunque en la actualidad se agregan artículos que han sido publicados en revistas que no retienen los derechos de manera exclusiva. El modelo de ArXiv ha sido discutido desde entonces y su viabilidad para otros campos del saber ha sido puesta en cuestión.

El año 1993 es trascendental para el movimiento de AA, puesto que excede el marco de estas iniciativas: la Organización Europea para la Investigación Nuclear (más conocida por su sigla, CERN) anuncia la posibilidad del uso libre y sin cargo de la tecnología WWW. Esta institución jugó un papel fundamental en el entramado de apoyo al movimiento en muchos aspectos, poniendo a disposición software libre, cumpliendo tareas de repositorio con sus propios servidores para los artículos de investigación y alojando las iniciativas europeas en el ámbito.

En 1994, Stevan Harnad, unos de los líderes más activos del movimiento, lanza la iniciativa por el autoarchivo. Básicamente, recoge la iniciativa ArXiv y la desarrolla para su aplicación en otros campos, ya no sólo para preprints sino para trabajos con revisión (Harnad, 2001). Su trabajo consistió en analizar el impacto que generaba en la comunidad científica el depósito de los artículos científicos en un archivo de acceso abierto (en aquel entonces, un sitio FTP). En su momento suscitó una gran discusión sobre todo el sistema de comunicación científica, que aún hoy continúa teniendo enorme vigencia.

En 1997, la Asociación de Bibliotecas de Investigación de los Estados Unidos (ARL, por sus siglas en inglés) pone en funcionamiento la iniciativa denominada Scholarly Publishing & Academic Resources Coalition (SPARC), una alianza internacional que comienza a trabajar para corregir el desequilibrio del sistema de edición científica y que va a transformarse en un catalizador de fuertes cambios. En ese mismo año, se lanza CogPrints, el primer depósito de artículos de investigación en las áreas de

psicología, neurociencias, lingüística, filosofía y ciencias de la computación. Como se dijo, comienza también el acceso libre a Medline a través de PubMed[1], gracias al lanzamiento de la iniciativa de los decanos en Estados Unidos, que aboga por el acceso libre a los resultados de la investigación científica en todos los campos.

A partir de 1998, el curso de los acontecimientos para el AA se acelera. Los consejos editoriales de algunas revistas científicas rompen con sus casas editoriales, por las serias divergencias en cuanto a la visibilidad de las revistas a través de la red. En torno a estos movimientos, la recién creada SPARC lanza una propuesta, llamada "Declaración de Independencia", concebida para asistir a la comunidad académica en la creación de revistas controladas por los mismos académicos (Suber, 2001-2008). También se debe destacar, no sólo por su importancia sino por ser la irrupción del mundo hispano en el movimiento, la Declaración de San José (Costa Rica), por los delegados del Sistema de Información en Ciencias de la Salud de Latinoamérica y el Caribe (BIREME). Precisamente, esta institución es la fundadora del PubMed hispano, denominado SciELO (por su nombre en inglés, Scientific Electronic Library On Line).

El surgimiento de diferentes acervos a partir de las distintas iniciativas trajo consigo dificultades de interoperabilidad y algunos inconvenientes surgidos de la necesidad de realizar búsquedas en más de un acervo; todo ello fomentó una ya famosa reunión en Santa Fe (Albuquerque, Estados Unidos) en 1999, que estableció la iniciativa Open Archives Initiative (OAI), destinada a delinear una serie de principios organizativos y especificaciones técnicas para permitir que los diversos sistemas de archivo y publicación fueran interoperables. La iniciativa OAI llevó a la aparición del OAI-PMH (Protocol of Metadata Harvesting) para facilitar el intercambio de los registros entre los acervos.

Un resultado adicional de la convención de Santa Fe fue la propuesta de desarrollo de software para facilitar la puesta en marcha de los repositorios, a la vez que se enunciaron las características o funcionalidades que debía reunir dicho software:

- un mecanismo de depósito;
- un sistema de almacenamiento a largo plazo;

[1] PubMed es un motor de búsqueda de libre acceso a la base de datos MEDLINE de citaciones y resúmenes de artículos de investigación biomédica, ofrecido por la Biblioteca Nacional de Medicina de los Estados Unidos.

- una política con respecto a la presentación de documentos y su conservación;
- una interfaz simple que permitiera a terceros recopilar registros de recursos provenientes de distintas fuentes (OAI-PMH).

En el año 2000, se crea un archivo central de literatura biomédica similar a PubMed, que se plasma con la creación, por parte de la Biblioteca Nacional de los Estados Unidos, de PubMed Central. También debe destacarse el lanzamiento de la primera iniciativa de un editor privado: BioMedCentral, que actualmente practica la "vía verde", es decir, que permite el depósito en repositorios del preprint, el postprint y el artículo de la revista.

A su vez, en el mismo año, varios científicos involucrados en el desarrollo de PubMed Central, fundaron un grupo llamado Public Library of Science (PLOS), que hizo circular una carta abierta en la que se le exigía un vuelco al sistema de comunicación científica. La carta planteaba que el registro de las ideas no debía ser controlado por los editores, sino pertenecer al dominio público y estar accesible a través de Internet. Los firmantes advertían que estaban dispuestos a dejar de publicar y arbitrar para las editoriales a menos que en septiembre del 2001 se comenzaran a hacer disponibles sus contenidos (luego de 6 meses de publicados) en PubMed Central u otro sitio similar. Esta carta fue firmada por más de 30.000 científicos de todo el mundo, pero la respuesta más contundente fue la de la propia PLOS, que se convirtió en una editorial de acceso abierto y lanzó sus dos primeras revistas en dicho sistema: *PLOS Medicine* y *PLOS Biology*. Hoy día, PLOS mantiene seis publicaciones periódicas y las tasas de publicación de los artículos son costeadas por los autores, por las instituciones que los albergan, o por los sponsors con los que cuentan para que, de este modo, los artículos estén libres para los lectores. También en el 2000, la Universidad de Southampton lanza EPrints, un sistema de publicación y depósito de archivos digitales, de código abierto y libre.

En diciembre del 2001, el Open Society Institute organizó una reunión en Budapest (Hungría), cuyo resultado fue la Iniciativa de Acceso Abierto de Budapest (Budapest Open Access Initiative, BOAI), que formalizó, en su declaración del 14 de febrero de 2002, los presupuestos del movimiento de acceso abierto. En la declaración se recomiendan las modalidades de publicación en revistas de acceso abierto, o bien a

través del autoarchivo en repositorios abiertos; en todos los casos, se promueve la disponibilidad gratuita en Internet, para que cualquier usuario la pueda leer, descargar, copiar, distribuir o imprimir, así como bucear dentro del artículo sin otras barreras financieras, legales o técnicas que las de acceso a la red, con especial atención a que *"la única función del copyright en este dominio, no puede ser otra que dar a los autores control sobre la integridad de su trabajo y el derecho a ser apropiadamente acreditados y citados"*[2] (BOAI, 2002). A la declaración de Budapest le siguieron las declaraciones de Bethesda y Berlín (Barrionuevo, 2009), apoyando la publicación de acceso abierto.

El movimiento de Acceso Abierto: estrategias

El movimiento de Acceso Abierto a la información se basa en dos estrategias fundamentales para garantizar el acceso y la diseminación sin restricciones económicas y legales de la información científico-técnica:

1. Las revistas de acceso abierto, la llamada "ruta o vía dorada".
2. Los repositorios de acceso abierto, la llamada "ruta o vía verde".

La figura 1.1 muestra las rutas de publicación posibles de acuerdo a estos preceptos.

[2] Texto en inglés: *"... and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited".*

Figura 1.1: Rutas de publicación posibles según la decisión del autor

Fuente: De Giusti, Marisa. (2013) "Bibliotecas y repositorios digitales: tecnología y aplicaciones". Curso de posgrado, Facultad de Informática (UNLP). Fuente original: Max Planck Society.

El éxito del movimiento de Acceso Abierto ha dado como resultado una creciente cantidad de repositorios de diferente índole que pueden agruparse de acuerdo a varios criterios:

- **Repositorios Institucionales**: son los que reúnen la documentación generada por una institución en el desarrollo de su actividad, principalmente las publicaciones de su personal: artículos, tesis, material docente, congresos. Ejemplos: Digital CSIC, SEDICI.
- **Repositorios Temáticos:** reúnen documentos altamente especializados, agrupados por materia o área temática. Ejemplos: RePEc (ciencias económicas), PubMed Central (ciencias de la salud), ArXiv (ciencias exactas).
- **Agregadores o Recolectores:** agrupan múltiples repositorios, lo que permite consultarlos simultáneamente mediante un único formulario de búsqueda. Ejemplos: La Referencia (Red Federada de Repositorios Institucionales de Publicaciones Científicas), NDLTD (Networked Digital Library of Theses and Dissertations).

- **Repositorios de Datos Científicos:** almacenan y preservan los datos científicos generados durante la investigación.
- **Repositorios Huérfanos:** agrupan las obras de los autores que no tienen un depósito en su institución o que no tienen afiliación.

Los repositorios también suelen clasificarse según los contenidos que alojan, que pueden ser muy diversos: artículos, tesis, materiales didácticos, fotografías, objetos de aprendizaje y otros. Hay, por ejemplo, repositorios que alojan revistas y que están implementados en desarrollos que originalmente no estaban pensados como tales. ROAR (Registry of Open Access Repositories) reporta 117 repositorios de revistas con un panorama de implementaciones diverso, según puede observarse en la figura 1.2.

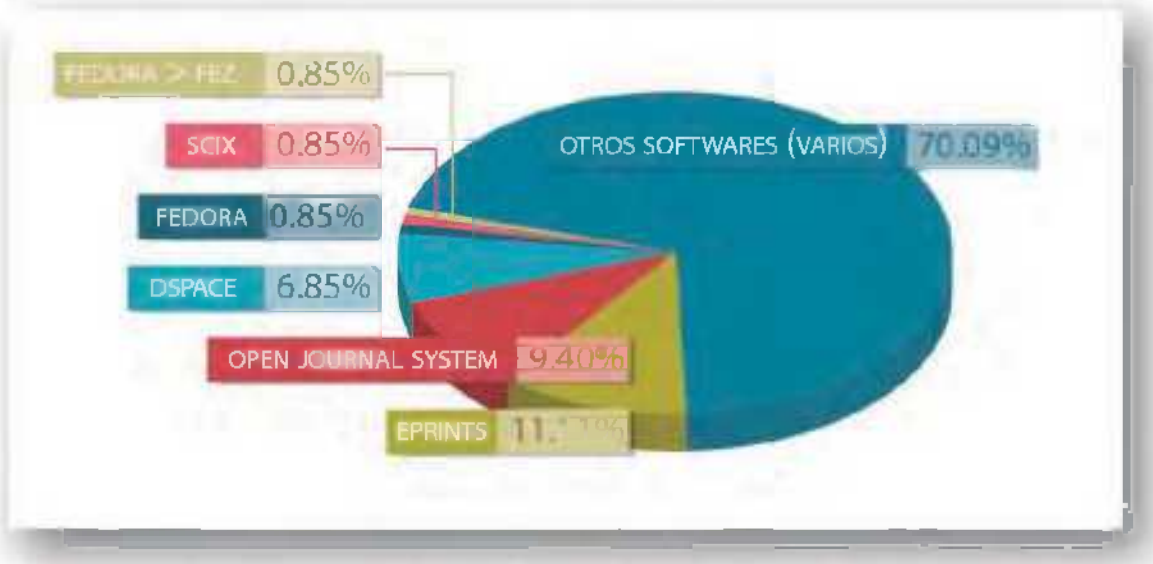


Figura 1.2: Softwares utilizados para la implementación de repositorios de revistas

Fuente: ROAR.

Es interesante hacer notar que ROAR muestra que más del 9% de los repositorios de revistas están implementados en Open Journal System (OJS), un excelente desarrollo para la gestión del circuito completo de una revista académica, pero que no cuenta, hasta el momento, con herramientas adecuadas para la preservación.

Las clasificaciones precedentes pueden ser un tanto arbitrarias. No obstante, sirven para tener en cuenta qué actividades y roles debe cumplir un repositorio para ser considerado como tal; en este punto, las dudas vinculadas a lo tecnológico pueden

condicionar la clasificación. En el capítulo 2 se profundizarán las definiciones de bibliotecas y repositorios digitales.

El poblamiento de los repositorios depende en gran medida de que los propios autores depositen sus trabajos (autoarchivo) y se ven muy favorecidos cuando existen mandatos institucionales que regulen la obligatoriedad de este proceso. La Universidad Nacional de La Plata, por ejemplo, a través de la resolución 78/11, estableció que las tesis de maestría y doctorado deben ser preservadas en formato digital, a través del Servicio de Difusión de la Creación Intelectual (SEDICI), y que el depósito debe ser un requisito obligatorio para los trámites del título pertinente, desde febrero de 2011 en adelante. En Argentina, la Ley 26.899 de Creación de Repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos fuerza a las instituciones educativas que reciben financiamiento del estado nacional a crear repositorios de acceso abierto:

> *"(...) en los que se depositará la producción científico-tecnológica resultante del trabajo, formación y/o proyectos, financiados total o parcialmente con fondos públicos, de sus investigadores, tecnólogos, docentes, becarios de posdoctorado y estudiantes de maestría y doctorado. Esta producción científico-tecnológica abarcará al conjunto de documentos (artículos de revistas, trabajos técnico-científicos, tesis académicas, entre otros), que sean resultado de la realización de actividades de investigación."*
> (Boletín Oficial de la República Argentina, 2013)

Esta ley ha impulsado grandemente la creación y el poblamiento de los repositorios de acceso abierto ya que como lo expresa su artículo 5:

> *"Los investigadores, tecnólogos, docentes, becarios de posdoctorado y estudiantes de maestría y doctorado cuya actividad de investigación sea financiada con fondos públicos, deberán depositar o autorizar expresamente el depósito de una copia de la versión final de su producción científico-tecnológica publicada o aceptada para publicación y/o que haya atravesado un proceso de aprobación por una autoridad competente o con jurisdicción en*

la materia, en los repositorios digitales de acceso abierto de sus instituciones, en un plazo no mayor a los seis (6) meses desde la fecha de su publicación oficial o de su aprobación. Los datos primarios de investigación deberán depositarse en repositorios o archivos institucionales digitales propios o compartidos y estar disponibles públicamente en un plazo no mayor a cinco (5) años del momento de su recolección, de acuerdo a las políticas establecidas por las instituciones." (Ibídem).

Las revistas de acceso abierto son revistas cuyos contenidos están disponibles libre y gratuitamente en Internet. Esta posibilidad de acceso sin restricciones está sustentada por el pago, por parte del autor, para que la obra sea accedida libremente por los lectores como se mencionó respecto de las revistas de la editorial PLOS. Se las denomina "genuinas" cuando disponen de todos sus artículos en libre acceso, e "híbridas" cuando proporcionan sólo algunos en forma libre.

Resulta importante, en este sentido, que los autores revisen las nociones respecto de derechos de autor para resguardar apropiadamente sus trabajos y esto se presenta como un fundamento más para la creación y mantenimiento de los repositorios institucionales. Con ellos, la institución tiene la posibilidad de contar con las obras de sus miembros, en un único lugar y, fundamentalmente, contar con datos fidedignos para la elaboración de estadísticas a las que antes sólo tenían acceso las empresas editoriales.

Los repositorios o archivos de acceso abierto constituyen una de las estrategias más viables para garantizar el acceso abierto a la información, como ya queda dicho. Si bien existen repositorios que sólo incluyen artículos arbitrados (preprints y postprints), otros, sobre todo los denominados repositorios institucionales, permiten además la presencia de ponencias de eventos, informes de investigación, conferencias, presentaciones de seminarios, tesis y hasta ordenanzas de una institución.

Es preciso hacer notar que el autoarchivo en los repositorios no se considera un sustituto de la publicación formal en una revista (sea de acceso abierto o por suscripción), sino una vía complementaria para garantizar la máxima visibilidad del trabajo científico, conservando sus derechos para publicarla por otras vías y aceptando que el repositorio preserve la obra y la difunda. Los repositorios usan de manera

habitual licencias de uso y las más frecuentemente utilizadas son las licencias Creative Commons (CC). Esta organización ofrece una variedad de licencias, las cuales pueden ser expuestas para que los autores elijan qué usos quieren que se dé a su obra en un repositorio; en todos los casos, la mención del nombre del autor es obligatoria.

Factores que empujaron el Acceso Abierto

Desde hace tres décadas, consultores, científicos, bibliotecarios y editores han señalado que el sistema tradicional de comunicación científica se encuentra en crisis, ya que no se cumplen los objetivos primarios de éste: favorecer la diseminación y el intercambio de los resultados científicos para avanzar y obtener un mayor progreso científico, técnico y social para toda la sociedad. Diversos y complejos son los factores que condicionan esta crisis. Entre ellos, se señalan como los más importantes el incremento sostenido de los precios de las revistas científicas, sobre todo en las áreas de ciencia, tecnología y medicina, suceso denominado por la literatura especializada, *"serial crisis"*.

Un segundo problema, que impacta grandemente en las áreas científico-técnicas de mayor movilidad, es la extensión del período de tiempo entre el envío de un artículo a una revista y su publicación definitiva. Algunos países han tomado recaudos y cambiado las reglas de sus sistemas de premios y promociones para reducir esa brecha; podría citarse aquí el caso de España, país que, en el área de informática, reconoce un conjunto de congresos, y por tanto las presentaciones en ellos, con una valoración superior incluso a la obtenida por el investigador, tras la publicación de un artículo en una revista arbitrada.

A los aspectos antes señalados, se suma la escalada de fusiones y adquisiciones de empresas editoriales —las más pequeñas desaparecen en manos de las más grandes—, por lo que se establece un mercado sin competencia. Otros aspectos, que pueden señalarse como síntomas de la crisis del sistema, son las crecientes restricciones que establecen las legislaciones actuales de derecho de autor sobre el acceso y la diseminación de la información científica, que han desvirtuado los objetivos primarios de la comunicación científica y del propio derecho de autor, del mismo modo que aquellos relativos al sistema de recompensa científica, enfocado más a la publicación

en revistas "de impacto" que a la amplia diseminación de los resultados científicos.

En estos últimos años ha ido creciendo el reconocimiento de que casi toda la investigación académica se financia con fondos públicos, por lo que entonces es razonable que se exija maximizar la diseminación de los resultados a toda la sociedad y de ahí que la necesidad de existan repositorios que posibiliten el acceso abierto sea cada vez mayor.

En este sentido, las potencialidades de las tecnologías de la información y la comunicación (TIC), cuyo exponente máximo es Internet, han facilitado la creación de revistas electrónicas y otras plataformas que tienen el potencial de permitir un acceso más amplio a la información.

La sinergia entre los diversos aspectos antes señalados ha contribuido a fortalecer toda una corriente de pensamiento y acción transdisciplinaria e internacional a favor de la ampliación del acceso a la información científica sin barreras económicas ni legales. Este movimiento constituye una alternativa válida para solucionar las restricciones en el acceso a la literatura científica.

La crisis de las revistas ha impactado con mayor fuerza en las publicaciones vinculadas a la ciencia, medicina y tecnología, como se dijo, en las cuales, además, resulta esencial el acceso inmediato a los nuevos conocimientos. Es por ello que la mayoría de las iniciativas relacionadas con el movimiento de acceso abierto comenzaron en estos campos. De acuerdo con el índice de precios de algunas de las más importantes empresas editoriales del mundo, el costo promedio de una revista en ciencia, tecnología y medicina aumentó entre un 200 y un 300% entre 1990 y 2010 (Kyrillidou, Bland, 2009).

La situación precedente crea un gran obstáculo en los países en vías de desarrollo, especialmente en todo lo que hace al acceso a la información científica y su posterior uso y beneficio en aras de la sociedad, sobre todo si se piensa en la "ecuación" precedente vinculada a los países con PIB escaso y en áreas críticas como las vinculadas a la salud.

Sobre los derechos de autor y el sistema científico

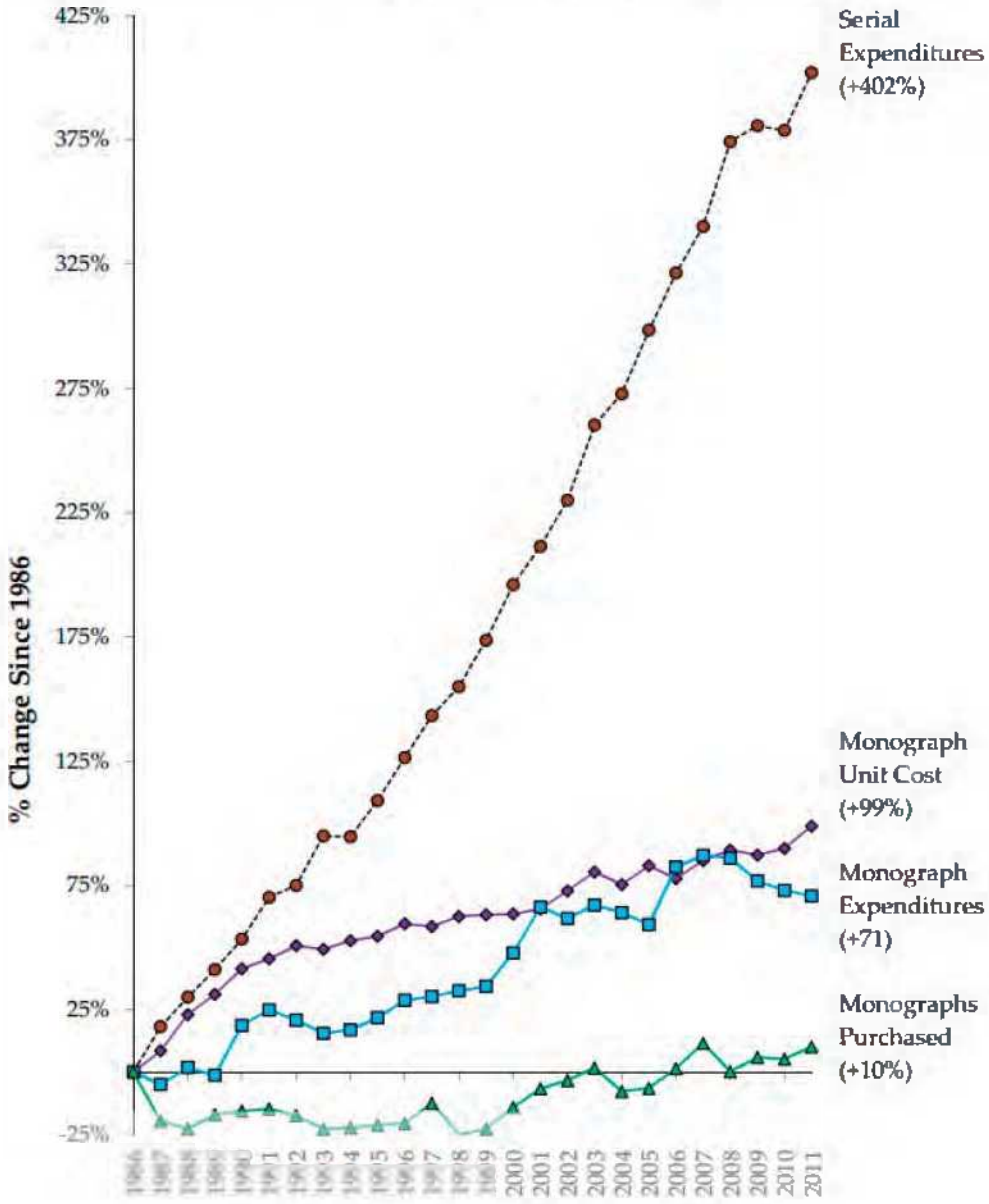
Las legislaciones de derecho de autor se crearon con la intención de proteger al

titular de los derechos patrimoniales de una obra (*copyright*) contra los usos indebidos que terceros pudieran hacer de éstas. Sin embargo, en el caso de las revistas científicas se da la paradoja de que, en un gran número de casos y ya de manera "rutinaria", todos los derechos patrimoniales pasan directamente a manos de los editores, por lo que el autor pierde el control sobre el uso posterior de su trabajo publicado. Cualquiera sea el uso que quiera darle a su creación, como base para otros estudios, brindarlo a estudiantes, o ponerlo en el sitio web de su biblioteca o institución, si la cesión ha sido firmada con carácter de exclusividad, corre el riesgo de infringir los acuerdos firmados con el editor. Vale preguntarse qué porcentaje de los editores solicitan hoy día la cesión exclusiva del derecho de autor, qué derechos retiene el autor a usar su propio trabajo, o, en el caso particular de los repositorios: ¿qué derecho tiene un autor a hacer depósito en un repositorio institucional de la propia institución que ha apoyado económicamente su trabajo?

Hasta hace un tiempo, esto no constituía una preocupación para los investigadores, acostumbrados a ceder su trabajo y no obtener ningún beneficio a cambio; hoy día, el advenimiento de las TIC, y especialmente Internet, que habilitan una mayor difusión, tienen, como contracara, restricciones cada vez mayores, impuestas por las legislaciones de derecho de autor. Un ejemplo más que curioso es el que cita Sánchez Tarragó (2007): en Estados Unidos, estas leyes limitan el "uso justo" institucional a sólo cinco artículos publicados en los últimos cinco años de cualquier revista. Una vez que ese límite es alcanzado, cualquier artículo adicional debe pagarse al editor, sea por concepto de préstamo interbibliotecario o por distribución de documentos. Otro detalle que menciona la autora cubana es que existe una práctica común por parte de los editores de prohibir el uso de suscripciones electrónicas para préstamo interbibliotecario, dado que las bibliotecas van aumentando las suscripciones a las publicaciones en desmedro de las en papel, y así la disponibilidad va descendiendo.

Resulta interesante pensar cómo decae entonces el acervo de las bibliotecas, afectado tanto por el costo de las publicaciones periódicas y algunas prácticas adversas de las editoriales, como la que se cita en el párrafo precedente. Como ilustración de lo anterior, cabe observar, en la figura 1.3, los gráficos incluidos a continuación de la Association of Research Libraries (ARL).

Monograph & Serial Costs in ARL Libraries, 1986-2011*



NOTE: Data for monograph and serials expenditures was not collected in 2011-12.

Source: *ARL Statistics 2010-11* Association of Research Libraries, Washington, D.C.
*Includes electronic resources from 1999-2011.

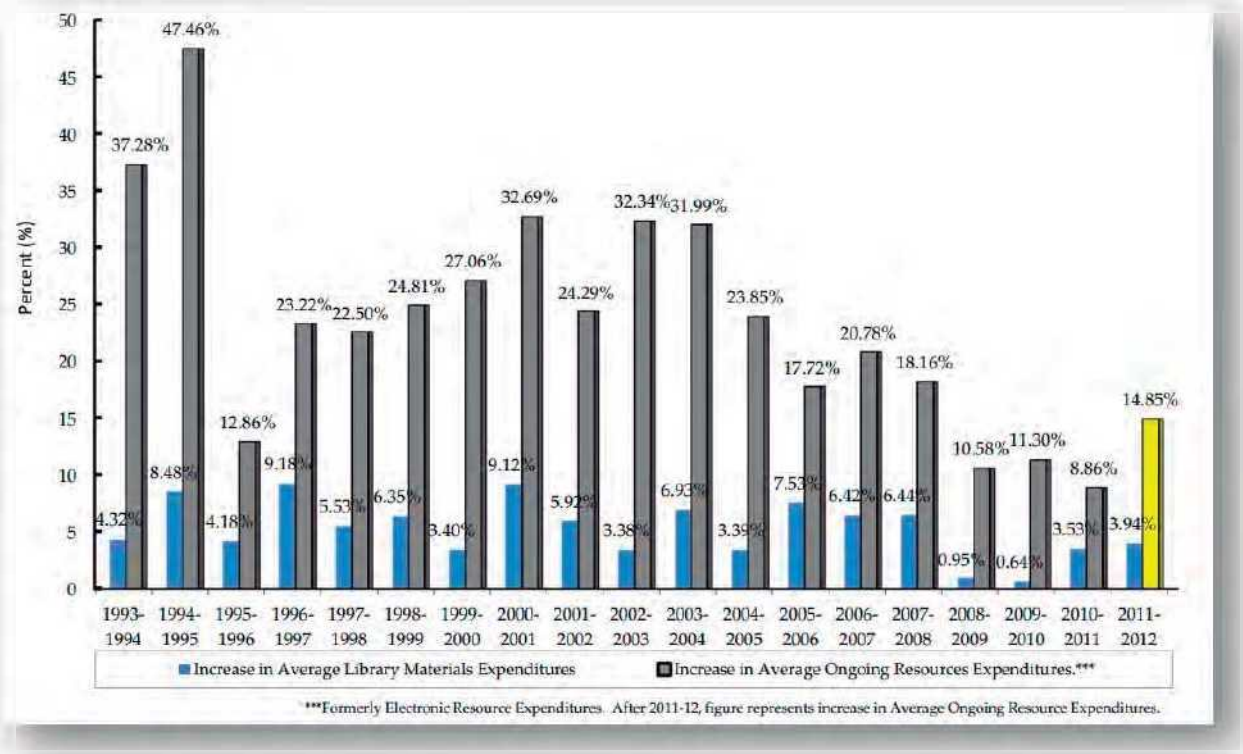


Figura 1.3: Gráficos de la ARL en los que se muestran, en primer término, el costo de trabajos monográficos y publicaciones periódicas en la ARL, para el período 1986-2011, y en segundo término, los gastos en recursos frente a los gastos en materiales

Fuente: ARL Statistics 2010-2011.

El grupo de trabajo SHERPA (Securing a Hybrid Environment for Research Preservation and Access) mantiene el proyecto RoMEO (Rights Metadata for Open Archiving), que analiza los términos de *copyright* de las editoriales y el autoarchivo. Según la política de cada editorial, RoMEO hace una clasificación por colores: verde (la editorial autoriza el depósito de la versión pre o postprint), azul (permite autoarchivo de postprint), amarillo (autoarchivo de preprint) y blanco (la editorial no permite el depósito posterior del artículo de ninguna forma). Actualmente, de la lista de 1546 editoriales registradas en su base de datos, más del 72% permiten algún tipo de autoarchivo (SHERPA/RoMEO, 2014). Esto es una clara muestra de los cambios que ha empujado el movimiento de AA.

Es importante remarcar que esta situación ha ido variando a lo largo del tiempo y que algunos de los términos referidos al autoarchivo se escriben, en los contratos de edición, con letra pequeña y es necesario prestar atención sobre esto, ya que para depositar en abierto es relevante tener en cuenta dichos términos y reconsiderar cuidadosamente las revistas en las que se decide publicar. RoMEO facilita la búsqueda

por revista o por editorial e incluye, en aquellas que está disponible, el texto donde se define la política de la revista sobre los derechos de autor. En todos los casos, se aconseja consultar siempre la página de la propia revista.

Si bien la elección de la revista en la que publicar un artículo, en función de los términos de su política de derechos, no resulta un factor relevante para el autor, se debe prestar especial atención a los usos permitidos. Algunos trabajos dedicados al estudio y análisis de este problema (Hoorn, 2005; Hoorn, Van der Graaf, 2006), incluyen encuestas a autores y muestran que la cuestión de los derechos no despierta demasiado interés entre los investigadores a la hora de publicar; esto es un error, ya que permitir el acceso y la posibilidad de usar lo publicado es una responsabilidad inherente al quehacer académico.

Desde todo punto de vista, la discusión precedente no es ajena a las instituciones que albergan y mantienen a los investigadores y que ven dificultada la posibilidad de dar a conocer los logros obtenidos. Es alarmante que hoy en día los términos de las autorizaciones que las editoriales ponen para el uso de los trabajos para con las instituciones, se limiten a la cantidad de artículos sin costo adicional a utilizar por año, o la posibilidad de utilizar las suscripciones electrónicas para préstamo interbibliotecario (Albert, 2006).

Factor de impacto

El Factor de Impacto (IF, por sus siglas en inglés) de una publicación se ha usado desde los años 60 para medir el impacto de un trabajo científico y premiar (o castigar) a los investigadores involucrados. El uso del factor de impacto de una revista para fines de evaluación científica genera polémicas, ya que muchas veces se asume como representativo de todos los artículos que se publican en la revista y, por tanto, como una medida cuantitativa —y objetiva— de la calidad del resultado científico publicado. Más aún, de esta suposición se desprenden algunas prácticas perversas como las autocitas o las estrategias editoriales para subir el factor de impacto (Guiu, García-Ramos García, 2008). Evidentemente, los comportamientos de unos y otros resultan tergiversados y se pierden los objetivos prioritarios; es decir, a pesar del interés de la comunidad científica por hacer públicos los resultados de sus trabajos, hay un empuje decidido de uno y otro lado para difundirlos en determinadas publicaciones, que

garanticen unos resultados a expensas del acceso y con el agregado de unos costos altísimos como se vio precedentemente.

Todo esto conduce a por lo menos dos conclusiones: que existe un consenso con respecto a que el progreso del conocimiento es un proceso social y colectivo y que, por tanto, la investigación financiada con fondos públicos debe permanecer en el dominio público para producir cambios, y que las políticas de acceso abierto a los resultados de investigación descansan todavía en frágiles pilares, tanto legales como normativos. Por estas razones es que se debe crear una conciencia real al respecto, que comprometa a los involucrados directos, pero dicho proceso implica también gestar cambios radicales en las políticas de evaluación.

Difusión, visibilidad

El Open Citation Project (OpCit), que comenzó en octubre de 1999 y terminó a finales de 2002, tuvo como objetivo brindar una herramienta de enlace y análisis de citas para archivos abiertos, con el fin de desarrollar soluciones para cuantificar el impacto y las citas de referencia, y explorar las relaciones críticas entre las citaciones de los documentos en línea. Si bien el proyecto ya llegó a su fin, aún se puede acceder a algunos de sus servicios. Este proyecto trabajó, entre otras actividades, en el desarrollo de un índice de citas experimental para el acceso abierto a la literatura científica, basado en la recopilación de metadatos de e-prints; igualmente, detectaba y analizaba las referencias en los más de 200.000 artículos de física del Laboratorio Los Álamos.

Numerosos estudios se han desarrollado durante los últimos diez años, dedicados a mostrar la visibilidad del trabajo científico publicado en abierto, ya que existe la convicción generalizada de que los trabajos en abierto logran más descargas y citas y por tanto tienen mayor impacto. Las afirmaciones en tal sentido, a partir de estudios bibliométricos, incluyen al propio Instituto de Información Científica (ISI, por sus siglas en inglés), responsable por el factor de impacto. Numerosos trabajos apoyan las afirmaciones en pos de la publicación en abierto para dar mayor visibilidad a la obra y a su autor (Antelman, 2004; Harnad, Brody, 2004; McVeigh, 2004; McVeigh, Testa, 2004; Shadbolt, Brody, Carra, Harnad, 2006; Eysenbach, 2006).

Hoy en día existen muchas más formas de medir el impacto de una obra que la

medición bibliométrica representada por el factor de impacto. Entre los muchos trabajos que se encuentran en este sentido, el de Isabel Bernal (2013) tiene un extenso apartado titulado "Críticas al modelo tradicional de medida de la producción científica y nuevos criterios de medición", en el que expone en detalle no sólo las objeciones a los criterios tradicionales de evaluación científica, sino otros nuevos factores de reputación, impacto, influencia y productividad como los indicadores de Webometrics[3], Altmetrics y otras múltiples fuentes de recogida de datos y plataformas abiertas y gratuitas. La autora se explaya sobre repositorios y revistas de acceso abierto como fuentes de datos y particulariza las opciones de un investigador que, por ejemplo, deposita sus obras en Digital CSIC, repositorio que es recorrido e indexado por numerosos buscadores, especialmente académicos como Google Scholar, CiteSeer, Citebase, etc. La autora también pone ejemplos de cantidad de citas y descargas de las obras y las posibilidades al momento de obtener estadísticas de la producción científica y académica. Trae a colación la consolidación del concepto de impacto social de la investigación y la incorporación del acceso abierto como criterio de evaluación, haciéndose eco de la necesidad de un sistema de evaluación *"más matizado que no dependa exclusivamente del lugar de publicación y el número de citas y un nuevo análisis del método de peer review; nuevos modelos de publicación y apoyo institucional en temas de copyright"*.

Por su parte, otro trabajo al respecto es el de Wagner (2010): una recopilación bibliográfica de más de 50 trabajos localizados en Google Scholar, SciFinder, Medline, Web of Science y otros, a texto completo y dedicados a analizar si el AA da ventajas en citas.

Como contraparte, el comportamiento de los autores en relación al autoarchivo no muestra todavía un compromiso fuerte con esta práctica. Dos artículos muy citados de Alma Swan (Swan, Brown, 2004; Swan, 2005) relatan los estudios sobre el comportamiento de los autores que realizó el JISC. Los estudios ponían de manifiesto un fuerte desconocimiento de las modalidades de publicación en revistas de acceso

[3] El objetivo original del Ranking Mundial de Universidades (y repositorios de AA) realizado por Webometrics, todavía válido, es apoyar las iniciativas de Acceso Abierto y promover el acceso global al conocimiento académico producido por las universidades de todo el mundo. El ranking pretende ser una herramienta útil que muestre el compromiso de las instituciones con el AA a través de indicadores web cuidadosamente seleccionados (véase www.webometrics.info/es). Recuperado el 30/05/2014.

abierto o del autoarchivo en repositorios institucionales, y las principales causas señaladas eran el desconocimiento de las ventajas del acceso abierto y, en particular en lo relativo al autoarchivo, las relacionadas con la preservación y la protección de derechos de autor frente al circuito tradicional de publicación. La encuesta también revelaba que un porcentaje muy elevado de autores (81%) aceptaría archivar en caso de existir mandato institucional. De allí se desprende que una política a nivel de país, que apoye estas prácticas y brinde las garantías y los incentivos para proporcionar el acceso abierto, es uno de los caminos más apropiados para lograrlo.

Sánchez Tarragó (2007) menciona también algunas instituciones con exitosas políticas de autoarchivo de carácter obligatorio, como la Universidad de Southampton en el Reino Unido, la Universidad Tecnológica de Queensland en Australia y la Universidad del Minho en Portugal, entre otras. Estas instituciones reportan un crecimiento exponencial en sus tasas de autoarchivo desde su implementación con carácter obligatorio.

Stevan Harnad, uno de los líderes del AA, ha propuesto una estrategia para hacer más efectivas las políticas de autoarchivo y la ha denominado "Depósito inmediato/Acceso Opcional". Consiste en exigir el depósito del texto completo del artículo arbitrado y los datos bibliográficos (metadatos) en un repositorio institucional inmediatamente después de su aceptación para publicarse. En este caso, sólo el depósito sería obligatorio, mientras que establecer los privilegios de acceso quedaría en manos del autor (que dependerá, por ejemplo, de sus acuerdos de derecho de autor con la editorial), aunque se recomienda encarecidamente el acceso abierto. De esta manera, la política de autoarchivo de la institución sería completamente independiente de las de los editores.

Harnad (2006) sostiene que ninguna institución de investigación puede mantener todas las revistas que sus investigadores necesitan, lo que redunda en una pérdida de impacto de la investigación, y remarca que los artículos abiertos reciben el doble de citas pero que sólo un 15% de ellos se autoarchivan. Las únicas instituciones que se aproximan al 100% de autoarchivo son las que han elaborado un mandato institucional al respecto. Las encuestas muestran que el 95% de los autores cumplen en este caso con el autoarchivo, y Harnad se extiende sobre los lineamientos de un modelo de acceso abierto para el propio beneficio de las instituciones.

Existen numerosos y variados trabajos que describen el estado de situación en mandatos y políticas de acceso abierto. Uno interesante es el referido al estado de situación de las instituciones de España. La razón de exponer aquí la situación de España deviene del hecho de que ese país siempre ha estado —y está— muy comprometido con el AA. Para el año 2012, fecha del artículo mencionado, tan sólo un 21% de las instituciones no contaban con alguna forma de mandato. El informe final y diagnóstico culminado por la RedCLARA, también de 2012, referido a América Latina, por su parte, enuncia: "...72% de las instituciones recomienda el acceso abierto y sólo el 19% tiene mandato institucional", lo que muestra la importancia capital de los repositorios institucionales y de los mandatos para los países de este continente.

Otro gran avance en este sentido lo constituye el Horizonte 2020 sobre AA de la Unión Europea de Innovación e Investigación, cuyo objetivo es optimizar el impacto de la investigación científica financiada con fondos públicos, tanto a nivel europeo como en los países miembros, ya que se lo considera esencial para aumentar el rendimiento económico y mejorar la competitividad a través del conocimiento.

Para un análisis de los mandatos existentes se recomienda ver ROARMAP (Registry of Open Access Repositories Mandatory Archiving Policies). Se trata de una página web desarrollada por la Universidad de Southampton (Reino Unido), que ofrece una lista de las políticas de acceso abierto desarrolladas por universidades e institutos de investigación. En esta página las instituciones se autoregistran y está complementada por las políticas localizadas por los responsables del sitio de ROARMAP. Si bien no es el número real y total de políticas, su información resulta muy útil y brinda un panorama más que claro al respecto.

Todo lo anterior apunta a remarcar que uno de los pilares fundamentales del movimiento de Acceso Abierto es la adopción de políticas institucionales o nacionales que coaccionen a los autores a depositar sus trabajos en repositorios temáticos o institucionales de acceso abierto. Sin embargo, hasta ahora, el éxito de estas políticas de autoarchivo es variable, ya que algunas de ellas no exigen depositar los trabajos, sino que invitan o exhortan, por lo que la respuesta de los autores aún es insuficiente. Se evidencia, además, que a la par del establecimiento de la política debe existir todo un esfuerzo de persuasión, difusión y entrenamiento al respecto.

En este sentido, las políticas formales académicas y científicas deberían estimular a

los investigadores a utilizar licencias de acceso abierto y/o a no ceder de manera exclusiva los derechos sobre sus obras al momento de publicarlas en revistas científicas. Las políticas de autoarchivo, como se dijo, deberían ir acompañadas de transformaciones en el sistema de evaluación científica, de manera tal que aquellos autores que publiquen bajo modalidades de acceso abierto, se reconozcan y retribuyan en el orden científico y económico.

Bibliografía del capítulo

Albert, K. M. (2006). "Open access: implications for scholarly publishing and medical libraries". *Journal of the Medical Library Association*, 94 (3) 253-62.

Antelman, K. (2004). "Do Open-Access Articles Have a Greater Research Impact?", en *College & Research Libraries News. American Library Association*. pp.372-382. Recuperado el 9 de junio de 2014, de http://eprints.rclis.org/handle/10760/5463.

Argentina. Ley 26.899: Creación de Repositorios Digitales Institucionales de Acceso Abierto, Propios o Compartidos. Buenos Aires, *Boletín Oficial de la República Argentina*, lunes 9 de diciembre de 2013, año CXXI, número 32.781, p. 3. Recuperado el 9 de junio de 2014, de http://www.boletinoficial.gov.ar/.

ArXiv (1991). Recuperado el 8 de junio de 2014, de http://arxiv.org/.

Barrionuevo, L. (2009). "Acceso Abierto". *Glossarium-BITri*. Recuperado el 8 de junio de 2014, de http://glossarium.bitrum.unileon.es/Home/acceso-abierto.

Bernal, I. (2013). "Digital CSIC: desarrollo de contenidos. Gestión de Copyright. Impacto de la Ciencia en Acceso Abierto". Curso impartido del 20 al 22 de marzo de 2013 en el Centro de Ciencias Humanas y Sociales del CSIC. Recuperado el 9 de junio de 2014, de http://digital.csic.es/handle/10261/73245.

BioMedCentral (2000). Recuperado el 8 de junio de 2014, de http://www.biomedcentral.com/journals.

Budapest Open Access Initiative (BOAI) (2002). Budapest Open Access Initiative. Recuperado el 8 de junio de 2014, de http://www.budapestopenaccessinitiative.org/read Versión en español: http://www.geotropico.org/1_1_Documentos_BOAI.html.

Creative Commons (CC) (2002). Licencias Creative Commons. Recuperado el 9 de junio de 2014, de http://creativecommons.org/.

E-Prints (2000). Recuperado el 29 de mayo de 2014, de http://www.eprints.org/.

Eysenbach, G. (2006). "The Open Access Advantage". *Journal of Medical Internet Research* 8 (2) Recuperado el 9 de junio de 2014, de http://www.jmir.org/2006/2/e8.

Guiu, J. M.; García-Ramos García, R. (2008). "El factor de impacto y las decisiones editoriales". *Neurología: Publicación oficial de la Sociedad Española de Neurología*, 23 (6). p. 342-348.

Harnad, S. (2001). "For Whom the Gate Tolls?". Harnad, Stevan. *How and Why to Free the Refereed Research Literature Online Through Author/Institution Self-Archiving, Now*. Recuperado el 29 de abril de 2011, de

http://users.ecs.soton.ac.uk/harnad/Tp/resolution.htm

Harnad, S., Brody, T. (2004). "Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals". *D-Lib Magazine*, 10 (6). Recuperado el 16 de junio de 2014, de http://www.dlib.org/dlib/june04/harnad/06harnad.html.

Harnad, S. (2006). "Maximizing Research Impact through Institutional and National Open-Access Self-Archiving Mandates". CRIS2006. Open Access Institutional Repositories. *Current Research Information Systems*. Bergen, Norway, 11-13 de mayo. Recuperado el 9 de junio de 2014, de http://eprints.ecs.soton.ac.uk/12093/1/harnad-crisrev.html.

Hoorn, E. (2005). "Repositories, copyright and creative commons for scholarly communication". *Ariadne*, 25. Recuperado el 9 de junio de 2014, de http://www.ariadne.ac.uk/issue45/hoorn/.

Hoorn E., Van Der Graaf, M. (2006). "Copyright issues on open access research journals". *D-Lib Magazine*, 12 (2). Recuperado el 9 de junio de 2014, de http://www.dlib.org/dlib/february06/vandergraaf.html.

Joint Conferences on Digital Libraries (JCDL) (2001). Recuperado el 27 de mayo de 2014, de http://www.sigweb.org/conferences/sigweb-conference/11-conferences/24.

Kyrillidou, M.; Bland, L. (2009). *ARL Statistics 2007–2008*. Association of Research Libraries, Washington, DC.

Melero Melero, R.; Barrueco Cruz, J. M. (s/d). "Copyright y auto-archivo: hábitos de los autores. Tipos de licencias para la cesión no exclusiva de copyright. Publicar vs distribuir". Asociación Española de Documentación e Información. Recuperado el 29 de mayo de 2014, de http://www.sedic.es/autoformacion/acceso_abierto/3-copyright-autoarchivo.html#.

McVeigh M. E. (2004). "Open Access Journals in the ISI Citation Databases: Analysis of Impact Factors and Citation Patterns. A citation study from Thomson Scientific". Recuperado el 9 de junio de 2014, de http://ip-science.thomsonreuters.com/m/pdfs/openaccesscitations2.pdf.

National Coordinating Centers Delegates of the Latin American and Caribbean System on Health Sciences Information (1998). "Declaration of San José Towards the Virtual Health Library". En: *VI Meeting of Latin American and Caribbean System on Health Sciences Information, IV Pan American Congress on Health Sciences Information*, San José, Costa Rica. Recuperado el 29 de mayo de 2014, de http://www.bireme.br/bvs/por/ideclar.htm.

Open Citation Project (OpCit) (1999). Recuperado el 9 de junio de 2014, de http://opcit.eprints.org/.

Peter Suber (s/d). En Wikipedia. Recuperado el 29 de mayo de 2014, de http://en.wikipedia.org/wiki/Peter_Suber

PubMed (1996). Recuperado el 8 de junio de 2014, de http://www.ncbi.nlm.nih.gov/pubmed/.

PubMed Central (2000). Recuperado el 8 de junio de 2014, de http://www.ncbi.nlm.nih.gov/pmc/.

Registry of Open Access Repository (ROAR) (2003). Recuperado el 9 de junio de 2014, de http://roar.eprints.org/.

Rights Metadata for Open Archiving (RoMEO) (2006). Recuperado el 9 de junio de 2014, de http://www.sherpa.ac.uk/romeo/?la=es.

Registry of Open Access Repositories Mandatory Archiving Policies (ROARMAP) (2003). Recuperado el 9 de junio de 2014, de http://roarmap.eprints.org/.

Sánchez Tarragó, N. (2007). "El movimiento de acceso abierto a la información y las políticas nacionales e institucionales de autoarchivo". *Acimed.*, 16 (3). Recuperado el 14 de septiembre de 2011, de http://bvs.sld.cu/revistas/aci/vol16_3_07/aci01907.html.

Schwartz, C. (1999). "A Working Definition of Digital Library" Digital Library Federation Recuperado el 29 de abril de 2011., de http://www.clir.org/diglib/dldefinition.htm

Schwartz, C. (2000). "Digital Libraries: An Overview". *The Journal of Academic Librarianship,* Volume 26, Number 6, p. 385-393. Recuperado el 23 de junio de 2014, de http://home.kku.ac.th/hslib/malee/412725/document/DLOverview.pdf.

SciELO (1997). Recuperado el 29 de mayo de 2014, de http://www.scielo.org/php/index.php.

Self-Archiving (1997). Recuperado el 29 de mayo de 2014, de http://users.ecs.soton.ac.uk/harnad/Tp/resolution.htm

Shadbolt, N., Brody, T., Carra, L., Harnad, S. (2006). "The Open Research Web: A Preview of the Optimal and the Inevitable". En: Jacobs, N. (ed.). *Open Access: Key Strategic, Technical and Economic Aspects.* Oxford: Chandos Publishing. Recuperado el 9 de junio de 2014, de http://eprints.ecs.soton.ac.uk/12369/.

SHERPA/RoMEO (2000). Políticas de la editoriales. Recuperado el 9 de junio de 2014, de http://www.sherpa.ac.uk/romeo/search.php.

SHERPA/RoMEO (2014). Statistics for the 1563 publishers in the RoMEO database. Recuperado el 9 de junio de 2014, de http://www.sherpa.ac.uk/romeo/statistics.php?la=en&fIDnum=|&mode=simple.

Sistema de Información en Ciencias de la Salud de Latinoamérica y el Caribe (BIREME). (s/d). Recuperado el 29 de mayo de 2014, de http://regional.bvsalud.org/php/index.php.

Stevan Harnad (s/d). En Wikipedia. Recuperado el 29 de mayo de 2014, de http://en.wikipedia.org/wiki/Stevan_Harnad.

Suber, P. (2001-2008). "Journal declarations of independence". En: Lists Related to The Open Access Movement. Recuperado el 23 de junio de 2014, de http://legacy.earlham.edu/~peters/fos/lists.htm#declarations.

Suber, P. (2009). Timeline of the Open Access Movement. Recuperado el 8 de junio de 2014, de http://www.earlham.edu/~peters/fos/timeline.htm.

Swan, A., Brown, S. (2004). "Authors and open access publishing". *Learned Publishing*, 17 (3), pp. 219-224. Recuperado el 9 de junio de 2014, http://eprints.ecs.soton.ac.uk/11003/ .

Swan, A. (2005). "Open access self-archiving: An Introduction". *Technical Report, JISC, HEFCE.* Recuperado el 9 de junio de 2014, de http://eprints.ecs.soton.ac.uk/11006/.

Testa, J.; McVeigh, M. E. (2004). "The Impact of Open Access Journals. A Citation Study from Thomson ISI". Recuperado el 9 de junio de 2014, de http://www.lib.uiowa.edu/scholarly/documents/ISI_impact-oa-journals.pdf.

Universidad Nacional de La Plata (2011). Resolución no. 78/11. N° de expediente: 100-8234/11. Recuperado el 9 de junio de 2014, de http://sedici.unlp.edu.ar/handle/10915/18184.

Wagner, A. B. (2010). "Open Access Citation Advantage: An Annotated Bibliography". *Issues in Science and Technology Librarianship*. Recuperado el 9 de junio de 2014, de http://www.istl.org/10-winter/article2.html.

Webometrics (s/d). Recuperado el 9 de junio de 2014, de www.webometrics.info/es.

Capítulo 2 | Conceptos

«Lo importante es nunca dejar de cuestionarse.»

ALBERT EINSTEIN

Síntesis: Este capítulo está dedicado a los conceptos principales que involucran a las bibliotecas digitales y repositorios institucionales; especialmente, se pone el acento en las características que diferencian un repositorio institucional de cualquier otro sistema de acopio de documentos. Las definiciones que se recogen aquí están presentes para entender de manera cabal el objeto que se va a evaluar posteriormente; tras ellas se describen elementos centrales de los repositorios, como los metadatos, y se enuncian las funciones esperadas que deben cumplir las tecnologías con las que se implemente un repositorio. A través de la enunciación de las funciones se espera simplificar la comprensión de los modelos presentados en el capítulo 3 y de los elementos a evaluar en el capítulo 4.

Bibliotecas Digitales (BD): panorama

El concepto de Biblioteca Digital (BD) ha tenido un crecimiento sostenido desde los albores del año 2000, empujado por el incremento de los recursos de cómputo, las redes y el decremento paralelo de los costos para acceder a servicios de este tipo. Si bien la web es, en la actualidad, el primer lugar de búsqueda elegido por la mayoría de las personas de diferentes extracciones, como antes lo fueran las bibliotecas, hay que destacar un punto fundamental que comparten todas las bibliotecas, más allá de la naturaleza de los objetos (físicos o digitales) que contienen, que es el concepto de selección. Quiere decirse que en cualquier biblioteca, siempre habrá un subconjunto de objetos de información seleccionado (en oposición a otros contenidos excluidos), segregado, disponible, preservado, y cuyo acceso está además favorecido por servicios añadidos, como la posibilidad de búsqueda de información, entre otros.

Bibliotecas Digitales: concepto refinado y definiciones varias

La motivación principal de una BD debe ser el provecho de su uso para la búsqueda y el avance en el conocimiento de las personas. En esto, en nada se distingue de una biblioteca tradicional, pero involucra metodologías que devienen de la naturaleza digital de sus acervos y en esto sí se separa de la biblioteca tradicional. A los fines de comprender más cabalmente qué es una BD, se realizará a continuación un breve recorrido por las distintas significaciones y alcances del término a lo largo de estos años.

Un extracto del documento de la NSF (National Science Foundation), que se reproduce a continuación, marca el origen del término en 1994. En aquel momento, esta área del conocimiento se concentraba en el desarrollo de herramientas informáticas e infraestructura capaz de manejar información digital y las incumbencias estaban delimitadas al área informática:

> *"Las bibliotecas digitales básicamente almacenan materiales en formato electrónico y manipulan grandes colecciones de estos materiales con eficacia. La investigación dentro de las bibliotecas digitales es investigación dentro de sistemas de información en red y se concentra en cómo desarrollar la infraestructura para manejar de manera efectiva la masa de información presente en la red. La cuestión tecnológica clave es cómo buscar y mostrar una selección deseada a partir de colecciones muy grandes."* (NSF, 1999)[4]

En la década del 90, organismos estatales de los Estados Unidos como la mencionada NSF, DARPA y NASA llevaron a cabo dos iniciativas[5]: DLI-1 y DLI-2 (Digital Library Initiative), cuyo objetivo fue desarrollar e implementar modelos de bibliotecas digitales. El primero de ellos, DLI-1, se desarrolló entre 1994 y 1998, y se

[4] Texto en inglés: *"Digital Libraries basically store materials in electronic format and manipulate large collections of those materials effectively. Research into digital libraries is research into network information systems, concentrating on how to develop the necessary infrastructure to effectively mass-manipulate the information on the Net. The key technological issues are how to search and display desired selections from and across large collections."*

[5] En el programa original de la iniciativa podía leerse lo siguiente: *"The Initiative's focus is to dramatically advance the means to collect, store, and make [information] available for searching, retrieval, and processing via communication networks –all in user-friendly ways."*

caracterizó por el desarrollo de buscadores, capacidades de procesamiento, digitalización, implementación de infraestructura y redes de comunicación. Dado el avance logrado en este proyecto, se decidió continuarlo, y así se dio lugar a DLI-2, que se desarrolló entre 1999 y 2003. Esta iniciativa se caracterizó por el desarrollo de la interoperabilidad y la conexión entre bibliotecas digitales.

El portal de la JCDL (Joint Conference on Digital Libraries), que se realiza desde 2001 a la fecha, es un importante foro internacional centrado en las bibliotecas digitales y las cuestiones técnicas, prácticas y sociales que atañen a ellas. Cabe citarlo aquí, pues es un buen lugar para comenzar a bucear en definiciones reconocidas y encontrar la convergencia de un término esquivo. Refiriéndose a JCDL, sus responsables se extienden en los muchos significados que abarca el concepto de las "bibliotecas digitales"[6], y así aseveran que

> *"incluye (pero no por ello se limita a) nuevas formas de instituciones de información; sistemas de información operativa con todo tipo de contenido digital; modelos teóricos de medios de información, géneros de documentos y publicación electrónica. Las BD se distinguen también de los sistemas de recuperación de información, ya que incluyen más tipos de medios y proporcionan funcionalidades adicionales y servicios, e incluyen todas las fases del ciclo de vida de la información, desde la creación y a través del uso."* (JCDL)

En 1994 aparece un primer estudio analítico de los elementos en las definiciones: *Analytical Review of the Library of the Future* de K. M. Drabenstott (1994), obra que abunda en las funciones y objetivos que se visualizaban entonces para las BD. Otros estudios, citados como contraparte por Tramullas (2002), resultan muy interesantes porque se centran en los elementos tecnológicos de los proyectos de BD. Otra referencia obligada en el área es el trabajo de Arms, Bianchi y Overly (1997).

[6] Texto en inglés: *"JCDL encompasses the many meanings of the term "digital libraries", including (but not limited to) new forms of information institutions; operational information systems with all manner of digital content; new means of selecting, collecting, organizing, and distributing digital content; and theoretical models of information media, including document genres and electronic publishing. Digital libraries are distinguished from information retrieval systems because they include more types of media, provide additional functionality and services, and include other stages of the information life cycle, from creation through use".*

Si se continúa evaluando cronológicamente el concepto, puede verse que para Lesk (1997): *"las bibliotecas digitales son colecciones organizadas de información digital. Combinan la estructura y concurrencia de la información, que siempre han tenido las bibliotecas y los archivos, con la representación digital que han hecho posible las computadoras"*[7].

Por su parte, Borgman (1999) intenta explicar el significado y la interpretación de la frase "biblioteca digital", analizando varias definiciones propuestas por investigadores y comunidades vinculadas al área. La autora distingue dos sentidos distintos: una definición tecnológica que establece que las bibliotecas digitales

> *"son un conjunto de recursos electrónicos y capacidades técnicas asociadas para crear, buscar y utilizar la información (...), son una extensión y mejora de sistemas de almacenamiento y recuperación que manipulan los datos digitales en cualquier medio. El contenido de las bibliotecas digitales incluye los datos y metadatos"* (p. 234)

en contraste con un punto de vista social, que establece que: *"las bibliotecas digitales son construidas, recopiladas y organizadas, por (y para) una comunidad de usuarios, y sus capacidades funcionales de apoyo a las necesidades de información y usos de la comunidad"* (p. 234)[8].

La definición de Borgman contiene un espectro interesante de elementos sobre los cuales ha de volver una y otra vez esta tesis, ya que constituyen los sujetos o candidatos para la evaluación: recursos electrónicos, capacidades técnicas, recuperación de información, metadatos, usuarios.

En el 2000, Arms propuso una definición informal: *"una biblioteca digital es una*

---

[7] Texto en inglés: *"Digital libraries are organized collections of digital information. They combine the structuring and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible."* (p. XIX).

[8] Textos en inglés: *"Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching and using information...they are an extension and enhacement of information storage and retrieval systems that manipulate digital data in any medium. The contents of digital libraries include data and metadata"*. (p. 234). *"Digital libraries are constructed, collected and organized, by (and for) a community of users, and their functional capabilities support the information needs and uses of that community."* (p. 234).

colección gestionada de información, con servicios asociados, donde la información es almacenada en formato digital y es accesible en toda la red"[9]. Esta definición enfatiza claramente los aspectos de la gestión de los contenidos.

La Digital Library Federation (DLF), organización establecida en los Estados Unidos en 1995 y dedicada a la creación, el mantenimiento, la expansión y distribución de colecciones de materiales digitales accesibles para estudiantes y un público amplio, considera a las BD como organizaciones que proveen recursos que incluyen a personas especializadas y estas organizaciones seleccionan, estructuran, preservan y dan acceso a colecciones digitales destinadas a comunidades designadas (Schwartz, 1999, 2000).

Así, el concepto de BD es, tal cual afirma Tramullas Saz (2002), un concepto que se definió cuando las BD llegaron a un nivel de madurez tal que el objeto de estudio estuvo lo suficientemente claro por sus propias funciones y herramientas, así como por los componentes tecnológicos.

Otro trabajo fundacional es el de Candy Shwartz (2000), que recorre un conjunto de definiciones formales e informales para mostrar los recursos de una BD, a quiénes sirven los mismos y qué materiales y funcionalidades brindan. El eje desde el cual parte su trabajo es la necesidad de proveer un contexto a través de la perspectiva de los componentes del trabajo de una biblioteca digital.

Se ha dejado en último lugar al grupo DELOS, para finalizar este decurso con las reflexiones —extensas en el tiempo y de probada fiabilidad— de este grupo. DELOS ha actuado como una una asociación de BD en los últimos diez años (tres años como un grupo de trabajo, otros tres como una red temática y cuatro como una red de excelencia). DELOS[10] ha realizado una contribución sustancial a la creación, en Europa, de una comunidad de investigación en bibliotecas digitales.

Dado el estado de madurez y experiencia acumulada en la década previa, en el año

---

[9] Texto en inglés: *"A digital library is a managed collection of information, with associated services, where the information is stored in digital formats and accesible over the network."*

[10] DELOS (Network of Excellence on Digital Libraries) fue fundada por la Comisión Europea y sus principales objetivos eran la investigación, cuyos resultados perteneciesen al dominio público, y la transferencia de tecnología, a través de acuerdos de cooperación con las partes interesadas. La red concebía a las BD (a través de redes posibilitadas por Internet) como el medio para que cualquier persona pudiera acceder al conocimiento que se encontraba alojado en bibliotecas, museos y colecciones. Con ese fin, realizó un programa conjunto de actividades dirigidas a la integración y coordinación de los esfuerzos de investigación en curso de los principales equipos europeos que trabajaban en áreas relacionadas con la BD.

2007, el grupo DELOS publicó el llamado "Digital Library Manifesto". El Manifiesto de la Biblioteca Digital establece un marco conceptual con tres niveles: Biblioteca Digital (DL, por sus siglas en inglés) es la organización que recopila, gestiona, preserva y ofrece contenidos digitales; Sistema de Biblioteca Digital (DLS, por sus siglas en inglés) es el sistema de software que proporciona la funcionalidad requerida por una BD particular, y Sistema de Gestión de Biblioteca Digital (DLMS, por sus siglas en inglés) se refiere a la plataforma: sistema operativo, bases de datos, interfaz de usuario.

Como corolario de este recorrido, cuyas citas y definiciones podrían extenderse a muchas otras, es posible vislumbrar la evolución de objeto de estudio y aseverar que una mirada sobre el presente panorama sirve para mostrar que la evaluación de un repositorio institucional deberá hacerse sobre un conjunto amplio de elementos que integran ese conjunto complejo que es una BD. Así, todas las nociones vinculadas a las BD necesitan complejizarse para superar la dicotomía de definiciones dadas desde diferentes puntos de vista: el ámbito de investigación y el ámbito bibliotecario. El área de trabajo amerita definiciones y prácticas más elaboradas porque obliga a la participación de distintos campos del saber.

Repositorios Institucionales (RI)

La Universidad Autónoma de Madrid define un repositorio institucional (RI) como un conjunto de servicios web centralizados, creados para organizar, gestionar, preservar y ofrecer acceso libre a la producción científica, académica o de cualquier otra naturaleza cultural, en soporte digital, generada por los miembros de una institución. Las principales características de un RI, según esta conceptualización, son:

- su *naturaleza institucional*, entendiendo por institución a una organización educativa y de investigación, cuyo punto de partida fueron las universidades;
- su *carácter científico, acumulativo y perpetuo*;
- su *carácter abierto e interoperable* con otros sistemas.

A lo dicho, resulta importante agregar una característica de todos ellos en conjunto: la diversidad. Si existe un elemento que los nuclea es, precisamente, que ninguno se parece a otro.

Deslindes terminológicos y aclaraciones

Puede parecer que en los párrafos anteriores se confunden y aúnan las nociones de BD y de RI. Los RI comparten muchas características con las BD, pero se podrían distinguir por algunos principios: por ejemplo, que los repositorios institucionales están diseñados principalmente para recoger, preservar y poner a disposición la producción académica de una institución; alternativamente, las bibliotecas digitales pueden estar organizadas en torno a otros principios: temas, disciplinas, o incluso tipos de documentos en particular, pero como se ha visto los repositorios también pueden organizarse con estas modalidades.

Los RI y las BD también difieren en cuanto a cómo adquirir contenidos. Mientras que las colecciones que figuran en las bibliotecas digitales son generalmente el resultado de esfuerzos deliberados de desarrollo de la colección por parte de los profesionales de la biblioteca, los repositorios institucionales son típicamente dependientes de las contribuciones voluntarias de los investigadores. Sin embargo, como se ha visto, a pesar de las bajas tasas de autoarchivo, éste y el depósito mediado han pasado a ser también una característica de estos últimos.

Otra diferencia inicial, entre los RI y las BD, era que los repositorios fueron pensados en sus inicios fundamentalmente como lugares para almacenar los materiales, con servicios mínimos ofrecidos a los usuarios; por su parte, las bibliotecas digitales ofrecían muchos más servicios a los usuarios, incluyendo el apoyo del personal en la búsqueda de información adicional e incluso en la interpretación. No obstante, en la actualidad esto tampoco distingue unos de otras, ya que con el objetivo de incrementar el interés de la comunidad por el depósito y la necesidad institucional de visibilizar las obras en el repositorio, los RI ofrecen muchos servicios con distintos alcances: asesoramiento sobre derecho de autor y derechos en general, listado de publicaciones por autores, información sobre descargas de las obras, tutoriales para mejorar las búsquedas y el uso del repositorio, etc.

Como se ha dicho, los RI pertenecen a una institución académica o de investigación, y se pretende que los materiales de la casa que representan la producción intelectual de esa organización estén presentes. Debido a esto, un RI es, necesariamente, una colección de documentos y objetos, por lo general de varios tipos y formatos. Investigadores afiliados a la organización patrocinadora del RI pueden (y deben,

cuando existe mandato) depositar los textos, los conjuntos de datos, los archivos de sonido, imágenes o cualquier número de otros artículos. Estos documentos (de acuerdo a la política de contenidos del RI) pueden estar en cualquier etapa del proceso de producción académica: preprints, postprints, material que no ha pasado procesos de referato, todo lo cual también dependerá siempre de la política de la institución.

Para los intereses de esta tesis, vinculados a las tecnologías y al acceso a los contenidos, la noción de RI resulta la más adecuada, dado que el objeto de estudio es un repositorio institucional. Así, en las páginas y capítulos siguientes se hablará utilizando este término, pero hay que hacer notar que, sin embargo, la aplicabilidad metodológica presentada podrá alcanzar a unos y otros siempre que se busquen objetivos similares.

Párrafo aparte merecen los repositorios de objetos de aprendizaje, ya que mientras que el desarrollo de repositorios con contenidos de investigación representa una migración relativamente intuitiva de prácticas de publicación tradicional, en el ámbito de la enseñanza es posible observar una transición menos intuitiva. La elaboración de material didáctico en forma digital abarca tanto material institucional y de autor de muchas especies distintas: material de cátedra, notas, colecciones de imágenes, animaciones e incluso libros de texto, además de clases, programas, actividades prácticas y otros materiales complementarios, por lo que captar el material de aprendizaje puede ser más complicado. Los derechos de autor presentan un gran obstáculo en este sentido y las instituciones no están tan fuertes como para exenciones, incluso en los derechos de autor de libros de texto a los que pueden haber contribuido, como sí lo están con los trabajos de investigación escritos por sus propios académicos. El argumento que puede dar mayor fortaleza a estas prácticas es la posibilidad del reúso por parte de los colegas de la institución.

Los objetos de aprendizaje, como puede verse, son un grupo heterogéneo de materiales que varían enormemente en su formato, en los requisitos de los metadatos, y en tamaño. Agruparlos en un mismo repositorio presenta muchos desafíos. Las ventajas de hacerlo, sin embargo, son las mismas que las que se aplican a los trabajos de investigación. Implica realizar un uso más eficiente de los recursos de la institución, preservar los contenidos digitales, ofrecer una visión completa de esta producción, brindar herramientas de apoyo para búsquedas pertinentes, y permitir la

interoperabilidad con otros repositorios similares. No obstante, las instituciones todavía no ven que el agregado y agrupamiento de una colección de objetos de aprendizaje agregue valor y visibilidad a su producción científica. Debe tenerse en cuenta también que mientras que los materiales de investigación tienden a ser muy leídos por los demás miembros de una comunidad disciplinaria en todo el mundo, el valor de los objetos de aprendizaje radica en su capacidad para ser reutilizados, hecho en el que aún no se ha insistido lo bastante en las comunidades educativas que podrían disponer de ellos.

Componentes principales de un repositorio

Con el objeto de comprender cómo es y cómo funciona un repositorio, se detallarán a continuación sus principales componentes.

Metadatos

En la construcción de RI hay, al menos, dos componentes principales. Uno de ellos es la tecnología y el otro son los metadatos. Los metadatos puede decirse que son "los datos sobre datos". Como comúnmente se describe, la creación de metadatos consiste en la producción de información estructurada que permite la gestión adecuada, el mantenimiento, la trazabilidad de los contenidos digitales e incluso los derechos sobre los recursos (metadatos administrativos), y dan la posibilidad de navegación sobre los recursos, describiendo su estructura interna y relaciones con otros recursos (metadatos estructurales), la posibilidad de intercambio con sistemas similares (interoperabilidad), la preservación y la localización de los contenidos (metadatos descriptivos), gracias a una descripción adecuada. También pueden informar sobre los recursos de software y hardware, y los formatos (metadatos técnicos). Las clasificaciones de los metadatos son numerosas y para una explicación inicial puede verse, entre otros muchos, Eva Méndez (2008), trabajo que se destaca porque contrasta los metadatos tradicionales con los de preservación, que ocuparán buena parte de este estudio.

Los metadatos se ordenan en esquemas y así existen numerosos esquemas de metadatos, como los planos y los jerárquicos; también hay metadatos que se usan para determinadas actividades del repositorio: por ejemplo, en el protocolo OAI-PMH,

como un "mínimo común denominador" para el intercambio de registros, se utiliza el esquema de metadatos Dublin Core simple, que cuenta con 15 elementos.

La complejidad creciente de los objetos digitales y la posibilidad de que se encuentren compuestos de múltiples archivos, está trayendo nuevos esquemas de metadatos que se van sumando a la catalogación mínima inicial que formaba parte de los objetos de las bibliotecas en papel. Estos esquemas están adaptados a las diferentes dimensiones de los objetos (gracias a los metadatos descriptivos, los metadatos técnicos, los metadatos estructurales y los administrativos). Para dar un ejemplo concreto, un esquema de metadatos jerárquico como METS está estructurado en siete secciones, como sigue:

I. *Cabecera*: son los metadatos que describen el documento.
II. *Sección descriptiva*: metadatos externos descriptivos (como un registro MARC) o bien metadatos descriptivos integrados internamente, o ambas cosas.
III. *Metadatos administrativos*: esta sección contiene información que describe cómo los archivos se crean y se almacenan, los derechos de propiedad intelectual, etc.
IV. *Archivo*: muestra todos los archivos de contenido que conforman el objeto digital.
V. *Mapa estructural*: describe una estructura jerárquica para el objeto, y los elementos de enlaces a archivos de contenido y los metadatos relacionados.
VI. *Vínculos estructurales*: esta sección registra la existencia de enlaces entre los nodos de la jerarquía que se indica en el mapa estructural.
VII. *Comportamiento*: sección que se puede utilizar para asociar comportamientos ejecutables con el contenido del objeto METS.

También cabe observar que ciertos tipos de contenido, por ejemplo los objetos de aprendizaje, utilizan estándares especiales, como Learning Object Metadata (LOM). LOM realiza una tarea similar a METS en el apoyo a la catalogación de los objetos compuestos, pero dentro de un contexto pedagógico. El estándar LOM tiene nueve categorías (general, del ciclo de vida, meta-metadatos, técnicos, educativos, derechos, relación, anotación y clasificación). Los metadatos educativos y de anotación son claramente las categorías específicas de los objetos de aprendizaje. Este trabajo no

abundará en extensión ni en explicaciones sobre los formatos, sino que, de manera más natural, dichos aspectos se expondrán en la propia experimentación, como dados por el uso del repositorio institucional de prueba, SEDICI.

Tecnologías expresadas como funciones generales del repositorio institucional

Los requerimientos tecnológicos deben ser uno de los últimos elementos a considerar en la planeación de un repositorio institucional. La pregunta a formularse es cuál es el software más adecuado para los servicios que pretende brindar la institución con el repositorio. Además, se debe tener en cuenta qué tecnologías utiliza la institución, con qué recursos se cuenta, incluidas las capacidades humanas, y otro punto a considerar es qué desarrollos utilizan otras instituciones similares. Además de estas definiciones, que parten de la institución y pensando que se tienen los elementos esenciales de hardware, software y sistemas de back-up, el repositorio, para considerarse tal, debe:

- ser capaz de brindar un servicio sencillo para el autoarchivo y la ingesta (o ingreso) de contenidos con diferentes tipologías;
- gestionar los derechos de las obras expuestas;
- organizar en colecciones y administrar los contenidos archivados, agregando información especializada, atendiendo a las diferentes tipologías; esto presupone manejar internamente diferentes esquemas de metadatos;
- manejar diferentes flujos de trabajo de acuerdo a la colección y a los permisos o roles de los distintos usuarios (investigadores, administradores, gestores);
- identificar de manera unívoca los contenidos: utilizar algún identificador persistente que asegure la localización permanente del material;
- conservar los contenidos a perpetuidad (o durante el plazo que establezcan las políticas) y que en esta conservación se realicen las transformaciones necesarias para que cualquier usuario final pueda interpretar y comprender los contenidos;
- permitir la colecta de sus registros a través de OAI-PMH (función de *data provider*); esto supone trabajar con un esquema básico de metadatos Dublin Core;
- asegurar la integridad de los contenidos (por ejemplo, usar *checksum*);
- interoperar con otros sistemas universitarios de gestión de documentos,

utilizando distintos protocolos: Sword, OpenURL, OAI-ORE, etc.;

- ofrecer interfaces amigables tanto para el ingreso de material, como para la consulta y la difusión de los contenidos.

Básicamente, las tecnologías que se utilicen en el repositorio y los responsables del mismo deberán asegurar los contenidos digitales a perpetuidad, no sólo salvaguardados sino identificables de manera inequívoca, sin alteraciones y legibles por su comunidad. Pensar al repositorio por sus funciones permitirá comprender de manera más eficiente los modelos que se exponen en el capítulo 3 y las formas de evaluación propuestas en el capítulo 4.

Bibliografía del capítulo

Arms, W. Y.; Blanchi, C.; Overly, E. A. (1997). "An Architecture for Information in Digital Libraries". *D-Lib Magazine*, February. Recuperado el 17 de septiembre de 2011, de http://www.dlib.org/dlib/february97/cnri/02arms1.html.

Arms, W. Y. (2000). *Digital Libraries*. Cambridge. MA., MIT Press.

Bawden, D.; Rowlands I. (1999a). "Digital Libraries: Assumptions and Concepts". *Libri*. 49, p. 181-191.

Bawden, D.; Rowlands, I. (1999b). "Digital Libraries: a conceptual framework". *Libri*. 49, p. 192-202.

Borgman, C. L. (1999). "What are digital libraries? Competing visions". *Information Processing & Management*, 35 (3), p. 227-243.

Digital Library Manifesto (2010). "Digital Library Manifesto"(2007). Recuperado el 23 de junio de 2014, de http://www.ifla.org/publications/iflaunesco-manifesto-for-digital-libraries.

Drabenstott, K. M. (1994). *Analytical Review of the Library of the Future*. Editor: Council on Library Resources Washington. Recuperado el 9 de junio de 2014, de http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.116.9658

Lesk, M. (1997). *Practical digital libraries: Books, bytes, and bucks*. San Francisco: Morgan Kaufmann.

Méndez Rodríguez, E. M. (2008). "Metadatos para la preservación digital: PREMIS". *VIII Workshop Rebiun*, Murcia, España, 20 y 21 de octubre.

National Science Foundation (NSF) (1999). *Digital Libraries Initiative: Available Research*, US Federal Government: http://dli2.nsf.gov/dlione/ citado en: Seadle, M., Greifeneder, E. (2007). "Defining a digital library". *Library Hi Tech* 25 (2). p. 169-173

Schwartz, C. (1999). "A Working Definition of Digital Library". Digital Library Federation. Recuperado el 29 de abril de 2011, de http://www.clir.org/diglib/dldefinition.htm.

Schwartz, C. (2000). "Digital Libraries: An Overview". *The Journal of Academic Librarianship*, Volume 26, Number 6, p. 385-393. Recuperado el 23 de junio de 2014, de http://home.kku.ac.th/hslib/malee/412725/document/DLOverview.pdf.

Tramullas Saz, J. (2002). "Propuestas de concepto y definición de la biblioteca digital". *Actas de las III Jornadas de Bibliotecas Digitales: (JBIDI'02)*. Madrid, España.

Capítulo 3 | Modelos para un Repositorio Institucional

Síntesis: En este capítulo se presentan algunos de los modelos propuestos para una biblioteca digital. En el desarrollo de la exposición se usará copiosamente el término biblioteca digital ya que es el que se utilizaba durante el desarrollo de algunos de estos modelos. No obstante, de acuerdo a los deslindes hechos en el capítulo 2 se afirma, más allá de la terminología, que su utilidad es la misma para representar un repositorio institucional, tal es el caso de SEDICI, que será el objeto de la evaluación posterior. Se presentarán sólo tres modelos, los que se completarán con los modelos de evaluación del capítulo 4. Se ha realizado esta selección atendiendo a la significación de los modelos elegidos, sus referentes y cómo los modelos presentados pueden proporcionar una base teórica que sustente en forma adecuada la propuesta experimental del capítulo 5.

Modelos de representación para un RI

Entre los muchos existentes, los tres modelos de representación de un RI elegidos para su comentario y análisis en esta tesis son:

- Modelo Bawden y Rowlands (1999)
- Modelo de Delos (2004-2007)
- ISO 14721 (2012)

Para entender cabalmente un repositorio institucional se requiere no sólo captar su concepto y caracterización, sino que también es necesario relacionarlo con los conceptos teóricos de modelo y evaluación. El estudio combinado de estas dos facetas lleva a comprender la combinación precisa: teoría de los modelos, modelos de evaluación y por último lleva también a la comprensión del modelo de evaluación del repositorio.

Un modelo es un conjunto de elementos esenciales que logra representar un aspecto de la realidad; por lo general, deriva de la teoría pero también de la realidad

experimentada o de la simple abstracción a modo de descripción verbal, visual, lógica o matemática. Un buen modelo es aquel que representa con el suficiente detalle (para los fines específicos) un sistema dado y permite inferir resultados, acciones e incluso paradigmas tras su análisis. En el caso de los repositorios, se entiende por modelo el conjunto de elementos esenciales que conforman una representación del mismo en cuanto a las funciones esperadas.

Para De Andrade (2006), existen diversos tipos de modelos, entre los cuales se puede aludir a los explicativos, físicos, formales, teóricos, analógicos, simbólicos, taxonómicos, exploratorios, descriptivos, predictivos, normativos, cuantitativos, cualitativos, experimentales, lineales, duales o cibernéticos.

El desarrollo de la investigación sobre el amplio campo de las bibliotecas digitales, en el cual han participado grupos de investigación muy diversos, ha traído como consecuencia la existencia de, por lo menos, dos perspectivas diferentes, que se han visto reflejadas en definiciones, concepciones y modelos distintos de lo que es una BD. De cualquier modo, la identificación de los elementos nucleares subyacentes en las definiciones facilita la comparación entre las mismas y, en este sentido, se ha adoptado una propuesta que muestra modelos que se concentran en las estructuras conceptuales, las características, la arquitectura o componentes, y los objetivos o funciones de los RI.

Marco conceptual: Modelo de Bawden y Rowlands

El propósito del modelo de Bawden y Rowlands es proponer —y parcialmente validar— un marco conceptual de alto nivel para ayudar a la mejor comprensión de la idea de "biblioteca digital". El marco se basa en una serie de términos esenciales para comprender el concepto y sus componentes, que los mismos autores explicaran en "Digital Libraries: assumptions and concepts". Allí presentan una extensión de algunas de las definiciones ya comentadas en el capítulo 2, por ejemplo el punto de vista de las tres vías de Borgman (1999):

1) Las BD como contenido, colecciones y comunicación.

2) Las BD como instituciones o servicios.

3) Las BD como bases de datos.

Bawden y Rowlands se basan en el trabajo de Yates (1989), que resalta tres aspectos cruciales de las BD: documentos, trabajo y tecnología. Las BD almacenan y dan acceso a documentos, los documentos son creados y mantenidos usando tecnologías, las cuales para su desarrollo requieren del tiempo de las personas, y ambas partes — documentos y tecnologías— son desarrolladas en la institución para apoyar el trabajo de los investigadores que usan la BD y el del propio staff que provee la posibilidad de lectura y otros servicios agregados. Esta propuesta enfatiza la interconexión y asimismo la importancia de las colecciones gestionadas para una comunidad designada, insistiendo en que la tecnología, o la infraestructura *per se* no constituye una BD. Los tópicos más importantes de hoy en día referidos a los RI, como interoperabilidad, escalabilidad, búsquedas semánticas y preservación, pueden verse como una coordinación de esas tres esferas.

Bawden y Rowlands analizan el modelo de Yates y renombran cada una de las tres esferas previamente mencionadas como: *informacional* (documentos), *sistemas* (tecnología) y *social* (trabajo); desglosan los temas de cada una, logran un agrupamiento más significativo de los factores que hacen al área y ponen de manifiesto dónde están los solapamientos y la interacción:

1. ***Dominio social***: factores humanos, factores organizacionales, factores de gestión del RI, políticas, impacto sobre la cadena de transferencia de la información.
2. ***Dominio informacional***: organización del conocimiento para su fácil localización (por ejemplo, metadatos), impacto sobre la cadena de transferencia de la información.
3. ***Dominio del sistema***: factores humanos, factores del sistema, organización del conocimiento para su fácil localización (agentes), impacto sobre la cadena de transferencia de la información.

Los autores indagan qué temas de investigación se abordan en los tres dominios, y así puede verse que en el dominio social destacan el consenso que parece emerger en los entornos de BD, que abarcaría dos componentes distintos: aprender a acceder, evaluar y utilizar los recursos de información, y aprender a dominar y construir sobre las ideas contenidas en esos recursos. Aunque se trata de un trabajo realizado en los

albores de las BD, reconoce ya el impacto de éstas en las formas de trabajo: la interacción y la ubicuidad de las BD, el problema de pensar métodos para la adquisición, la selección, el acceso y el enorme problema de la preservación digital, especialmente en el enfoque de materiales nacidos digitales y que por distintos cambios tecnológicos pueden quedar rápidamente obsoletos, así como el problema de derechos en las migraciones y las emulaciones, de las que se hablará luego. Las transformaciones necesarias de los materiales digitales para mantener su accesibilidad y comprensión implican, además, la necesidad de verificar la integridad de la información, de autenticar los cambios y, naturalmente, de dejar traza de ellos y de sus responsables: todo esto fuerza a un sistema de permisos y, por tanto, de comunidades de usuarios diversos.

En el dominio informacional, los autores destacan las actividades de investigación vinculadas al acceso, la búsqueda, la exploración y la navegación y en estas actividades los metadatos juegan un rol central; también tratan la necesidad de evaluar los sistemas de recuperación de información y las métricas que debieran estudiarse (por ejemplo, precisión y *recall*) para medir la efectividad. En cuanto al dominio de los sistemas, los tópicos centrales pasan por la escalabilidad (la integración en redes locales e internacionales) y la interoperabilidad sintáctica y semántica.

Los autores tratan también los siguientes conceptos:

- *biblioteca electrónica*: un lugar físico donde los usuarios pueden acceder a servicios electrónicos, que es un modelo incremental desde la biblioteca tradicional;
- *biblioteca híbrida*: da acceso a información de un rango de medios y formatos, y
- *biblioteca digital*: basada en una institución y que no necesita tener una localización física.

A partir de estas consideraciones los autores se inclinan por la biblioteca digital, ya que consideran que tiene la posibilidad de alterar el circuito de publicación tradicional y sintetizan sus principales puntos de vista con el marco conceptual que puede verse en la figura 3.1.

Figura 3.1: Marco conceptual de una biblioteca digital

Fuente: Bawden y Rowlands (1999).

Para los autores, ese marco conceptual resumía muchos aspectos importantes referidos a la investigación en BD, vinculando ideas y prácticas y concibiendo la BD de manera dinámica, para proveer una herramienta útil para comprender el área. Es necesario tener en cuenta que aunque este marco conceptual es de 1999, su aporte ya prefigura todas las esferas internas y externas de los repositorios.

Modelo de DELOS

El comentario y análisis de este modelo se basa en un resumen del Manifiesto de DELOS (Candela, L. *et al.*, 2007), en el que se ofrecen las bases que fundamentan las bibliotecas digitales, identificando los conceptos que son su piedra angular y generando un modelo robusto que encapsula toda la riqueza de las perspectivas que conllevan las BD.

El Manifiesto reconoce una visión de las BD que es parangonable al concepto de "Espacio de Información" (Information Space), similar al que se establece en el campo del Computer Supported Collaborative Work (CSCW) y de "Inhabited Spaces" (Snowdon *et al.*, 2004). Este último es un concepto próximo a la idea de una BD ubicua, lo cual constituye un requisito para un sistema CSCW.

El Manifiesto se basa en numerosas investigaciones y experiencias, y establece un marco conceptual con tres sistemas que denomina marco de trabajo en tres niveles (Three Tier Framework), constituido por la **Biblioteca Digital** (Digital Library), el **Sistema de la Biblioteca Digital** (Digital Library System) y la **Gestión del Sistema de la Biblioteca Digital** (Digital Library Management System). El primero de los sistemas, DL (por sus siglas en inglés), es la organización que recoge, gestiona y preserva a largo plazo los contenidos digitales y la que también los ofrece a las distintas comunidades de usuarios. El sistema DLS es básicamente el sistema de software, la arquitectura requerida para brindar la funcionalidad de la biblioteca, y el DLMS es la plataforma: sistema operativo, bases de datos, interfaz de usuario que brinda la funcionalidad básica y también la posibilidad de integración con el software especializado. Mientras que el concepto de DL es abstracto, los de DLS y DLMS capturan realizaciones concretas de sistemas de software. En la figura 3.2 puede observarse como interactúan estos tres conceptos de acuerdo a DELOS.

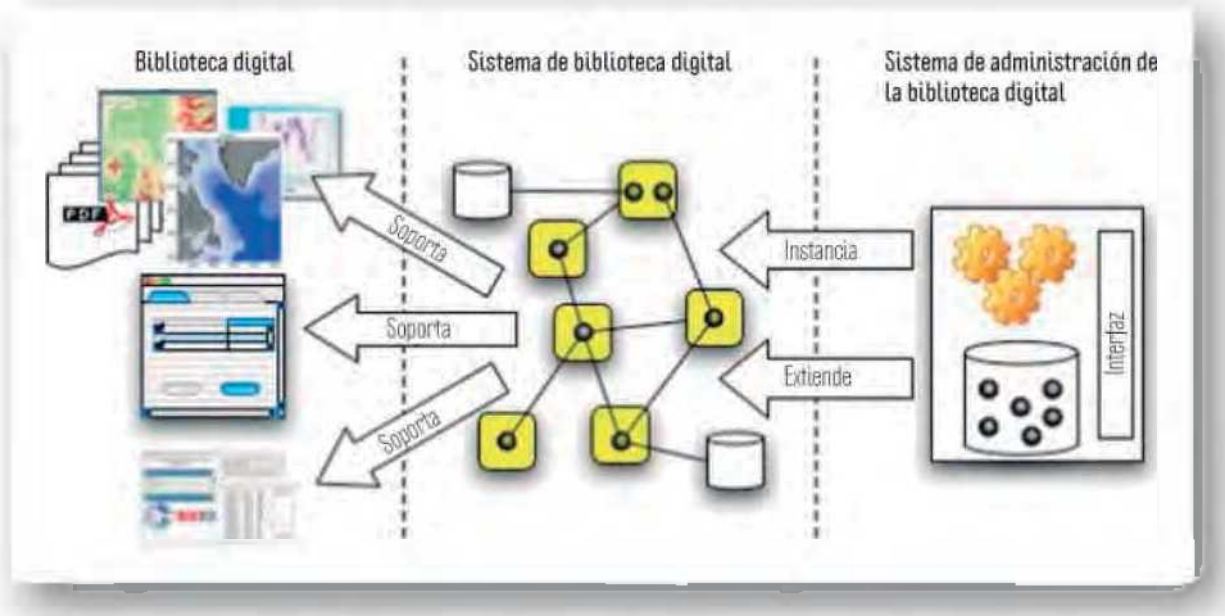


Figura 3.2: Arquitectura propuesta por DELOS para una biblioteca digital con sus tres niveles: DL, DLS y DLMS

Fuente: Candela, L. *et al.* (2007).

A pesar de la aparente riqueza y la diversidad de las actuales BD, hay sólo un pequeño número de conceptos básicos definido por todos los sistemas. Esos conceptos

proporcionan una base para las bibliotecas digitales. Cinco de ellos aparecen en la definición de la BD: el *contenido*, el *usuario*, la *funcionalidad*, la *calidad*, y la *política*; el sexto surge en la definición de DLS: la *arquitectura*. Los seis conceptos influyen en el marco conceptual establecido por DELOS. De los mencionados conceptos se daban las siguientes definiciones, las cuales se mantienen aquí sin alteración.

Contenido

El concepto de contenido abarca los datos y la información que maneja la BD y que pone a disposición de sus usuarios. Se compone de un conjunto de objetos de información organizados en colecciones. El contenido es un concepto general que se utiliza para agregar todas las formas de objetos de información que una biblioteca digital recopila, administra y difunde, e incluye objetos primarios, anotaciones y metadatos. Por ejemplo, los metadatos tienen un papel central en el manejo y uso de objetos de información, ya que proporcionan información esencial para su interpretación sintáctica, semántica y contextual.

Usuario

El concepto de usuario cubre los diversos actores (humanos o máquinas) que tienen derecho a interactuar con las BD. Las bibliotecas digitales conectan a los usuarios con la información y los apoyan en su capacidad para consumir y hacer un uso creativo de ella para generar nueva información. El de usuario es un concepto global que incluye todos los conceptos relacionados con la representación y la gestión de las entidades de "actores" dentro de una BD y a los distintos perfiles corresponden distintos permisos: administradores, lectores, creadores de información, entre otros.

Funcionalidad

El concepto de funcionalidad alude a los servicios que una BD ofrece a sus usuarios. Si bien las expectativas de servicios son muchas, mínimamente deben permitir incorporar nuevos objetos, buscar y navegar la información disponible. Lo deseable es que la funcionalidad cubra las necesidades de información de la comunidad a la que ofrece los servicios, lo que habitualmente se denomina "comunidad designada".

Calidad

El concepto de calidad representa los parámetros que pueden ser utilizados para caracterizar y evaluar el contenido y el comportamiento de una BD. La calidad puede estar asociada no sólo con cada tipo de contenido o funcionalidad, sino también con los objetos de información o servicios específicos. Algunos de estos parámetros son de naturaleza objetiva y pueden ser medidos de forma automática, mientras que otros son de carácter subjetivo y sólo se pueden medir a través de evaluaciones de los usuarios.

Política

El concepto de política representa el conjunto o conjuntos de condiciones, reglas y reglamentos que rigen la interacción entre la BD y los usuarios. Ejemplos de políticas incluyen el comportamiento permitido a los usuarios, la gestión de derechos, privacidad y confidencialidad. Las políticas son de diferente tipo y no todas se definen en la BD; por ejemplo, la BD puede definir la política de metadatos, pero no la de los materiales que se van a preservar, lo que puede ser resorte de decisión de la institución que alberga a la BD.

Arquitectura

El concepto de arquitectura se refiere a la entidad Sistema de Bibliotecas Digitales y representa un mapeo de la funcionalidad y el contenido ofrecido por una BD en el hardware y el software que la componen. Hay dos razones principales para tener la arquitectura como un concepto fundamental: i) las BD son formas complejas y avanzadas de los sistemas de información, y ii) la interoperabilidad entre BD es reconocida como un desafío para la investigación en este campo. Es indispensable, entonces, una arquitectura capaz de hacer frente a los dos problemas mencionados.

Los conceptos de contenido, usuario, funcionalidad y política comparten características similares en muchos aspectos y todos son conceptos que se refieren a entidades internas de una BD que puede ser detectada por el mundo exterior. La introducción del concepto de *recursos*, que subsume a los cuatro anteriores, nos permite razonar acerca de estas características en una manera consistente.

Los seis conceptos fundamentales mencionados (contenido, usuario, funcionalidad,

calidad, política y arquitectura), que se encuentran en el centro de listas de distribución, se deben considerar en conjunto con las cuatro formas principales en que los actores interactúan con los sistemas de BD: usuarios finales (creadores de información, consumidores y bibliotecarios), diseñadores que definen, personalizan y mantienen la BD alineada con la información y las necesidades funcionales de sus potenciales usuarios finales. Para lograrlo, interactúan con el DLMS proveyendo los parámetros de configuración personal y de contenidos. Los parámetros funcionales instancian aspectos de la funcionalidad de la BD que van a ser percibidos por los usuarios finales, incluyendo las características del formato del conjunto de resultados, el lenguaje de consulta, los perfiles de usuario, y el documento/modelo de datos utilizado. Los parámetros de configuración de contenido especifican los recursos de terceros explotados por la BD; por ejemplo, los repositorios de contenido, las ontologías, los esquemas de clasificación, los ficheros de autoridad y diccionarios geográficos. Los valores de estos parámetros configuran la forma en que la BD será presentada a los usuarios finales, ya que determinan la instancia particular de DLS. Por supuesto, estos parámetros no necesariamente tienen que ser fijados para toda la vida de la BD: pueden ser reconfigurados para permitirle responder a las expectativas cambiantes de los usuarios y los cambios en todos los aspectos de las políticas de contenido.

Los administradores seleccionan los componentes de software necesario para crear el DLS, reflejando las expectativas de los usuarios finales y los diseñadores, así como los requerimientos impuestos por la disponibilidad de recursos. Los administradores interactúan con el DLMS para proveer los parámetros de configuración de la arquitectura, como los componentes de software elegido. Su tarea es determinar los parámetros de configuración de la arquitectura que mejor se ajustan al DLS para asegurar el más alto nivel de calidad. Estos parámetros pueden ser cambiados a lo largo de la vida de la BD, pero su cambio puede ocasionar una funcionalidad diferente o niveles distintos de calidad. Los desarrolladores desarrollan los componentes de software de los DLS y DLMS para asegurar que los diferentes niveles y tipos de funcionalidad estén siempre disponibles.

El universo de las BD es complejo y la representación de sus múltiples elementos depende de la introducción de marcos conceptuales con distintos niveles de

abstracción, saber:

- ***Reference Model (Modelo de Referencia)***: consiste en un conjunto mínimo de conceptos unificados, axiomas y relaciones dentro del dominio de un problema en particular y es independiente de estándares específicos, tecnologías, implementaciones u otros detalles concretos. Las BD precisan obtener su correspondiente modelo de referencia a fin de consolidar la diversidad de aproximaciones dentro de un todo cohesivo y consistente, para ofrecer un mecanismo que permita la comparación de diferentes BD y proveer así una base común de comunicación entre las comunidades de BD.
- ***Reference Architecture (Arquitectura de Referencia)***: es un patrón de diseño que indica una solución abstracta para implementar los conceptos y las relaciones identificadas en el Modelo de Referencia. Puede haber múltiples arquitecturas de referencia que indiquen cómo diseñar un DLS construido según el Modelo de Referencia. Por ejemplo, se podría tener una arquitectura para BD que soportan localmente recursos federados y múltiples organizaciones, y otra para BD con aplicaciones especiales sobre los recursos propios.
- ***Concrete Architecture (Arquitectura Concreta)***: en este nivel, la Arquitectura de Referencia es actualizada, reemplazando los mecanismos delineados en la Arquitectura de Referencia con estándares concretos y especificaciones.

Los tres marcos de referencia son el resultado de un proceso de abstracción que ha tenido en cuenta los objetivos, los requisitos, las motivaciones y, en general, el mercado de la biblioteca digital, como se muestra en el lado izquierdo de la figura 3.3, y las mejores prácticas de investigación (lado derecho). La misma figura describe también los compromisos y las consideraciones de los tres niveles de abstracción propuestos por DELOS.

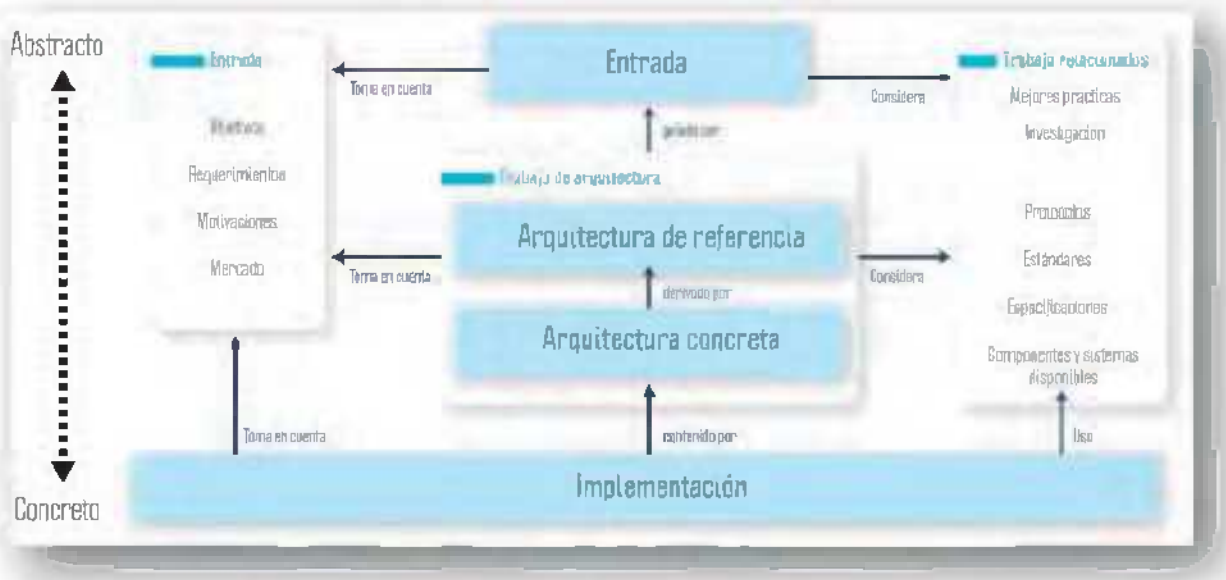


Figura 3.3: El universo de la biblioteca digital

Fuente: Candela, L. *et al.* (2007) (Imagen original en inglés).

Modelo de Referencia OAIS

En el año 2003 se publicó por primera vez la norma ISO:14721, denominada Space Data and Information Transfer Systems o ISO Reference Model. Open Archival Information System (OAIS), que modela las partes que componen un sistema de archivo de información, como un RI. Es un marco de trabajo conceptual, no es prescriptivo y sirve para identificar las características necesarias de un sistema de archivo de información constituido por personas y sistemas, que ha aceptado la responsabilidad de preservar información y hacerla disponible para una comunidad determinada (o "comunidad designada", según la norma), es decir, el grupo de usuarios capaces de comprender la información en cuestión.

El modelo se sitúa en un contexto de *Productores* (Producers) de información, *Consumidores* (Consumers) que buscan recuperar y usar esa información, y *Gestión* (Management), organización más amplia que alberga el Sistema de Archivo de Información (Archival Information System). Claramente, OAIS puede ser un modelo abstracto muy interesante a la hora de evaluar una BD o un RI, que, como ya se ha mencionado, es un Sistema de Archivo de Información.

En el modelo, una persona con una base adecuada de conocimientos (por ejemplo,

la capacidad de leer español si el contenido está en esa lengua) extrae y entiende la información transmitida por los datos. Se reconoce que la base de conocimientos necesaria para descifrar los datos y comprender la información puede no estar disponible entre los consumidores, y de ahí la necesidad de información adicional de la representación para cerrar esa brecha. El modelo introduce así la terminología de un objeto de datos (por ejemplo, un *bitstream*) que, cuando se interpreta utilizando la información de representación (por ejemplo, el estándar ASCII), se convierte en un objeto de información (por ejemplo, un archivo de texto).

El problema que surge de esta conceptualización es la naturaleza recursiva de la representación de la información, pues debe asumirse alguna base mínima de conocimientos entre los consumidores. La idea de una comunidad designada aparece así como forma de asegurar que esta base de conocimientos mínima se mantiene.

La norma del año 2003 fue revisada y se continuaron los lineamientos propuestos para la nueva versión: ISO 14721:2012, denominada Space Data and Information Transfer Systems — Open archival information system (OAIS) — Reference model.

Modelo de Información

El elemento central de la norma es el *Paquete de Información* o Information Package (IP), que se refiere al conjunto que conforman el objeto digital y todos sus metadatos. La versión 2012 de la norma define el IP de acuerdo a lo presentado en la figura 3.4.

Figura 3.4: Estructura del Paquete de Información según OAIS

Fuente: ISO 14721:2012.

La norma define el IP como un contenedor conceptual con dos tipos de información: de contenido y de preservación. La Información de Contenido (CI) es el objeto mismo que se desea mantener en el tiempo e incluye los Datos del Objeto (CDO) y su Información de Representación (RI). La Información Descriptiva de Preservación (PDI) debe brindar datos suficientes sobre la procedencia, el contexto, la referencia, la integridad y los derechos del objeto. La procedencia, más allá de describir la fuente, incluye los procesos que se han realizado sobre la información: su historia, cambios, versiones y responsables. El contexto muestra las relaciones con otras fuentes de información o contenidos. La referencia provee una identificación única del contenido. La integridad (o fijeza) provee una protección para que la información no sea alterada de manera intencional o no.

Los objetos de información se mueven a través del OAIS encapsulados. El paquete en su totalidad necesita ser "envuelto" por la Información de Empaquetado o Packaging Information (por ejemplo, un identificador) y, a su vez, el OAIS debe mantener la información descriptiva acerca del paquete para facilitar la búsqueda y

localización. De ahí que la figura 3.4 se complejice a lo largo de este trabajo para poner de manifiesto las partes constitutivas del paquete de información, así como la información descriptiva que debe guardar el OAIS para su búsqueda y recuperación.

El IP puede subclasificarse en tres subtipos, según su función en el proceso de archivo:

- **Archival Information Package (AIP)**: contiene, como mínimo, suficiente información de un objeto para garantizar la preservación a largo plazo. Busca mantener la mayor calidad posible de información descriptiva de preservación y de representación de los objetos representados o contenidos. Es la variante que preserva el OAIS y de las tres es la que tiene el DIP más detallado.
- **Submission Information Package (SIP)**: es el paquete que proviene del productor y que se va a incorporar al OAIS. Suele contener menos información que el AIP, su PDI es insuficiente y puede estar estructurado de otra manera que el AIP.
- **Dissemination Information Package (DIP)**: es el paquete que se entrega a un consumidor en respuesta a una solicitud. La información de empaquetado toma muchas formas, dado que los usos de OAIS son diversos; puede ser tan completo como los AIP a partir de los cuales se construye o ser sólo una breve descripción del paquete. Básicamente es una versión del AIP a la medida del consumidor.

Todas las variantes del paquete de información, por ejemplo, los SIP pueden incluir contenido destinado a dividirse en varios AIP y varios AIP pueden ser empaquetados dentro de un único DIP para su diseminación a los consumidores. Puede ser conveniente también almacenar varios AIP dentro de un AIP mayor.

Modelo Funcional

Las reflexiones de este apartado están basadas en el trabajo de De Giusti *et al.* (2012). La figura 3.5 muestra las seis entidades funcionales y los paquetes de información del modelo:

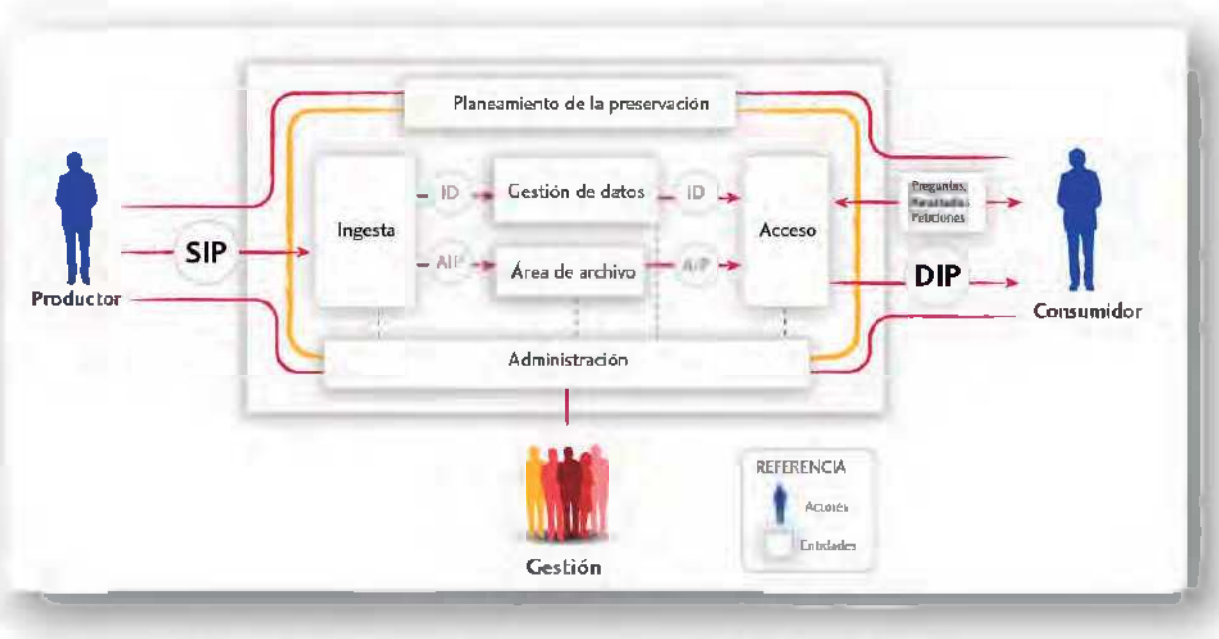


Figura 3.5: Entidades propuestas según el Modelo OAIS

Fuente: ISO 14721:2012.

Entidad Ingesta[11] (término original de la norma: ***Ingest***)

1. Provee los servicios y funciones para incorporar al archivo un SIP producido por los productores o bajo el control de la administración.
2. Realiza el aseguramiento de calidad/validación de los SIP.
3. Genera el AIP que cumple con los estándares de formato de datos y lo envía al *archival storage*.
4. Genera la información descriptiva y la envía al *data management*.
5. Coordina las actualizaciones en el *archival storage* y en el *data management* de la base de datos.

Entidad Almacén de Archivos (término original de la norma: ***Archival Storage***)

Provee los servicios y funciones para el almacenamiento, mantenimiento y recuperación de los AIP.

1. Recibe el AIP de la entidad *ingest* y lo almacena; gestiona las jerarquías de almacenamiento (por ejemplo, pone el AIP en un medio apropiado de almacenamiento); configura niveles especiales de servicio, seguridad y protección

[11] Si bien la RAE no reconoce este uso de la palabra, la literatura del área utiliza "ingesta" e "ingestión" y aquí se adopta este término.

(por ejemplo, back-ups); provee estadísticas de inventario, capacidad disponible, etc.

2. Transforma los datos que constituyen la información de empaquetado para reproducir el AIP en el tiempo.

3. Realiza una verificación de errores; provee un mecanismo estándar para el seguimiento y verificación de la validez de los datos; provee un mecanismo de recuperación ante desastres; provee un mecanismo de duplicación de los contenidos en un lugar físico separado.

4. Provee copia de los AIP almacenados a la entidad *access* a pedido.

Entidad Gestión de Datos (término original de la norma: ***Data Management***)

1. Provee los servicios y funciones para el mantenimiento de la base de datos de información descriptiva e información del sistema.

2. Recibe solicitudes de la entidad *access* y genera un conjunto de resultados.

3. Recibe pedidos de las entidades *ingest*, *access* y *administration* y genera reportes.

4. Actualiza la base de datos con información descriptiva que recibe de *ingest* y con información del sistema y revisiones actualizadas que recibe de *administration*.

Entidad Administración (término original de la norma: ***Administration***)

1. Provee los servicios y funciones para la operación global del OAIS.

2. Negocia la política y los acuerdos con los productores.

3. Monitorea la funcionalidad del OAIS: gestiona su configuración; controla los cambios de la configuración y mantiene su integridad y trazabilidad. Audita las operaciones del sistema, performance y uso. Envía reportes al *data management* y recibe reportes de esa entidad. Sintetiza todos los reportes y provee información sobre performance del OAIS e inventario, y envía esta información a *preservation planning* para establecer políticas y estándares. Recibe los paquetes de migración para *preservation planning*.

4. Recibe los pedidos de cambio, procedimientos y herramientas para la actualización del archivo.

5. Realiza la actualización de la información archivada, envía un pedido de

diseminación a *access*, actualizando los contenidos de los DIP y resuministrando los SIP a *ingest*.

6. Provee mecanismos para restringir/permitir acceso a los elementos del archivo.
7. Es responsable de enviar información para establecer estándares y políticas. Desarrolla políticas de gestión de archivo por jerarquías, incluyendo políticas de migración. Es responsable de la recuperación ante desastres.
8. Verifica que los AIP y SIP suministrados sigan las especificaciones necesarias. Verifica el PDI según los usos de la comunidad designada.
9. Revisa periódicamente los contenidos del archivo para determinar si los datos están disponibles.
10. Crea/mantiene/borra las cuentas de acceso de los consumidores.

Entidad Planeamiento de la Preservación (término original de la norma: ***Preservation Planning***)

1. Interactúa con los consumidores y productores de archivos. Proporciona reportes, alertas de requisitos y estándares independientes. Monitorea la comunidad designada en busca de cambios en los requerimientos.
2. Monitorea la tecnología: identifica tecnologías que pueden causar obsolescencia, estándares y plataformas, para seguir la aparición de algunas nuevas y declinar las antiguas.
3. Desarrolla y recomienda estrategias de preservación y estándares, que envía a *administration*.
4. Desarrolla nuevos IP y planes de migración y prototipos, para implementar políticas y directivas de administración de IP.

Entidad Acceso (término original de la norma: ***Access***)

1. Proporciona una interfaz única de usuario para el acceso a la información de los archivos. Tiene tres categorías: los *query requests*, los *result sets* y los *report requests*.
2. Acepta los requerimientos de los paquetes de diseminación recuperados de los AIP de la entidad *archival storage* y transmite un *report request* al *data management* generando un DIP.

3. Entrega las respuestas en línea y fuera de línea de los consumidores.

Existen, además, servicios comunes que soportan a las seis entidades, que, aunque no aparecen en la figura, incluyen sistemas operativos, servicios de red y seguridad. Si bien buena parte del modelo funcional podría producirse extrayendo los procesos y los flujos de trabajo de las bibliotecas tradicionales y los archivos, lo distintivo de este modelo se asienta en el énfasis en dos funciones principales: la preservación de la información y la garantía de acceso a la misma.

En la ya mencionada figura 3.5 se observan claramente los tres actores y las seis entidades. El proceso puede iniciarse cuando el productor (un actor) suministra el recurso (paquete de entrada) o SIP a través de *ingest*, que luego se convierte en AIP y concluye en la entidad de *archival storage*. El flujo puede continuar cuando el consumidor busca información en el sistema, que es entregada como un DIP a través de la entidad *access*, ya que ha sido preservada en el sistema previamente. Al mismo tiempo, los datos relacionados con los artículos y el repositorio mismo se mantienen organizados a través de la entidad *data management*. Luego, la entidad *administration,* adjunta a la gestión, es decir, los administradores y el responsable del repositorio, se relaciona con las secciones de *data management* y *preservation planning*. Esto permite una gestión estructural y también ayuda a mantener los AIP a lo largo del tiempo. El módulo de *preservation planning* desarrolla estrategias y normas de preservación, monitorea las últimas novedades y avances en el campo, y los cambios en la comunidad designada, para que toda la información nueva que se solicite se pueda adjuntar a los AIP correspondientes.

Bibliografía del capítulo

Borgman, C. L. (1999). "What are digital libraries? Competing visions". *Information Processing & Management*, 35 (3), p. 227-243.

Candela, L.; Castelli, D.; Pagano, P.; Thanos, C.; Ioannidis, Y.; Koutrika, G.; Seamus Ross; Hans-Jörg Schek; Heiko Schuldt (2007). "Setting the Foundations of Digital Libraries. The DELOS Manifesto". *D-Lib Magazine*, 13 (3-4).

Consultative Committee for Space Data Systems (2012). *Recommendation for Space Data System Practices: Reference Model for an Open Archival Information System (OAIS) : Recommended Practice*. Washington, DC, USA. Recuperado el 11 de junio de 2014, de http://public.ccsds.org/publications/archive/650x0m2.pdf.

De Andrade Matins, G. (2006). "Hablando sobre teorías y modelos en las ciencias contables". *Actualidad contable FACES*, 9, (13) p. 42-53.

De Giusti, M. R.; Lira, A. J.; Villarreal, G. L.; Texier, J. D. (2012). "Las actividades y el planeamiento de la preservación en un repositorio institucional". *II Conferencia Internacional de Acceso Abierto, Comunicación Científica y Preservación Digital (BIREDIAL)*, Colombia. Recuperado el 29 de junio de 2014, de http://sedici.unlp.edu.ar/handle/10915/26045.

ISO 14721:2012 (2012) Space data and information transfer systems — Open archival information system (OAIS) — Reference model. Recuperado el 26 de junio de 2014, de http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=57284.

Snowdon, D. N.; Churchill, E. F.; Frecon, E. (eds). (2004). *Inhabited Information Spaces living with your data*. Springer, London.

Yates, J. (1989). *Control Through Communication: The Rise of System in American Management*, JHU Press.

Capítulo 4 | Tendencias y modelos de evaluación para los repositorios institucionales

Síntesis: Del mismo modo que en el capítulo 3, en este se relatan modelos de evaluación que facilitan la comprensión de la elección experimental, particularmente en los aspectos más concretos, a saber: cuál será el objetivo de la evaluación, sobre qué aspectos/recursos se realizará la evaluación, qué se medirá, cómo se evaluarán los resultados obtenidos (comparaciones, patrones, contrastes). Como podrá observarse estos modelos tienden a encontrar una propuesta de evaluación que resulte objetiva y de la cual se puedan obtener resultados que orienten sobre el estado de situación y por lo tanto conduzcan a acciones de mejora. Los modelos analizados en este capítulo también son tres: 1) Saracevic y Lisa Covi (2000), Saracevic (2001), Saracevic (2004); 2) Fuhr y otros (2001) y 3) Fuhr y otros (2007).

Saracevic y Lisa Covi (2000), Saracevic (2001)

Los autores de este modelo analizan los constructos, el contexto y los criterios empleados en las bibliotecas digitales. Se plantean interrogantes como ¿por qué hay que evaluarlas? ¿con qué objeto? ¿qué se debe evaluar, a qué nivel, y con qué criterios?

Partiendo de la visión de Borgman, ya reseñada en el capítulo 3, los autores hablan de la existencia de múltiples comunidades vinculadas a la práctica en las BD, pero se centran en dos de ellas: la comunidad de la investigación y la comunidad de la práctica, pues cada una de ellas tiene una definición diferente de la BD y esto afecta la naturaleza conceptual de la evaluación.

Saracevic rescata tres estudios importantes vinculados a la comunidad de la investigación, y vale aclarar que, si bien reconoce la importancia de proyectos referidos a BD en otros países, se centra en la investigación realizada en los Estados Unidos. Estos tres estudios son:

1) La evaluación que tuvo lugar en el marco del Proyecto de la Biblioteca Digital de Alejandría (ADL) en la Universidad de California, Santa Bárbara (Hill *et al.*, 2000). El enfoque incluía una serie de estudios de usuario que comprendía diferentes comunidades y se concentraban en diferentes características de diseño relacionadas

con su usabilidad y funcionalidad. Los resultados sirvieron como base para especificar una lista de requerimientos para las nuevas interfaces. Se concentró en los usuarios y sus interacciones a través de la interface, teniendo como principales criterios la usabilidad y la funcionalidad.

2) En el marco del proyecto DLI-1 en la Universidad de California en Berkeley, se realizaron una serie de entrevistas con los futuros usuarios (Schiff *et al.*, 1997). Se usó como marco la teoría sociológica de Pierre Bourdieu (1990), sobre las relaciones entre los agentes individuales y los campos de orientación conductual. Se arribó a la conclusión de que *"el investigar el escenario social para el cual se pretende implementar una biblioteca digital provee de una comprensión valiosa. Los criterios para el estudio de los usuarios fueron el entorno social y sus acciones"*. Sin embargo, no queda claro que la teoría del "habitus" de Bourdieu pueda ser inmediatamente aplicada y usada para evaluar BD.

3) En el proyecto DLI-1[12] en la Universidad de Illinois, los investigadores académicos estudiaron la forma en que los lectores usaban artículos de publicaciones científicas tanto en entornos impresos como digitales; cómo los lectores *"movilizaron el trabajo (...) a medida que identifican, recuperan, leen y usan el material de los artículos de su interés"* (Bishop, 1999). Los criterios fueron el trabajo y el uso de material recuperado por parte de los usuarios.

Estos tres proyectos u otros similares sobre el comportamiento de los usuarios en relación a las BD y a la información en general son útiles para proponer nuevos criterios para la educación de los usuarios y distintos requerimientos para las BD, pero no son sistemáticos y no ofrecen medidas objetivas.

En relación a la comunidad de la práctica el enfoque se centra en la construcción de bibliotecas operativas, funcionales, el mantenimiento y operación de ellas y los servicios provistos a los usuarios. Se destacan al menos tres proyectos en este sentido:

1. El proyecto Perseus (Marchionini, Crane, 1994), dedicado a proveer información sobre el mundo griego; la evaluación tomaba en cuenta como criterios: 1) aprendizaje, 2) enseñanza, 3) sistema (desempeño, edición, interfaz) y 4) contenido (amplitud,

[12] Este proyecto ya mencionado de la NSF tuvo seis universidades participantes en la fase 1: la Universidad de California en Berkeley y en Santa Bárbara, la Universidad Carnegie Mellon, la Universidad de Illinois, la Universidad de Michigan y la de Stanford.

precisión).

2. El proyecto Peak (Keefer Riva, 2001), dedicado a evaluar el uso y una amplia variedad de aspectos económicos que involucró no sólo a bibliotecas sino a la editorial Elsevier; entre los parámetros de evaluación incluía: 1) acceso (en diferentes grupos de usuarios), 2) precio y 3) ingresos y costos. Este proyecto se centró en la eficiencia como factor preponderante de la evaluación.

3. El proyecto MESL (Cornell University, 1999), que involucró a varias universidades, relacionado con el uso educativo de imágenes digitalizadas de museos. Básicamente, se encuestó a los usuarios, diseñadores, desarrolladores y operadores para la evaluación. Los criterios para las preguntas incluyeron la funcionalidad (facilidad/dificultad de búsqueda), las necesidades de capacitación, la integración con otros servicios, el desarrollo técnico, el soporte, pero la evaluación no fue formal.

Saracevic relata varios esfuerzos de evaluación más, sobre todo para poner a la vista los criterios de evaluación utilizados y tras esto se inclina por una evaluación orientada a sistemas, ya que se trata, en definitiva, de sistemas de información. La evaluación en estos casos siempre está ligada a algún aspecto del desempeño, respondiendo, en un principio, desde este punto a la necesidad de evaluación.

Para establecer un vocabulario y conceptos comunes, Saracevic define un sistema como un conjunto de elementos en interacción. Un sistema creado por el hombre, tal como una biblioteca digital, tiene un aspecto adicional: los elementos, o componentes, interactúan para desempeñar ciertas funciones o procesos para alcanzar objetivos dados. Además, existe en un entorno, o una serie de entornos (que también pueden ser considerados como sistemas, y alguien podría pensar en ellos como contextos), e interactúa con ellos.

En este contexto, la evaluación significa una valoración del desempeño o del funcionamiento de un sistema, o parte del mismo, en relación a cierto objetivo. El desempeño puede ser evaluado en cuanto a:

1. ***Efectividad***: ¿cuán bien desempeña un sistema (o cualquiera de sus partes) aquello para lo que fue designado?
2. ***Eficiencia***: ¿a qué costo (los costos pueden ser financieros o involucrar tiempo o esfuerzos) lo realiza?
3. ***Ambos***: una combinación de ambos (por ejemplo costo-efectividad).

Una evaluación debe especificar cuál de los tres aspectos será evaluado. En el mencionado trabajo se explicita la evaluación de la efectividad. Para un mismo sistema, la evaluación puede ser realizada en diferentes niveles, en relación con diferentes elecciones de objetivos y usando una variedad de métodos, y puede ser orientada hacia distintas metas y audiencias. Para ser considerada una evaluación en regla, debe satisfacer ciertos requerimientos:

1. ***Constructo***: ¿Qué evaluar? ¿Qué significa en concreto una biblioteca digital? ¿Qué engloba? ¿Qué elementos (componentes, partes, procesos) se deben involucrar en la evaluación?
2. ***Contexto de evaluación***: Selección de una meta, marco, punto de vista o nivel(es) de evaluación. ¿Cuál es el nivel de evaluación? ¿Qué se considera crítico para un nivel dado? En última instancia, ¿qué objetivos deben seleccionarse para ese nivel?
3. ***Criterios que reflejen el desempeño en relación a los objetivos elegidos***: ¿En qué parámetros del desempeño conviene concentrarse? ¿Qué dimensión o característica se debe evaluar?
4. ***Medidas que reflejen los criterios elegidos para registrar el desempeño***: ¿Qué medidas específicas deben usarse para un criterio dado?
5. ***Metodología para realizar la evaluación***: ¿Qué instrumentos de evaluación se deben usar? ¿Qué muestras? ¿Qué procedimientos se deben usar para la recolección de datos, y cuáles para el análisis de los datos?

La especificación clara de cada uno de estos niveles es un requisito para cualquier evaluación de una biblioteca o repositorio digital. La primera pregunta (¿qué evaluar?) Saracevic la responde atendiendo definiciones ya vistas y llega a la conclusión de que los elementos pueden ser:

- colecciones digitales, recursos;
- selección, recopilación, propiedades, medios;
- distribución, conexiones, vínculos;
- organización, estructura, almacenamiento;
- interpretación, representación, metadatos;
- administración;

- preservación, persistencia;
- acceso;
- redes físicas;
- distribución;
- interfaces, interacción;
- búsqueda, recuperación;
- servicios;
- disponibilidad;
- asistencia, referencia;
- uso, usuarios, comunidades;
- seguridad, privacidad, políticas, aspectos legales, licencias;
- administración, operaciones, personal;
- costos, economía;
- integración, cooperación con otros recursos, bibliotecas o servicios.

Claramente, de entre todos esos elementos hay que elegir qué evaluar. También es preciso establecer un contexto para la evaluación, decidir a qué nivel evaluar: básicamente, elegir un par, un elemento elegido para ser evaluado y un elemento elegido sobre su desempeño. Saracevic divide las evaluaciones en siete niveles: los tres primeros centrados en el usuario, los tres últimos en el sistema y el del medio, la interface entre ambos; para cada nivel sugiere unas preguntas sobre el desempeño.

Los estudios centrados en los usuarios (nivel social, institucional e individual) vinculados a las necesidades y demandas, resultan difíciles de objetivar por razones de la propia diversidad de los objetivos individuales y/o comunitarios y esto dificulta la evaluación, en opinión de Saracevic. Las preguntas del nivel intermedio son las esperadas: qué tan bien soporta la interfaz el acceso, la búsqueda, la navegación, exploración y ¿cómo interactúa con la BD?

Las preguntas de desempeño para los ejes centrados en el sistema (ingeniería, procesamiento y contenido) parecen más fáciles de replicar y generalizar; las preguntas que plantea son las siguientes:

Ingeniería: ¿Qué tan bien se desempeñan el hardware, las redes y las configuraciones relacionadas?

Procesamiento: ¿Qué tan bien se desempeñan los procedimientos, técnicas,

algoritmos, operaciones y demás?

Contenido: ¿Qué tan bien están seleccionados, representados, organizados, estructurados y administrados la colección o los recursos información? Aunque esto también es bastante sistemático, las preguntas relacionadas son "¿qué tan bien?", "¿para quién?" y "¿con qué propósito?".

En cuanto a los criterios para la evaluación, el autor plantea que, hasta entonces, en las evaluaciones se elegía un nivel (por ejemplo, individual) y el criterio más prominente (por ejemplo, la usabilidad); también insiste en evitar las dicotomías e intentar hacer que los enfoques centrados en unos y otros aspectos trabajen bien juntos. Los criterios que describe son los tradicionales de las bibliotecas sobre colección, información, uso y estándares; algunos de los tradicionales para RI: relevancia, índice, búsqueda, etc.; y criterios tradicionales para las interfaces: usabilidad, funcionalidad, navegabilidad, exploración, ayudas.

El autor advierte sobre la necesidad de que en la evaluación se tenga como objetivo la uniformidad para el acceso y el uso (interoperabilidad) y otro punto que señala como crítico es la persistencia de la información.

Saracevic (2004)

En este trabajo, Saracevic sostiene que la literatura sobre evaluación de BD se divide en dos tipos:

1. ***Metaliteratura***: comprende trabajos que sugieren conceptos de evaluación, modelos, aproximaciones, metodologías, pero que no contienen datos.
2. ***Objeto***: trabajos que reportan evaluaciones y que contienen datos.

El trabajo aquí reseñado es justamente de esta última especie, pues recoge una extensa bibliografía y los reportes de evaluaciones los estructura de acuerdo a cuatro aspectos:

1. ***Constructo de la evaluación***: qué se va a evaluar: elementos, componentes, partes.
2. ***Contexto***: selección de objetivo, marco conceptual, punto de vista o nivel de la

evaluación.

3. ***Criterio***: sobre qué parámetros de performance se concentra la evaluación; qué dimensión o características se van a evaluar.
4. ***Metodología***: qué medidas e instrumentos de medidas se van a usar. Qué procedimientos se usarán para la recolección y el análisis de los datos.

A continuación, se realiza una revisión de estos aspectos para una mejor comprensión de la evaluación posterior, objeto de esta tesis, y de los resultados obtenidos en ella.

Constructos

Saracevic se refiere aquí a los ítems a evaluar, es decir, al sujeto de la evaluación. Así, esto puede estructurarse en dos grandes constructos iniciales a evaluar: la entidad en sí y los procesos. Esquemáticamente:

1. Una BD específica, como una entidad en su conjunto.
   a. Evaluación de bibliotecas digitales específicas: Perseus, Hal, RePEc, etc.
   b. Evaluación de algunos aspectos operacionales, por ejemplo: capacidades de búsqueda y navegación; materiales didácticos; interface, etc., en una biblioteca dada. Saracevic menciona, por ejemplo, que una BD que apoya efectivamente a los usuarios académicos debe abordar las conductas y actividades de los usuarios que participan en la investigación. Mediante el uso de grupos de enfoque, entrevistas semi-estructuradas y cuestionarios, este estudio concluye que los usuarios académicos se verán beneficiados con las interfaces de usuario adaptables y flexibles que permiten una fácil navegación en un panorama de información compleja (Payettea, Riegerb, 1998).
   c. Evaluación de múltiples bibliotecas digitales. Por ejemplo, el proyecto SOUP (Jones *et al.*, 1999).
2. Procesos: varios procesos evaluados, pero sin relacionarlos con una BD específica, lo que hace difícil una posterior generalización. El autor expone una clasificación arbitraria.
   a. Evaluación de distintas representaciones para usar en las BD: por áreas, materias.
   b. Evaluación de distintas herramientas: generación de enlaces, método de

recuperación de imágenes, etc.

c. Evaluación de servicios: por ejemplo, referencia.
d. Evaluación de un esquema de evaluación: por ejemplo, Fuhr (2001), que se analizará posteriormente en este capítulo.
e. Estudio del comportamiento de los usuarios en relación a las BD: patrones de uso de *service logs*, preferencias de los usuarios en cuanto a búsquedas en bases a texto completo.

Contexto

Se refiere al marco de trabajo general para la evaluación, incluyendo aproximaciones, orientación, nivel y objetivos. Las BD son sistemas sociales, institucionales y técnicos complejos. Ninguna evaluación es capaz de abordar todos estos aspectos en conjunto, y por lo tanto deben usarse diferentes aproximaciones o estrategias para diferentes objetivos de evaluación. Saracevic presenta diferentes objetivos, ordenados en forma descendente de uso en los diferentes reportes que ha analizado:

a. ***Aproximaciones centradas en el sistema***: son las que prevalecen. Involucran estudios sobre aspectos de la performance. Incluye cálculos de la efectividad y/o eficiencia sobre características de algunos diseños específicos o componentes tecnológicos. Aplicadas en un número de estudios con resultados que pueden informar elecciones específicas en diseños u operaciones.

b. ***Aproximaciones centradas en las personas***: también ampliamente aplicadas. Involucran estudios del comportamiento con respecto a determinadas necesidades de información, como *seeking*, *browsing*, *searching* o *performance* para la completitud de determinadas tareas. Usadas en un número de estudios dan luz sobre el comportamiento humano, los requerimientos, las necesidades o dificultades encontradas. Generan, de manera indirecta, implicaciones para el diseño.

c. ***Aproximaciones centradas en la usabilidad***: involucran cálculos de diferentes características, particularmente en relación a los portales por los usuarios. Es un intermedio entre las dos primeras aproximaciones presentadas.

d. ***Aproximaciones etnográficas***: involucran estudios de cultura y costumbres en relación al ambiente de las BD. También, estudios de impacto en la comunidad por

la presencia de la BD. Se ha aplicado con éxito en unos pocos estudios, especialmente de impacto.

e. ***Aproximaciones sociológicas y económicas*** (costos, beneficios: PEAK): muy pobremente usadas.

Criterios

Saracevic se refiere aquí al estándar elegido para evaluar. Los criterios se usan para desarrollar medidas. Desde la década del 50, los sistemas de información usan la relevancia como criterio para evaluar colecciones, servicios o referencia. Es un criterio básico. Sin embargo se han tratado de desarrollar otras métricas. A continuación, se revisan algunos de los criterios más utilizados.

Usabilidad: no tiene una definición uniforme. Es un criterio muy general que cubre una gran cantidad de terreno e incluye muchos criterios específicos, por lo que podría decirse que es un metatérmino. ISO define la usabilidad *"como la extensión en la cual un producto puede ser usado por usuarios específicos para lograr metas específicas con efectividad, eficiencia y satisfacción en un contexto específico de uso"* (ISO, 9241-11: 1998). Este sería el marco en el cual la usabilidad es usada para la evaluación de BD. A continuación, se presenta una lista de los criterios de usabilidad específicos en varios estudios:

- **Content** (de un portal o sitio)
  - accesibilidad, disponibilidad
  - claridad
  - complejidad (organización, estructura)
  - cobertura
  - calidad
  - confiabilidad
- **Proceso** (para llevar a cabo tareas de búsqueda, exploración, navegación, etc.)
  - capacidad de aprendizaje
  - esfuerzo/tiempo
  - facilidad de uso
- **Formato**
  - atractivo

- consistente
- rotulado
- comunicacionalmente hábil

- **Mediciones globales**
  - satisfacción
  - éxito
  - relevancia, utilidad de los resultados
  - impacto
  - calidad de experiencia
  - barreras

Características de los sistemas: como las bibliotecas digitales son sistemas pueden usarse muchos criterios de evaluación de sistemas. Algunos conciernen a la performance de la tecnología, otros a la de algunos procesos o algoritmos usando la tecnología. Sirva de ejemplo el siguiente detalle:

- **Performance de la tecnología**
  - tiempo de respuesta
  - tiempo de procesamiento/velocidad
  - capacidad, carga
- **Performance del proceso/algoritmo**
  - relevancia de los resultados obtenidos
  - similaridad
  - funcionalidad
  - flexibilidad
  - tasa de errores
  - optimización
- **Performance del sistema global**
  - mantenibilidad
  - escalabilidad
  - interoperabilidad

Utilización

- uso de materiales
- estadísticas de uso

- quién usa, qué usa, cuándo lo usa
- razones de uso

Etnográficos

Estos criterios se refieren, en general, a la medición de impacto, como el caso del proyecto Perseus.

Metodologías

El rango de métodos usados en la evaluación de las BD es amplio, por ejemplo:

- encuestas
- entrevistas
- grupos, estudio de casos
- análisis de accesos
- experimentación
- análisis de uso
- análisis de registros

Fuhr *et al.* (2001)

Los autores proponen una aproximación integral a una BD ya que la evaluación no debe estar restringida a facetas o aspectos aislados, sino ser realizada con un punto de vista amplio. Siguiendo a Borgman (1999), consideran los dos puntos de vista ya mencionados, el de la comunidad de investigación y el tradicional de la comunidad de la biblioteca. La primera de estas comunidades se enfoca en el contenido de la información que va a ofrecerse y que debe satisfacer a los grupos de usuarios: objetos digitales, arquitectura y uso; en cambio, para la comunidad de la biblioteca, las BD son organizaciones que ofrecen servicios de información en formato digital y el eje es cómo adaptar las estructuras existentes a las nuevas tecnologías y desafíos.

Al variar las definiciones, o el acento de las mismas, varían las preguntas sobre qué evaluar, cómo medir, quiénes precisan de la evaluación y cuándo es apropiado realizarla.

¿Qué evaluar?: por ejemplo, mientras un bibliotecario puede estar interesado en las colecciones, un informático puede interesarse por los aspectos tecnológicos, más

allá del contenido.

¿Cómo medir?: quien diseña un sistema puede poner el acento en medidas vinculadas a la eficiencia (uso del recurso de cómputo), mientras que otros están interesados en la efectividad (precisión y *recall*).

¿Quiénes precisan los resultados de la evaluación?: en muchos casos la evaluación se necesita para la toma de decisiones: selección de software, por ejemplo, o un bibliotecario que, atendiendo a las suscripciones, quiere ver qué ofrece la BD a sus usuarios.

¿Cuándo evaluar?: las evaluaciones pueden realizarse en cualquier momento y lugar; por ejemplo, un desarrollador puede querer evaluar diversos métodos posibles para una función dada y para ello la evaluación en su lugar de trabajo puede ser suficiente.

Como Saracevic y Covi (2001), los autores consideran oportuno definir constructo, contexto, criterios y metodologías y pretenden definir algunas de estas cuestiones. Tienen en cuenta también las experiencias del D-Lib Working Group on Digital Library Metrics (DLib), dedicado a establecer métricas consensuadas para medir y comparar la efectividad de componentes tecnológicos en un ambiente distribuido. Finalmente, recogen la experiencia de DELOS, de la cual ya se ha hablado aquí, para proveer un marco de referencia y mencionan al Working Group 2.1, entonces responsable de proveer las bases para la evaluación de las bibliotecas digitales. Si bien DELOS cesó sus actividades a finales de 2007, aún es posible encontrar las publicaciones de sus *workshops* en línea. En este caso, los autores referencian al IV Workshop de DELOS, dedicado a evaluación de BD y organizado por el Working Group 2.1 (Borgman, Sølvberg, Kovácsv, eds., 2002).

El trabajo menciona también los parámetros tradicionales de evaluación de sistemas de cómputo vinculados a la performance: efectividad y eficiencia. Los autores hacen referencia a la iniciativa Text REtrieval Conference (TREC), con su valiosa colección de documentos dedicados a la evaluación de efectividad de recuperación (en términos de *recall* y precisión) en distintas colecciones. En el campo de la Interacción Persona-Computador (Human Computer Interaction, HCI) recogen bibliografía de estudios que investigan un rango de métodos de evaluación de usabilidad y de intercambios entre usuarios y sistemas de cómputo. Continuando con los trabajos relacionados, describen

brevemente una aproximación centrada en el usuario y un trabajo que diferencia usabilidad, utilidad y performance. Tras una extensa revisión de propuestas de evaluación, remarcan los cambios en las bibliotecas, los problemas de satisfacción y las medidas de performance que han llevado a nuevas métricas cuantitativas y cualitativas, más allá de las tradicionales de efectividad y eficiencia.

Considerando que una BD es una clase especial de sistema de información que integra la tecnología, las colecciones, las personas y el medio para el cual está construida, los autores resaltan la importancia de la integración de todos los elementos y opinan que las evaluaciones que ignoran las otras dimensiones, no son capaces de demostrar la validez de los resultados en relación a otros aspectos. A continuación presentan su aproximación holística para la evaluación de una BD, que puede observarse en la figura 4.1.

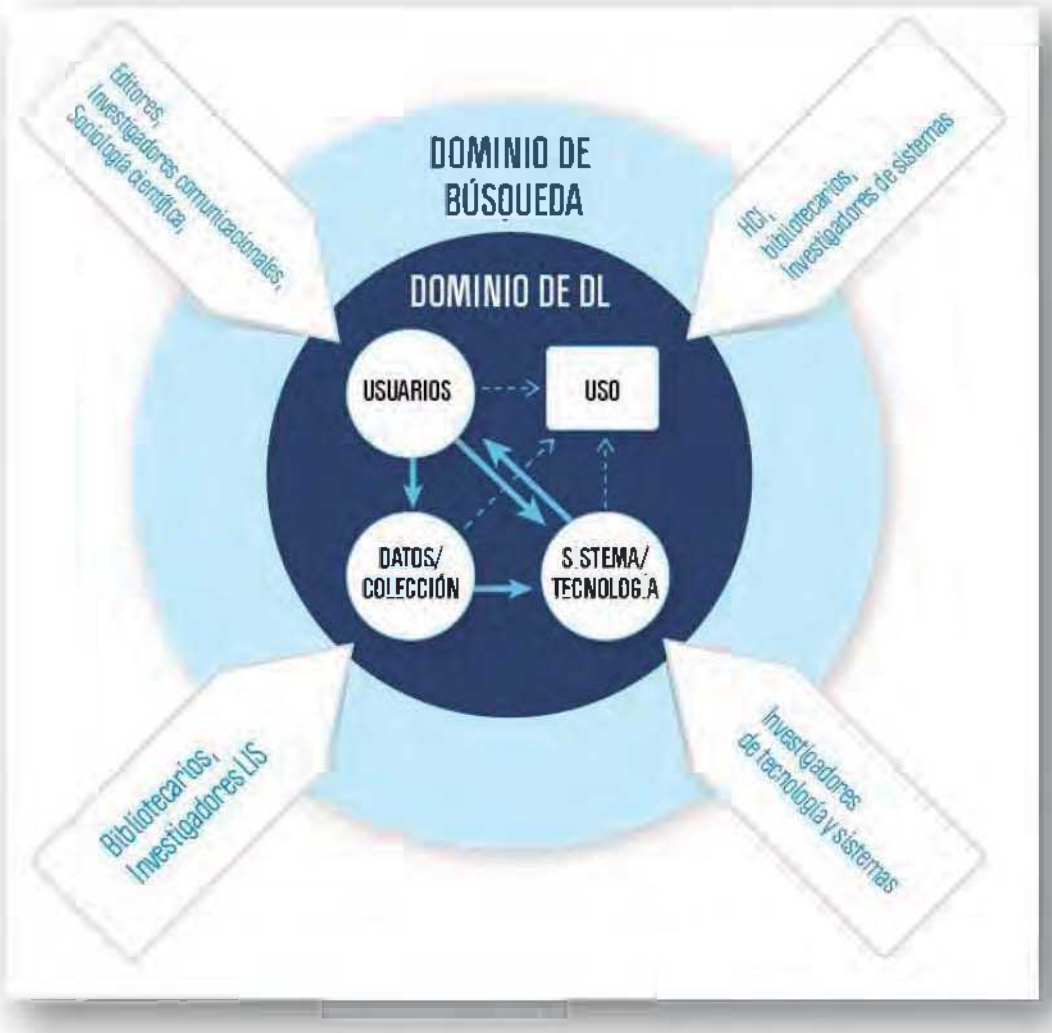


Figura 4.1: Esquema generalizado de una BD

Fuente: Fuhr *et al.* (2001).

Su propuesta incluye, como puede verse, tres componentes: usuarios, colecciones/datos y tecnología. Los datos/colecciones predeterminan tanto a la tecnología como a los usuarios. Las flechas finas muestran las interacciones entre personas y tecnología, y las flechas punteadas muestran la contribución colectiva de los usuarios, la colección y la tecnología al uso global. Desde el dominio de la BD es posible moverse hacia afuera al dominio de la investigación y usar las relaciones entre las áreas principales de investigación (usuarios, usos, colecciones y tecnologías) para crear un conjunto de requerimientos de investigación para un banco de pruebas de una BD.

Los autores identifican los parámetros más relevantes en la caracterización de las

tres dimensiones de dominio de la BD. La mayoría de los parámetros mencionados son binarios o están restringidos a un conjunto acotado de valores siguiendo la tabla 4.1.

Tabla 4.1: Criterios de evaluación para las tres dimensiones de una BD

DATOS/COLECCIÓN	TECNOLOGÍA	USUARIOS
Contenido: ninguno, parcial, completo. Audio, texto, video, etc. *Metacontenido*: indexación, tesauro, clasificación, metadatos, etc. *Gestión*: derechos, flujo de trabajo, gestión de usuarios, mantenimiento, preservación, frecuencia de mantenimiento, corrección de errores, etc.	*Tecnología vinculada al usuario*: creación de documentos (ingreso), interface, búsqueda, navegación, etc. *Acceso a la información*: recuperación, filtrado, extracción, minería de datos, eficacia, eficiencia, etc. *Estructura tecnológica*: repositorio, protocolos. *Tecnología de los documentos*: formatos.	*Quién*: interno, investigador, general. *Qué*: áreas de interés, distribución. *Cómo*: búsqueda de información: búsqueda, *surfing*. *Porqué*: propósito: uso, análisis, distribución.

Fuente: ISO 14721: 2012.

Primera encuesta

Con el fin de obtener un primer conjunto de información sobre las colecciones de bibliotecas digitales y colecciones de ensayo, los autores diseñaron un cuestionario de dos partes. El cuestionario A (para las bibliotecas digitales y colecciones de prueba disponibles), vinculado con la disponibilidad de las BD existentes y las colecciones de prueba que podrían ser utilizadas para la investigación en el campo de las bibliotecas digitales. El Cuestionario B (para colecciones digitales de prueba de la biblioteca deseada), que investigaba las necesidades futuras de las colecciones de prueba de bibliotecas digitales.

Las preguntas de ambos cuestionarios se compusieron en una manera similar, y la redacción sólo se modificó para reflejar su meta diferente: BD existentes *versus* requisitos y necesidades de investigación. Un conjunto de preguntas se corresponde con cada una de las tres principales categorías del esquema: los usuarios y usos,

datos/colecciones y tecnología. Además, el cuestionario incluyó preguntas sobre el género y el dominio de trabajo de la personas que respondieron. Las respuestas se realizaron seleccionando una o varias de una lista de opciones y con la posibilidad de dar un comentario. El cuestionario A contenía 31 preguntas y el cuestionario B, 21 preguntas. La encuesta fue anunciada en varias listas de correo para la audiencia principal de los investigadores y desarrolladores de bibliotecas digitales, y la estimación fue que respondió un 4% de la audiencia objetivo. Casi el 70% de los encuestados eran del dominio de la investigación, y los usuarios de las bibliotecas digitales evaluados también tenían más del doble de peso en el ámbito de la investigación que en cualquier otro dominio.

Estos estudios mostraron que, mientras que la clasificación propuesta parece ser adecuada para la caracterización de la BD, la redacción de las preguntas es muy problemática. Debido al enfoque integral del esquema de clasificación, muchas áreas de investigación están cubiertas, y estas áreas tienen sus diversos sistemas a largo plazo y su propio uso del lenguaje, lo que dificulta la creación de cuestionarios que sean igualmente comprensibles por los investigadores de diferentes áreas.

Los autores hablan también de MetaLibrary BD (Constantopoulos, Sølvberg, 2001), un esfuerzo del grupo de trabajo de DELOS que se encontraba en funcionamiento al momento del trabajo de Fuhr. Era una base de datos de la encuesta extensible, en la que cada colección de BD/banco de pruebas podía registrarse y proporcionar información y actualizarla. Las preguntas fueron rediseñadas y para los autores ofrecían más oportunidades para respuestas en texto libre. Las preguntas se asignaban en una jerarquía sugerida por el sistema de clasificación propuesto. Por lo tanto, los nodos de esta jerarquía representaban un área de investigación o una funcionalidad de la BD. Se pretendía que fuera una herramienta potente de búsqueda de experiencias en algunas de esas áreas. El trabajo se fijaba como una meta futura, con el fin de facilitar el mantenimiento de la meta-biblioteca, la posibilidad de cambiar a un sistema de recolección (*harvesting*), en la que cada proveedor de una BD de prueba mantuviera un registro de metadatos y el cosechador recogiera los registros a intervalos regulares y, de este modo, realimentar el banco de preguntas para crear colecciones de las pruebas pertinentes, similares a las iniciativas de TREC (Text REtrieval Conference) y CLEF (Conference and Labs of the Evaluation Forum). Resulta claro que este trabajo era una

propuesta muy ambiciosa, más aún en vista del año en que fuera escrito; sin embargo, la mirada integradora resulta excelente hoy en día, en que la tecnología ha avanzado para la obtención automática de perfiles de contenidos, formatos, usuarios, estándares de metadatos, etc.

Durante los años 2005 y 2006, Norbert Fuhr continuó abordando los temas de evaluación de BD y particularmente ahondó en la experimentación con grupos de bibliotecas produciendo nuevos trabajos en el área como, por ejemplo, Albrechtsen *et al.* (2005, 2006a y 2006b).

Fuhr, Tsakonas, Agosti, Hansen (2007)

Estos autores, como primer paso, describen la BD como un sistema de información especial constituido por distintos componentes, tal como ya se ha expresado en otros modelos previos: una colección o colecciones, un sistema informático que ofrece diversos servicios sobre la colección, personas y el medioambiente para la utilización de los recursos para los cuales se creó la BD, es decir que el modelo para crear una BD responde a estos usuarios y sus necesidades. El atractivo de las colecciones y su facilidad de uso determinarán la utilización de la BD. Dado el estado de avance de las BD, reconocen que cada vez son mayores las expectativas y requerimientos sobre la calidad de las mismas, por lo cual se precisan procedimientos de evaluación. Reconocen como antecedentes tanto al trabajo expuesto precedentemente (Fuhr, 2001) y en general a las determinaciones del *workshop* de DELOS celebrado en Padua (Agosti, Schek, Türker, eds., 2004).

El trabajo presenta una sección dedicada a conceptos y suposiciones en relación a la evaluación de BD, una referida al estado de la práctica, una revisión del marco conceptual de evaluación de DELOS y un nuevo marco, el cual opera dinámicamente en diversos niveles. Finalmente, realizan algunas recomendaciones y elaboran sus conclusiones.

Conceptos básicos y suposiciones en torno a la evaluación de BD

Para los autores, si bien las BD pueden ser utilizadas con diferentes propósitos, la razón central se enfoca en el acceso a la información: encontrar determinados

contenidos, recuperar información específica, localizar ítems conocidos, acceder a materiales que el usuario no conoce lo suficiente, etc. La provisión de este acceso para los usuarios está vinculada a las tecnologías de recuperación de información, que ya para 2007 tenían una gran formalización. Bajo esta concepción, la noción central de la evaluación es la *relevancia*, que sirve para entender las interacciones entre usuarios, utilización, necesidades de información y contenido informacional; sin embargo, la relevancia no toma en cuenta la satisfacción del usuario, la calidad de la información, las relaciones entre los documentos o la confiabilidad de la información. Como un dato aún más importante, se abstrae completamente de todos los contextos posibles que se debieran atender en una investigación. Esto incluye los distintos tipos de contextos en los cuales el ítem de información, el usuario y el productor interactúan entre sí.

Los autores citan además como métricas muy usadas en la evaluación (dada una noción objetiva como la relevancia) a las ya mencionadas *recall* y precisión. Estas medidas, que ya se han mencionado en los otros modelos, han sido definidas aquí para modelizar la completitud y la exactitud de los sistemas, respectivamente, pero no modelizan la satisfacción de los usuarios, la pertinencia de los resultados o la efectividad del sistema, dado un contexto de tarea determinado. En la evaluación de una BD, como una herramienta orientada al contenido, la performance de recuperación de información es importante, pero no es lo único que se debe evaluar para la aceptabilidad del sistema global y en la evaluación se debe tener algo más en cuenta. Asimismo, aproximaciones basadas en ensayos de laboratorio no alcanzan y son necesarios esquemas específicos en el campo de las BD.

Por lo expuesto, los autores sostienen que el soporte tecnológico para muchas de las interacciones del usuario en el proceso de acceso a la información, están implícitas en el contexto. La mayor parte de la persistencia o continuidad la provee el usuario en interacción con el sistema, más que el sistema y el alcance de un componente de recuperación de información, lo que los lleva a sostener la necesidad de utilizar nuevas métricas que tengan en cuenta las interacciones remarcadas.

Los autores se enfocan hacia un nuevo concepto de relevancia que tome en cuenta el contexto y la información de acceso al sistema para modelar a los usuarios de la información, los productores, las sesiones de acceso.

Este trabajo recoge varios modelos que ya han sido comentados: el modelo OAIS y

las experiencias de DELOS antes y después del *workshop* de Padua, en el relato de Candela (2007). Describe también el Modelo 5S (Gonçalves *et al.*, 2004), cuyo nombre refiere a *Streams* y *Structures* (para la construcción de objetos digitales), *Spaces* (para la descripción de los objetos y sus relaciones), *Scenarios* (para la definición de cómo los servicios y las actividades cambian el estado del sistema) y *Societies* (para la interconexión de roles y actividades dentro de la comunidad de usuarios). El modelo está basado en un formalismo matemático y ha sido usado en distintos casos de estudio, incluyendo la generación de una taxonomía de términos referidos a las BD. Los modelos pueden extenderse para incluir aspectos relativos a la evaluación y la calidad. Algunas propiedades como accesibilidad, preservación e integridad se han definido a través del modelo y pueden usarse como métricas de evaluación.

Los autores mencionan el marco conceptual de Nicholson (2004) para la evaluación integral de los servicios de biblioteca, que propone una matriz de evaluación con las dimensiones de la perspectiva (interna, sistema de bibliotecas y externa, por el usuario) y de los tópicos (sistema de bibliotecas y el uso). Una matriz de este tipo, afirman, puede ayudar a los evaluadores para seleccionar objetivos para las mediciones y los métodos para medir los objetivos seleccionados. Según el punto de vista de Nicholson, la evaluación implica medida y los criterios de evaluación se calculan como combinaciones de algunas de estas medidas. Introduce los puntos de vista de los distintos niveles de usuarios y así es posible, entre cambios propuestos por las evaluaciones y nuevas mediciones sobre impacto de los cambios, tener una mejora continua.

Además de los modelos y marcos precedentes, los autores mencionan la existencia de metodologías y herramientas prácticas que pueden ser utilizadas para medir el valor de los servicios de la BD. Por ejemplo, la metodología LibQUAL+, un estudio de mercado completo de las percepciones de los usuarios para identificar las faltas en la prestación de servicios, muy utilizado en Estados Unidos, dentro de la Asociación de Bibliotecas de Investigación (ARL) y que aún está en vigencia. También mencionan el *toolkit* eVALUEd, cuyo objetivo era —y es— producir un modelo transferible para la evaluación de la BD y proporcionar capacitación y difusión.

Estos modelos se diferencian de los modelos de evaluación vistos de las BD, debido a la relación de los resultados del funcionamiento de la BD con ciertas comunidades

atendidas, por ejemplo, el impacto o la satisfacción de los resultados para determinadas comunidades geográficas. Como resultado de la perspectiva que se ofrece, se considera la inherente dificultad que acarrean las actividades de evaluación. Esta visión fue reafirmada por la primera encuesta realizada por el Working Group sobre la evaluación de DELOS, como se ve en Fuhr (2001). Es de destacar que, en la evaluación de las condiciones actuales en su totalidad, se trata con un proceso multifacético que requiere contribuciones de variados actores y de distintas comunidades.

El trabajo incluye, además, un apartado dedicado a deslindar los términos vinculados a la evaluación de BD y, para culminar, expone una muy rica bibliografía que muestra la continuidad de un conjunto de autores y donde queda en evidencia la actividad significativa que llevó adelante el grupo DELOS. Vale la pena repetir que los *workshops* del grupo DELOS se encuentran aún en línea, con sus contenidos accesibles, por ejemplo en "DELOS Workshops" en el sitio de DBLP Computer Science Bibliography de la Universidad de Trier.

Reflexiones

Para culminar este capítulo resulta interesante, ante todo, remarcar la amplitud y actualidad de las propuestas de evaluación precedentes y luego hacer un brevísimo recorrido final centrado en diversos aspectos.

¿Por qué hacer una evaluación de un repositorio institucional? Ante todo, porque el propósito final del repositorio está vinculado a los usuarios y, en este sentido, es a todas luces necesario evaluar la calidad del trabajo realizado. Es decir, la construcción de un repositorio es un largo camino hasta un estado estable, donde su calidad alcanza las expectativas; además de eso, es necesario monitorear el avance o el progreso; y existe un punto aún más ambicioso vinculado a certificar la calidad del repositorio y a hacer una actividad de estudio (el mismo que debe realizarse antes de la implementación) dedicado a compararlo con el estado de otros repositorios. Hoy en día existen guías (Stellenbosch University, 2013; Barrueco Cruz *et al.* 2010-2011; Harvard Open Access Project, 2014), directrices (OpenAIREplus, 2008, 2010, 2013a, 2013b) y rankings web para evaluar la calidad de un RI. Asimismo, existen recomendaciones

para la certificación de confiabilidad de repositorios digitales (CCSDS 652.0-M-1, 2011).

Más allá del avance los modelos comentados, resulta importante reconocer qué modelos y métodos de evaluación abundan y cuáles están pensados con una amplitud que en muchos casos excede las posibilidades de realización; el modelo elegido deberá ceñirse a las funciones y funcionalidades que se quieren estudiar y la evaluación deberá versar sobre los puntos clave en el ciclo de vida del repositorio, sus objetos y sus usuarios; así, el diagnóstico, por pequeño que sea, resultará en una mejora (Fuhr lo relataba en el aspecto del usuario y los usos, pero vale para otros aspectos). Por lo demás, la actividad de modelización y evaluación deberá ser continua para mejorar efectivamente la calidad y los servicios brindados.

Bibliografía del capítulo

Albrechtsen, H.; Fuhr, N.; Klas, C.-P.; Micsik, A.; Kapidakis, S. A. (2005). "Digital library testbed framework for the evaluation of architectures, services and execution dynamics". En: *DELOS Research Activities 2005*, C. Thanos (ed.). DELOS Network of Excellence: p. 70-71.

Albrechtsen, H.; Fuhr, N.; Hansen, P.; Jacob, E.; Kapidakis, S.; Klas, C.-P.; Kovacs, L.; Kriewel, S.; Micsik, A.; Papatheodorou, Ch.; Tsakonas, G. (2006a). "A Logging scheme for comparative digital library evaluation". En *Proceedings of the 10th European Conference on Digital Libraries (2006)*. Thanos, C., Verdejo, M.F., Carrasco, R.C., Gonzalo, J. (eds.). Alicante (Spain). 17-22 September 2006. *Lecture Notes in Computer Science 4172*, Springer: p. 267-278.

Albrechtsen H.; Fuhr N.; Klas C.-P.; Micsik A.; Kapidakis S. (2006b). "A digital library testbed framework for the evaluation of architectures, services and execution dynamics". En *DELOS Research Activities 2006*, C. Thanos (Eds.). DELOS Network of Excellence. p. 111-113.

Barrueco Cruz, J. M.; Caballos Villar, A.; Campos Rodríguez, Á.; Casaldàliga, N.; Combarro Felpeto, P.; Cívico Martín, R.; Domènech, L.; García Gil, M. A.; Losada, M.; Morillo Moreno, J. C. (2010-2011). Guía para la evaluación de repositorios institucionales de Investigación. Recuperado el 13 de junio de 2014, de http://recolecta.fecyt.es/sites/default/files/contenido/documentos/GuiaEvaluacionRecolectav1.0-1.pdf

Borgman, C.; Sølvberg, I.; Kovács, L. (eds.) (2002). *Fourth DELOS Workshop. Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics*, Hungarian Academy of Sciences Computer and Automation Research Institute (MTA SZTAKI), Budapest, Hungary 6-7 June. Recuperado el 23 de junio de 2014, de http://www.ercim.eu/publication/ws-proceedings/DelNoe04.pdf

Bourdieu, P. (1990). *The logic of practice*. Stanford University Press.

Bishop, A. P. (1999). "Document structure and digital libraries: How researchers mobilize information in journal articles". *Information Processing & Management*. 35 (3) p. 255-279.

Constantopoulos, P.; Sølvberg I. T. (eds.) (2001). *Research and Advanced Technology for Digital Libraries. 5th European Conference*, ECDL 2001 Darmstadt, Germany, September 4-9.

Cornell University (1999). *MESL Technical Report*. Recuperado el 29 de septiembre de 2000,

de http://cidc.library.cornell.edu/gateway.htm

CCSDS 652.0-M-1 (2011). *Audit and Certification of Trustworthy Digital Repositories, (ISO Equivalent: 16363:2012).* Recuperado el 13 de junio de 2014, de http://public.ccsds.org/publications/archive/652x0m1.pdf.

D-Lib Working Group on Digital Library Metrics (s/d). Recuperado el 13 de junio de 2014, de http://www.dlib.org/metrics/public/.

eVALUEd (s/d). Recuperado el 13 de junio de 2014, de http://www.evalued.bcu.ac.uk/#.

Fuhr, N.; Hansen, P.; Mabe, M.; Micsik, A.; Solvberg, I. (2001). "Digital Libraries: A Generic Classification and Evaluation Scheme". *Proceedings of the 5th European conference on Research and Advanced Technology for Digital Libraries*, p. 187-99.

Fuhr, N.; Tsakonas, G.; Aalberg, T.; Agosti, M.; Hansen, P.; Kapidakis, S.; Klas K.; Kovács, L.; Landoni, M.; Micsik, A.; Papatheodorou, C.; Peters, C.; Sølvberg, I. (2007). "Evaluation of digital libraries". *International Journal on Digital Libraries.* 8(1), p. 21-38

Gonçalves, M. A.; Fox, E.; Kipp, N.; Watson, L. (2004). "Streams, Structures, Spaces, Scenarios, Societies (5S): a formal model for digital libraries". *ACM Transactions on Information Systems*, 22, p. 270-312

Harvard Open Access Project, FECYT, RECOLECTA, CRUE, REBIUN. (s/d). *Guía de buenas prácticas de políticas de acceso abierto para las universidades.* Recuperado el 13 de junio de 2014, de http://cyber.law.harvard.edu/hoap/Good_practices_for_university_open-access_policies.

Hill, L. L.; Carver, L.; Larsgaard, M.; Dolin, R.; Smith, T. R.; Frew, J.; Rae, M. A. (2000). "Alexandria Digital Library: User evaluation studies and system design". *Journal of the American Society for Information Science*, 51(3) p. 246-259.

ISO 9241-11:1998 (1998) "Ergonomic requirements for office work with visual display terminals (VDTs). Part 11: Guidance on usability". Recuperado el 23 de junio de 2014, de http://www.iso.org/iso/catalogue_detail.htm?csnumber=16883.

Jones, M. L. W.; Gay G. K.; Rieger, R. H. (1999). "Project Soup: Comparing Evaluations of Digital Collection Efforts". *D-Lib Magazine* 5 (11). Recuperado el 12 de junio de 2014, de http://www.dlib.org/dlib/november99/11jones.html.

Keefer Riva, A. (2001). "El proyecto Peak y sus implicaciones para el acceso a los artículos científicos". *El Profesional de la Información*, 10 (1-2), p. 28-30.

LibQUAL+ (s/d). Recuperado el 13 de junio de 2014, de https://www.libqual.org/home.

Marchionini, G.; Crane, G. (1994). "Evaluating hypermedia and learning: Methods and results from the Perseus Project". *ACM Transactions on Information Systems*, 12(1), p. 5-34.

Maristella Agosti; Hans-Jörg Schek; Can Türker (eds.) (2004). *Digital Library Architectures: Peer-to-Peer, Grid, and Service-Orientation, Pre-proceedings of the Sixth Thematic Workshop of the EU Network of Excellence DELOS*, S. Margherita di Pula, Cagliari, Italy.

Nicholson, S. (2004). "A conceptual framework for the holistic measurement and cumulative evaluation of library services". *J. Doc.* 60, p. 164-182.

OpenAIRE (2010). *Directrices OpenAIRE 2.0. Directrices para proveedores de contenido del espacio de información.* Recuperado el 13 de junio de 2014., de http://recolecta.fecyt.es/sites/default/files/contenido/documentos/OpenAIRE-Guidelines_v2-0_en.pdf.

OpenAIRE DRIVER Project (2008). *Directrices DRIVER 2.0. Directrices para proveedores de contenido - Exposición de recursos textuales con el protocolo OAI-PMH.* Recuperado el 13 de junio de 2014, de http://recolecta.fecyt.es/sites/default/files/contenido/documentos/DRIVER_2_1_Guidelines_Spanish.pdf.

OpenAIREplus (2013a). *Directrices OpenAIRE 3.0 para repositorios de documentos.* Recuperado el 13 de junio de 2014, de https://guidelines.openaire.eu/wiki/OpenAIRE_Guidelines:_For_Literature_repositories.

OpenAIREplus (2013b). *Directrices OpenAIRE 1.0 para el archivos de datos.* Recuperado el 13 de junio de 2014, de https://guidelines.openaire.eu/wiki/OpenAIRE_Guidelines:_For_Data_Archives.

Payettea S. D.; Riegerb, O. Y. (1998). "Supporting scholarly inquiry: Incorporating users in the design of the digital library". *The Journal of Academic Librarianship*. 24 (2) p. 121-129.

Perseus Project (1987). Recuperado el 9 de junio de 2014, de http://www.perseus.tufts.edu/hopper/.

Saracevic, T.; Covi, L. (2000). "Challenges for digital library evaluation". *Proceedings of the Annual Meeting-American Society for Information Science.* 37 p. 341-350.

Saracevic, T.; Covi, L. (2001). "Digital Library Evaluation: Toward an Evolution of Concepts". *Library Trends* 49 (3), p. 350-369.

Saracevic, T. (2004). "Evaluation of digital libraries: an overview". *DELOS Workshop on the evaluation of digital libraries. DELOS WP7.* Recuperado el 12 de junio de 2014, de http://dlib.ionio.gr/wp7/WS2004_Saracevic.pdf.

Schiff, L. R.; Van House, N. A.; Butler, M. H. (1997). "Understanding complex information environments: A social analysis of watershed planning". *Proceedings of the 2nd ACM International Conference on Digital Libraries*, p. 161-168.

Stellenbosch University (2013). *Guía para la puesta en marcha de un repositorio institucional con software DSpace.* Recuperado el 13 de junio de 2014, de http://recolecta.fecyt.es/sites/default/files/contenido/documentos/ir-guide-v18%5B1%5D.pdf.

Text REtrieval Conference (TREC) (1992). Recuperado el 13 de junio de 2014, de http://trec.nist.gov/.

Tsakonas, G.; Papatheodorou, Ch. (2008). "Exploring usefulness and usability in the evaluation of open access digital libraries". *Information Processing & Management*, 44 (3) p. 1234-1250.

Universität Trier (s/d). *DELOS Workshops.* DBLP Computer Science Bibliography. Recuperado el 13 de junio de 2014, de http://www.informatik.uni-trier.de/~LEY/db/conf/delos/index.html.

Capítulo 5 | Caso de estudio: SEDICI

Síntesis: La propuesta de esta tesis, como se dijera, es ofrecer un modelo de evaluación del contenido de un repositorio institucional y proponer, a partir de dicho modelo, posibles mejoras en la calidad del material ofrecido al usuario y, particularmente, en sus posibilidades de acceso, entendido en el sentido de que los documentos siempre estén disponibles y legibles para los usuarios. Este capítulo compendia los objetivos y propuestas principales de la tesis de modo de exponer adecuadamente la experimentación realizada sobre el RI elegido y darle el contexto adecuado para su mejor comprensión.

Estado de situación y justificación

Las instituciones educativas dedican, en la actualidad, enormes esfuerzos para la puesta en marcha y mantenimiento de repositorios institucionales, esto es, estructuras web que ofrecen servicios a sus usuarios y que albergan la producción institucional en sus diferentes facetas, académica, de investigación, de extensión, en una definición propia (de acuerdo a la institución) y caracterizada por la diversidad. El objetivo central de los repositorios es reunir, gestionar, organizar, preservar y difundir en un mismo sitio, esa producción, así como permitir la exposición de los registros de la misma, utilizando estándares que permitan la interoperabilidad con otros repositorios similares. Una comunidad de autores y miembros de la institución colabora con la administración del repositorio, poblándolo a través de procesos de autoarchivo y también a partir de la carga "mediada", realizada por personal del repositorio (administradores). Como parte esencial de un repositorio existe también una comunidad de usuarios, personas y máquinas que precisan de ciertos contenidos, esto es, la información guardada en el repositorio.

El crecimiento en contenidos de alrededor de 3000 repositorios del mundo[13]

[13] OpenDOAR registra 2579 repositorios y ROAR, 3583, al 5 de febrero de 2014.

muestra una notable proliferación de objetos digitales con sus representaciones y su medioambiente asociados, los cuales, hoy día, por la propia fragilidad del objeto digital (que no existe sin una representación y queda obsoleto fácilmente), llevan a la necesidad de mantenerlo a un alto costo, realizando un conjunto de acciones imprescindibles para asegurar que el objeto subsista en el tiempo y, fundamentalmente, que la comunidad de usuarios pueda tener acceso a él y a su interpretación. Las acciones, en este sentido, se inscriben dentro de las llamadas "acciones de preservación de objetos digitales". De allí que esta tesis se haya propuesto como objetivo principal el encontrar la mejor metodología para la evaluación de un RI y proponer, a partir de ella, una serie de tareas y estándares a cumplir por parte de los RI.

Proyectos y trabajos relacionados a preservación digital, evaluación y confiabilidad de repositorios

En la última década, ha surgido un número importante de iniciativas de investigación dedicadas a la preservación digital, debido al urgente problema que presenta esta actividad. Importantes autores como Cristoph Becker, Andreas Rauber, Stephan Strodl, Hannes Kulovitis y Hans Hofman (todos de la Universidad de Viena, Austria), entre otros, han realizado un sinnúmero de trabajos y participado en proyectos de alcance internacional. Ellos y muchos otros expertos han abundado en tópicos como el perfilamiento de repositorios, los criterios, el plan de preservación, las mediciones y los planes específicos para diferentes tipologías documentales. Cabe destacar que una importante parte de estos textos se encuentra en línea y accesible para su lectura, por ejemplo en Plato. Sección de documentación.

Además de metodología y proyectos, otro conjunto importante de autores se ha dedicado a las experiencias en migración (que transforma los objetos digitales a representaciones actualizadas o de mayor acceso) y emulación (que crea un medioambiente técnico donde los objetos pueden ser interpretados como en el medioambiente original), consideradas las dos estrategias de mayor importancia en cuanto a la preservación, así como a la puesta a punto de herramientas para la realización de estas conversiones (Rothenberg, 1999; Lawrence *et al.*, 2000; Hoeven *et*

al., 2005).

En torno a las acciones de preservación de un repositorio, y a la generación de un plan de preservación que tenga en cuenta las funcionalidades descritas en alto nivel por el modelo abstracto OAIS de la norma ISO 14721:2012, está claro que son necesarias las herramientas de caracterización del contenido, que es el objeto mismo de la preservación. Como se verá más adelante, herramientas como DROID y JHOVE realizan la identificación de los formatos de archivos, su validación y la caracterización de los objetos digitales.

En paralelo, también en esta década ha habido esfuerzos vinculados a la confiabilidad de los repositorios. La confiabilidad en el contexto de la preservación a largo plazo de los objetos digitales significa contar en las instituciones (y particularmente en el repositorio) con estrategias capaces de superar el cambio tecnológico continuo (Dobratz, Schoger, 2007).

En el año 2003, el Research Library Group (RLG)[14] y el National Archives and Records Administration (NARA)[15] se unieron para analizar la posibilidad de crear repositorios digitales seguros y poder acreditar esa capacidad. El objetivo era permitir la auditoría del repositorio, la evaluación y su certificación. El grupo de trabajo RLG-NARA, autores de *Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)*, documento que describe las métricas de un repositorio adecuado a OAIS, junto con las aportaciones del Center for Research Libraries (CRL)[16], el Digital Curation Centre (DCC)[17] y el Digital Preservation Europe, desarrolladores de Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)[18] y el NESTOR[19]

[14] The Research Libraries Group (RLG) es una organización sin fines de lucro de más de 150 universidades, bibliotecas de investigación, archivos, sociedades y museos dedicada a mejorar el acceso a la información académica y científica. Se fundó en 1974 y desde entonces es líder en desarrollar soluciones para la adquisición, acceso, entrega y preservación de estos materiales.

[15] La National Archives and Records Administration (NARA) es una agencia gubernamental independiente de los Estados Unidos, dedicada a la preservación de los registros de documentos gubernamentales e históricos y también de darles acceso público.

[16] El Center for Research Libraries (CRL) es un consorcio internacional de universidades y bibliotecas de investigación que preserva materiales en el área de Humanidades, como modo de respaldar la enseñanza e investigación en el área.

[17] La creación del Digital Curation Centre (DCC) fue clave en las estrategias de JISC (Joint International System Committee) para el acceso continuo y la preservación digital; su propósito, afrontar los desafíos en curaduría digital (*digital curation*) que no podían llevarse adelante por una institución aislada ni a través de una disciplina aislada.

[18] La Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) es una metodología y un conjunto de herramientas basadas en software asociado desarrollado por Digital Curation Centre

Working Group, autores de *Catalogue of Criteria for Trusted Digital Repositories* (2006), fijaron las condiciones para poder crear depósitos y archivos digitales seguros, auditables y certificables; luego, seleccionaron diez requisitos básicos que los repositorios digitales deberían cumplir para garantizar los resultados de su actividad en el tiempo, a saber:

1. Dedicación y compromiso con los objetos digitales.
2. Organización.
3. Legalidad.
4. Eficiencia y eficacia en las políticas.
5. Infraestructura técnica adecuada.
6. Adquisición.
7. Integridad, autenticidad y usabilidad en la conservación del objeto digital.
8. Gestión de metadatos y existencia de una pista de auditoría.
9. Difusión.
10. Planificación y actuación.

Tanto TRAC como DRAMBORA y NESTOR usan el modelo de referencia OAIS. Todos ellos nacen con el mismo objetivo y con la necesidad de establecer pautas para medir el cumplimiento (ideal) de este modelo completo en todas sus funcionalidades. Sin embargo, los criterios para la evaluación son diferentes en los tres proyectos. Con base en el estándar TRAC, en el año 2012 se publica la norma ISO 16363:2012, que define las prácticas recomendadas para la evaluación de la confiabilidad para todo tipo de repositorios digitales.

PLANETS (Preservation and Long-term Access through Networked Services) es un proyecto que comenzó en el año 2006 en Europa, con el objetivo de pensar los problemas y brindar soluciones útiles para la preservación a largo plazo y el aseguramiento del acceso a los acervos culturales presentes en instituciones europeas. Propusieron, para ello, una red de recursos tecnológicos compartidos, donde las

---

(DCC) and Digital Preservation Europe (DPE), para apoyar la evaluación de los repositorios de preservación digital.

[19] NESTOR (Network of Expertise in Long-Term STOrage of digital Resources) es una iniciativa fomentada por el Ministerio de Educación e Investigación de Alemania (http://www.dcc.ac.uk/resources/repository-audit-and-assessment/nestor#sthash.xuf56imV.dpuf).

herramientas a sumar debían ser abiertas e integrables en una red de servicios distribuidos. Con estos y otros muchos antecedentes en línea, PLANETS elaboró criterios que tienen en cuenta muchos aspectos de la institución que lleva adelante el plan: políticas de preservación, obligaciones legales, limitaciones organizativas y técnicas, requisitos de los usuarios y objetivos de preservación que ha definido la propia institución (los alcances del plan); también debe describirse el contexto de preservación, las estrategias evaluadas y la decisión resultante de cada estrategia. Hay un conjunto de pasos o acciones, junto con las responsabilidades y las normas y condiciones para la ejecución del plan de preservación, para la colección elegida. A condición de que las acciones y su implementación, así como el entorno técnico lo permitan, cualquier plan de acción termina definiendo un flujo de trabajo ejecutable.

Para los aspectos arriba mencionados, PLANETS incluye criterios de TRAC y NESTOR y medidas a realizar, todas las cuales pueden llevarse adelante de manera automática con la herramienta Plato. El plan ideal de preservación debería contener elementos que sirvieran para identificarlo, un estado dado (en ejecución, terminado...), disparadores (por ejemplo, cambios en la colección que pueden llevar a la necesidad de ejecutar el plan), una descripción de los objetivos de preservación de la institución misma, la colección, los responsables y las acciones a tener en cuenta.

Noción de preservación de la UNESCO

La preservación digital, de acuerdo al documento de UNESCO titulado *Directrices para la preservación del Patrimonio Digital*, del cual se recomienda su lectura, con especial atención a los capítulos 4 y 5, dedicados al deslinde terminológico, supone la selección y puesta en práctica de un conjunto evolutivo de estrategias, con objeto de lograr el tipo de accesibilidad anteriormente mencionado, considerando las necesidades de preservación de las diferentes capas de los objetos digitales. En dicho documento, puede leerse que *"la preservación digital puede definirse como el conjunto de los procesos destinados a garantizar la continuidad de los elementos del patrimonio digital durante todo el tiempo que se consideren necesarios"*. Y también:

"La mayor amenaza para la continuidad digital es la desaparición de los medios

de acceso. No puede decirse que se han conservado los objetos digitales si, al haber dejado de existir los medios de acceso a ellos, resulta imposible utilizarlos. El objetivo de la preservación de los objetos digitales es mantener su accesibilidad, es decir, la capacidad de tener acceso a su mensaje o propósito esencial y auténtico". (UNESCO, 2003: p. 37)

Las estrategias de las directrices de la UNESCO (p. 39-40) abarcan:

1. Colaborar con los productores (tanto autores como distribuidores) para consensuar normas y estándares que prolonguen la vida efectiva de los medios de acceso y reduzcan la variedad de problemas a resolver.
2. Que no es posible preservar todo y que hay que seleccionar el material a preservar, por lo cual un buen plan de preservación debe definir claramente su alcance para cada parte de la colección.
3. Debe existir seguridad en cuanto al lugar donde se guardan los materiales.
4. Utilizar durante toda la gestión de los materiales metadatos estructurados y otros documentos que faciliten el acceso y ayuden durante todo el proceso de preservación.
5. Proteger la integridad y la identidad de los datos.
6. Elegir los medios apropiados para proporcionar acceso a pesar de la existencia de constantes cambios en la tecnología.
7. Gestionar los programas de preservación para lograr los objetivos con la mayor economía de recursos posible y atendiendo al dinamismo necesario.

Esta propuesta

Como se desprende del panorama precedente, la puesta en práctica de acciones y de procesos dedicados a la preservación digital en un repositorio institucional depende de muchas decisiones, que van desde lo político hasta acciones concretas generadas de manera sistemática a partir de —en los mejores casos— un plan de preservación, dentro del cual los objetivos y alcances de la propuesta constituyen un primer paso. La preservación digital busca asegurar la conservación y disponibilidad de los objetos en soporte digital a largo plazo, como se dijera. Como consecuencia de este objetivo, que

resulta imprescindible para asegurar la calidad de un repositorio y es, en definitiva, central a la utilidad que debe brindar a los usuarios, se han desarrollado a lo largo del tiempo estándares y recomendaciones que atienden a este objetivo central de preservación.

Dentro de un sinnúmero de realizaciones, recomendaciones y estándares para gestionar la adecuada preservación de objetos digitales en un RI, se destaca el Modelo OAIS Reference Model for Open Archival Information Systems (OAIS), que, al decir de UNESCO en las antes mencionadas directrices, es:

> *"...la mejor tentativa existente para definir tanto un modelo conceptual de gestión de los objetos digitales de valor perdurable como un vocabulario aplicable al tema. Cualquier persona que prevea asumir una responsabilidad de gestión de objetos digitales debe tratar de comprender los conceptos tratados en el propio Modelo de Referencia"* (UNESCO, 2003: p. 45)

Se encuentra entonces ampliamente justificado, por dichos antecedentes de proyectos de importancia, tomar en este trabajo como referencia central el modelo OAIS. Su noción de paquete de información como contenedor conceptual, en sus variantes SIP, AIP y DIP, vistas en el capítulo 3, resulta muy adecuada para los propósitos aquí perseguidos, ya que es imprescindible para identificar los principales componentes y procesos implicados, en los distintos flujos de trabajo, para la preservación a largo plazo de los documentos en soporte digital. El modelo OAIS propone un marco conceptual destinado a cumplir con los objetivos de interoperabilidad y preservación y cubre, en este sentido, las acciones presentadas por UNESCO para que todo usuario (máquina o humano) del repositorio pueda entender y apropiarse eficazmente de la información contenida allí.

Metodología de trabajo

Objetivo de la evaluación

La presente evaluación sobre el repositorio institucional SEDICI va a analizar la

estructuración de sus contenidos bajo el criterio de usabilidad específico, que es la accesibilidad:

- Contenidos (del RI)
- Accesibilidad de los contenidos para la comunidad designada de usuarios.

Cabe advertir aquí que el significado de "accesibilidad" sigue, en este contexto, lo expresado por UNESCO y delineado en los párrafos precedentes. Obsérvese que al mencionar a los usuarios como la "comunidad designada", ésta podrá extenderse de modo de abarcar a usuarios con capacidades diferentes. Aunque el objetivo de este trabajo no es tan amplio, algunos estándares que se mencionarán sí son útiles para brindar accesibilidad en grado tal que todas las personas puedan utilizar los objetos digitales. En definitiva, si bien la propuesta inicial no forzará al uso de los mismos, queda indicado que este será el objetivo a más largo plazo. Debido a este uso más limitado del término, en numerosas ocasiones, en el decurso de este trabajo se utilizará el término "legibilidad" para especificar las posibilidades del usuario vinculadas a la capacidad de leer e interpretar correctamente el contenido.

Para establecer un vocabulario que clarifique conceptos en común, se utilizará —de entre los modelos ya presentados— el establecido por la norma ISO 14721, que define de manera más precisa las funciones esperadas para un repositorio o archivo digital y las presenta a partir de las seis entidades y sus funciones asociadas como se describió en el capítulo precedente. Este modelo ha devenido en el estándar para representar el intercambio de información entre un repositorio y su entorno. En particular, los elementos o componentes interactúan para desempeñar ciertas funciones de modo de alcanzar un objetivo central con respecto a una comunidad designada, la cual siempre debe poder acceder a los objetos digitales y también debe poder interpretarlos, no importa el tiempo que medie desde que el objeto fue colocado por vez primera en el repositorio y la consulta en cuestión.

En este contexto, como se dijera, la evaluación significa una valoración del desempeño o del funcionamiento de un sistema, o parte del mismo, en relación a cierto(s) objetivo(s). El análisis de la norma ISO 14721 permite aseverar que el flujo de datos necesario para cumplir los cometidos del repositorio con sus usuarios, se asegura a través de numerosas operaciones que se ponen en marcha desde el inicio de ese flujo

de trabajo por parte de un productor de contenido (un investigador, un tesista o un sistema automático) que suministra material al repositorio. Dentro ya del repositorio, se realiza un conjunto de operaciones que transforman ese paquete de información en otro paquete de información (más completo), con todas las características necesarias para poder realizar su preservación en el repositorio. Ese paquete también debe cumplir con los objetivos de interoperabilidad (en otras de sus formas: DIP), y, de hecho, las dos funciones que el modelo asegura podrán cumplirse con la propuesta normativa aquí expuesta. El flujo de trabajo culmina cuando se entrega el paquete solicitado al usuario (por ejemplo, un investigador que ha realizado una consulta al repositorio).

De esto surge claramente que existen muchas funciones que realizan las distintas entidades que representan en OAIS al propio repositorio y está claro que el incumplimiento de cualesquiera de ellas interrumpe el flujo de datos hacia el usuario. Por todo ello es que este trabajo se centrará muy especialmente en la evaluación de las funcionalidades del repositorio vinculadas a la preservación de los objetos digitales, dado que sin las mismas es imposible entregar al usuario el contenido requerido.

El desempeño del repositorio será evaluado también en cuanto a la efectividad: ¿cuán bien desempeña el RI (o cualquiera de sus partes) aquello para lo que fue designado? El parámetro de este desempeño será la capacidad del sistema de asegurar la preservación de los objetos digitales que alberga (o un subconjunto de los mismos, de acuerdo a los alcances de su plan de preservación). De las directrices de UNESCO ya mencionadas, se extractan los siguientes párrafos de sus dos primeros artículos, pues resultan de suma utilidad a este punto:

"Artículo 1 – Patrimonio digital

Cada vez más, los recursos que son fruto del saber o la expresión de los seres humanos, sean éstos de carácter cultural, educativo, científico o administrativo o engloben información técnica, jurídica, médica y de otras clases, se generan directamente en formato digital o se convierten a éste a partir de material analógico ya existente. Los productos «de origen digital» no existen en otro formato que no sea el electrónico original. Los objetos digitales pueden ser textos, bases de datos, imágenes fijas o en movimiento, grabaciones sonoras, material gráfico, programas informáticos o páginas web, entre otros muchos formatos

posibles dentro de un vasto repertorio de diversidad creciente. A menudo son efímeros, y su conservación requiere un trabajo específico en este sentido en los procesos de producción, mantenimiento y gestión.

Artículo 2 – Acceso al patrimonio digital

El objetivo de la conservación del patrimonio digital es que éste sea accesible para el público de modo permanente...".

Criterio de evaluación

Se considerará el par elemento-elegido-para-ser-evaluado/elemento-elegido-sobre-su- desempeño y, dado que se trabajará con un eje centrado en el sistema (ingeniería, procesamiento y contenido), se analizará qué tan bien están descritos, representados, estructurados y administrados los recursos de la colección para el propósito de mantener el acceso a lo largo del tiempo para una comunidad designada de usuarios.

Justificación de la elección del contenido y de los elementos a evaluar

En la gestión de documentos digitales un elemento central es la autenticidad. Los elementos relacionados con la autenticidad son la integridad, la identidad, y los elementos esenciales que, si se consiguen proteger en los documentos, garantizarán su preservación incluso después de procesos de migración o conversión. El proyecto Investigating the Significant Properties of Electronic Content Over Time (InSPECT)[20] se ha dedicado al análisis de estos elementos o características esenciales para la conservación de un objeto digital.

En este proyecto se planteó como objetivo que el documento digital (el objeto digital) debe ser conservado todo el tiempo, accesible y con significado. Las propiedades del objeto digital se muestran según la categoría contenido, contexto (metadatos), apariencia (color, forma), funcionamiento (interacción, funcionalidad) y estructura (paginación, secciones). Se debe decidir cuál de estos aspectos de cada categoría deben conservarse a lo largo del tiempo para una correcta planificación, y así

[20] Investigating the Significant Properties of Electronic Content over Time (InSPECT). Recuperado el 2 de diciembre de 2014 de http://www.significantproperties.org.uk/.

establecer buenas prácticas para la conservación de los objetos digitales. El proyecto se centró en demostrar que, en la conservación digital, los enfoques centrados en datos, es decir, preocupados por mantener el objeto de datos utilizable en el tiempo, ofrecen mejores perspectivas de éxito que aquellos que están centrados en el proceso, por ejemplo, tratar de mantener el software original y/o entornos de hardware operativos.

InSPECT se basó, para la definición de las propiedades, en el modelo OAIS y su relación con la Representación de la Información. Este punto nuevamente afirma las posibilidades del constructo de evaluación planteado en esta tesis. El modelo OAIS sirve como modelo funcional de un sistema de archivos. No es un modelo de metadatos en sí mismo pero define distintos tipos de objetos conceptuales, y algunos de ellos se definen específicamente para cumplir con los propósitos de la preservación y el acceso. Cada tipo de objeto de información se construye a partir de un modelo genérico simple, en el cual el objeto de datos puede ser interpretado usando la información de representación asociada.

Para OAIS, como se viera en anteriores capítulos, el Paquete de Información es un contenedor conceptual constituido por información de contenido e información descriptiva de preservación. La información de representación (puesto que un ítem puede estar compuesto de varios bloques o bundles y este a su vez de varios archivos o bitstreams, y cada bitstream tiene su propia información de representación) se utiliza para comprender la información que lleva el objeto digital, como puede verse en la figura 5.1.

Figura 5.1: Objeto de contenido y su representación: bundle text y bundle original

Fuente propia.

La información descriptiva sirve para localizar el objeto digital y la información de empaquetado es la que enlaza la información de contenido con la información de la PDI. Con esto en mente, parece claro que se evaluarán los objetos del repositorio como paquetes de información (con todas sus partes), o más específicamente el AIP (en términos de OAIS) que es la forma del paquete de información que aseguraría su preservación. Si el AIP está bien formado será intercambiable (interoperable), convenientemente transformado en un DIP, y permitirá todas las operaciones necesarias para su preservación. El paquete de información puede verse en la figura 5.2, la cual se realizó con base en la norma ISO 14721.

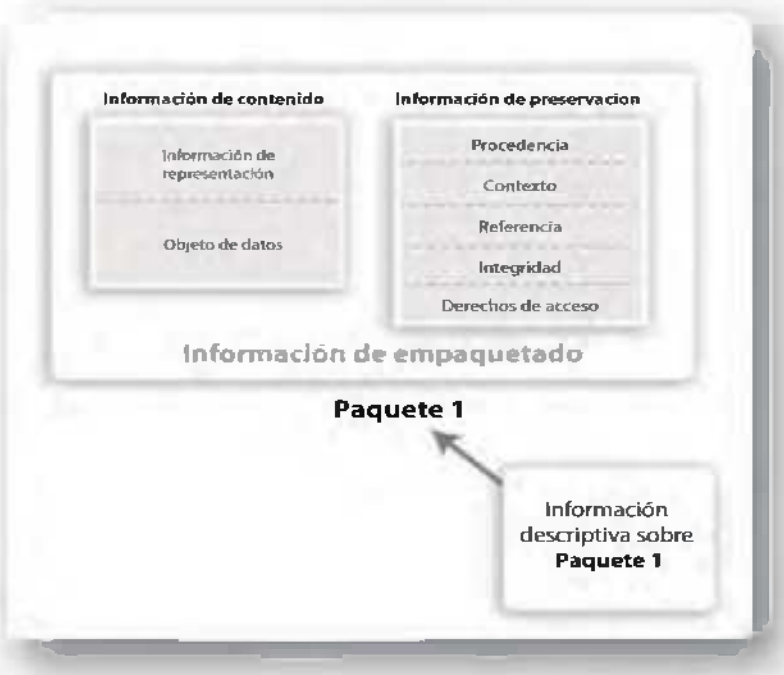


Figura 5.2: Paquete de Información del modelo OAIS y sus partes

Fuente: norma ISO 14721:2012.

Si se acepta que el Modelo de Referencia OAIS es, como establece UNESCO en sus directrices ya citadas, *"...para todos aquellos que diseñen, utilicen y evalúen aplicaciones reales. Su valor reside en que explica lo que es necesario a un elevado nivel conceptual, independientemente de los medios seleccionados para lograrlo"* (UNESCO, 2003: p. 45),

y si la norma establece que la unidad de intercambio de información es el "paquete de información", de allí se deduce que si este paquete está bien formado en cada una de sus variantes y según los distintos usos (ingesta, preservación, entrega) esto asegurará que los procesos, vinculados a las entidades planteadas por OAIS para el modelo funcional, realizarán correctamente sus funciones. Como las dos principales funciones de este modelo, plasmadas en esas entidades, son la preservación y la interoperabilidad, el repositorio que cumpla con esto, por ende, funcionará correctamente. De allí que se plantee aquí revisar el paquete de información en todas sus partes (puesto que es lo que se va a intercambiar, lo que los usuarios y sistemas informáticos van a querer preservar y asegurar su acceso), como modo de evaluar satisfactoriamente las actividades del repositorio.

De este modo, las tareas propuestas aquí están vinculadas a los elementos constitutivos del paquete de información: la *información de contenido* (CDO), la *información sobre la representación* de ese contenido (RI), la *información descriptiva de preservación* (PDI) y la *información descriptiva* (DI). La *información de empaquetado* (PI) no será considerada en este trabajo, por lo siguiente: para el modelo OAIS esta información es la que conecta los componentes dentro de una unidad identificable o medio específico. Por ejemplo, si la CI y la PDI están especificadas y son parte del contenido específico de un archivo TAR, la información de empaquetado puede incluir el nombre del archivo TAR y su codificación específica. La información de empaquetado puede tomar distintas formas dependiendo del medio de diseminación y de los requerimientos de los consumidores.

En el caso de DSpace, el software que soporta a SEDICI, repositorio elegido para la evaluación, esta información se encuentra inmersa dentro de su base de datos relacional (PostgreSQL) y forma parte también de su arquitectura interna, su modelo de objetos y sus métodos y funciones. Esto asegura una consistencia a nivel de base de datos, proporcionada por la definición de un esquema que brinda integridad referencial y el uso de transacciones que aseguran las propiedades ACID[21] durante todo el ciclo de vida de los datos, y hace innecesario considerarlo en el presente análisis.

La presente propuesta tiene como fin, entonces, generar un reporte sobre el estado

[21] ACID: acrónimo de Atomicity (atomicidad), Consistency (consistencia), Isolation (aislación) y Durability (durabilidad). Más información en http://es.wikipedia.org/wiki/ACID.

de los objetos del repositorio, considerándolos como paquetes de información. Se reportará si están bien formados (o no) y si serán preservables (o no), si cuentan con todos los elementos que la norma define para el paquete de información; también se validará que cada elemento esté bien formado en el sentido de que cumple con una serie de estándares o criterios que se van a definir a continuación. Si los paquetes de información en el repositorio se adecúan a los criterios establecidos, los objetos digitales del repositorio, y por tanto el repositorio mismo, "pasan" la evaluación.

Objetos Digitales (OD)

Los metadatos que son útiles a la preservación pueden verse más fácilmente en su utilidad en la figura 5.3, en la que se se representa, junto al Objeto Digital (OD), las acciones a realizar para su preservación, de las cuales se desprenden los metadatos necesarios para ello.

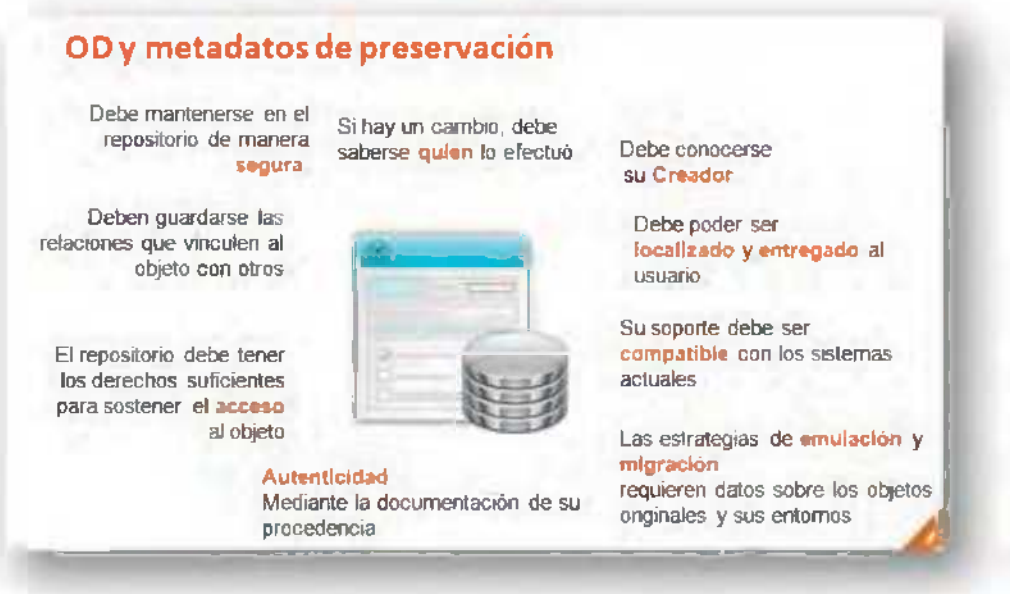


Figura 5.3: Objeto Digital (OD) y las acciones en el ciclo de vida para su preservación

Fuente propia.

Para llevar a cabo las acciones de preservación, de acuerdo a las características del OD, se proponen las siguientes operaciones:

1. **Perfilamiento automatizado de los objetos del repositorio**: esto involucra al objeto de contenido (CDO) con sus propiedades significativas y a la

información de representación de ese objeto (RI). El contraste aquí se realizará comparando los objetos, según su tipología, con el registro PRONOM[22].

2. **Revisión de los metadatos de preservación** que acompañan a los objetos digitales del repositorio y contraste con los metadatos de la PDI de OAIS.
3. **Revisión de la información descriptiva**: metadatos descriptivos que permiten la localización de los objetos digitales y contraste con los metadatos descriptivos que proponen las directrices DRIVER (2008).

De este modo, con la experimentación propuesta en los tres pasos precedentes se construirá la evaluación del RI elegido, la cual deberá, además, brindar un reporte adecuado y una recomendación de acciones a seguir en el caso de los objetos que no cumplan con los requerimientos.

Verificación de metadatos

Resulta claro que el paquete de información propuesto por la norma ISO 14721 es una estructura compleja y, sin embargo, podría afirmarse que tiene dos grandes áreas: una de ellas es la referida al contenido en sí mismo y su representación, la cual se trata más adelante en la sección de revisión de contenido y formatos. La otra área son los metadatos, que se despliegan de manera muy elemental aquí, para luego tratarse en detalle en los apartados dedicados a las revisiones de la PDI y la DI. A continuación, se realiza un comentario más detallado de las tres operaciones básicas que se llevarán a cabo como primera medida para evaluar correctamente el repositorio:

1. ***Perfilamiento automatizado de los objetos del repositorio***: como se mencionó anteriormente, la información de contenido y el perfilamiento de la misma vinculado a los distintos formatos en que se encuentra representada la información contenida en el repositorio y tratada en este experimento, es una de las partes del paquete de información. El perfilamiento ha sido realizado con la herramienta DROID, que contrasta el análisis que realiza sobre formatos con

[22] El registro de información técnica en línea PRONOM es un recurso imprescindible para todos aquellos que precisan información imparcial sobre formatos de archivos, productos de software y otros componentes técnicos requeridos para realizar la preservación a largo plazo de recursos electrónicos.

el registro internacional PRONOM y de este modo brinda una perspectiva en cuanto a los riesgos de los formatos incluidos en un repositorio y específicamente sobre el riesgo de pérdida de información. En el apartado de descripción de herramientas se hablará de DROID y de qué tipo de reportes brinda a los usuarios.

2. ***Metadatos de preservación:*** la información descriptiva de preservación (PDI) para OAIS debe brindar datos suficientes sobre la procedencia (*provenance*), el contexto (*context*), la referencia (*reference*), la integridad (*fixity*) y los derechos de acceso (*access rights*). La procedencia, más allá de describir la fuente, incluye los procesos que se han realizado sobre la información: la historia del objeto, cambios, versiones y responsables. El contexto muestra las relaciones con otras fuentes de información o contenidos. La referencia provee una identificación única del contenido. La integridad (o fijeza) provee una protección para que la información no sea alterada de manera intencional o no intencional. Los derechos de acceso proveen los términos de acceso, incluyendo preservación, distribución y uso de la información de contenido. Por ejemplo, podría contener los permisos de los autores o derechohabientes para las operaciones de transformación necesarias para la preservación, la licencia de distribución que se da al repositorio y la licencia de uso (acceso a los contenidos). Los elementos de la PDI serán chequeados (y eventualmente corregidos) tras tareas de curation o consultas directas a la base Solr.
3. ***Metadatos descriptivos***: las directrices DRIVER, desarrolladas en el ámbito del proyecto DRIVER (Digital Repository Infrastructure Vision for European Research), cuya versión actual 2.0 es de noviembre de 2008, son las directrices para proveedores de contenido sobre exposición de recursos textuales con el protocolo OAI-PMH utilizadas en Europa, y se tomaron como base en el Sistema Nacional de Repositorios Digitales en Argentina, para la conformación de una red interoperable de repositorios digitales en ciencia y tecnología, lo que valida la elección de estas directrices para la selección de los metadatos descriptivos a considerar obligatorios en el repositorio. La determinación de la existencia o la falta de alguno de estos metadatos será determinada por consultas directas a la base Solr.

¿Cómo aplicar la metodología?

La metodología implementada en la evaluación del RI podrá aplicarse del siguiente modo:

1) Perfilamiento automático de los archivos del repositorio y generación de reporte: esta tarea se llevará adelante, como se dijo, con la herramienta DROID. El reporte que ofrece DROID, como se verá a continuación, es global, es decir que analiza agrupamientos de archivos por formatos y no hace reportes sobre objetos individuales. Por lo tanto, para generar las tareas de corrección sobre el repositorio será necesario conocer qué ítems tienen problemas y para identificarlos se cruzará la información de DROID con consultas sobre el *assetstore*.

2 y 3) La revisión de la PDI de SEDICI y la revisión de metadatos descriptivos, así como la generación de un reporte que dé cuenta de la adecuación o no con los patrones establecidos (PDI de OAIS, y metadatos descriptivos según directrices DRIVER) serán analizados a través de un validador de desarrollo propio, que toma la forma de una tarea de curation y consultas sobre la base de datos Solr; con la PDI se utilizarán ambos métodos y con la DI sólo consultas.

4) Reporte final: el reporte se muestra en cada una de las secciones relativas al análisis de contenido y formatos, de la PDI y de la información descriptiva. En el capítulo final, referido a las conclusiones de este trabajo, se hará una breve reseña global basada en los reportes parciales detallados.

Descripción de herramientas

DROID

DROID (Digital Record Object Identification) es una herramienta de software desarrollada por The National Archives que permite llevar a cabo, de manera automática, la identificación por lotes de los formatos del conjunto de archivos suministrado por el usuario. Fue desarrollada por el Departamento de Preservación

Digital, como parte de actividades aún más amplias de TNA en relación a la preservación digital. DROID está diseñada para cumplir con el requisito fundamental de cualquier repositorio digital, en cuanto a la identificación de los formatos de todos los objetos digitales almacenados, y para vincular esa identificación a un registro central de información técnica sobre formatos.

DROID utiliza firmas internas y externas (como el registro PRONOM) para identificar y reportar el formato específico de cada archivo y la versión de los archivos digitales. Estas firmas se almacenan en un archivo XML, generado a partir de la información registrada en PRONOM. Firmas nuevas y actualizadas se añaden regularmente a PRONOM y DROID se puede configurar para la descarga automática de los archivos de firmas actualizados.

Para usar DROID, primero se selecciona el conjunto de archivos que se desea perfilar. En el caso del presente experimento, se ha trabajado con un directorio que reúne todas las carpetas y archivos del *assetstore* (al momento de inicio de esta evaluación). Una vez que se han seleccionado los archivos, se le indica a DROID que comience la ejecución y, de manera muy rápida, realiza el perfilamiento de los archivos. El perfil obtenido como resultado puede guardarse para formar parte de un conjunto de registros de perfiles. Los perfiles son guardados como archivos XML comprimidos en formato ZIP.

A partir de los perfiles, es posible generar un reporte completo (*comprehensive report*) que reúne la totalidad de los reportes detallados en la tabla 5.1 (los cuales a su vez pueden constituirse en reportes independientes). Como puede verse a partir de la tabla, DROID entrega la suficiente información del perfil de los archivos como para saber si hay (o no) que realizar acciones de migración de formatos, ya que muestra claramente el estado de los formatos en relación al registro internacional. Además, muestra también las versiones de los formatos, lo que a veces puede llevar a acciones más simples de migración.

Tabla 5.1: Reportes provistos por DROID

REPORTES DE DROID	DESCRIPCIÓN DE LOS REPORTES
File count and sizes	Suma, tamaño total, mínimo, máximo y promedio de todos los archivos del perfil.
Total count of files and folders	Suma de todos los archivos y carpetas del perfil.
Total unreadable files	Suma de todos los archivos ilegibles.
Total unreadable folders	Suma de todas las carpetas ilegibles.
File count and sizes by file extension	Suma, tamaño total, mínimo, máximo y promedio de todos los archivos del perfil, separados según su extensión.
File count and sizes by file format PUID	Suma, tamaño total, mínimo, máximo y promedio de todos los archivos del perfil, separados de acuerdo a sus formatos PUID (PRONOM Unique IDentifiers).
File count and sizes by MIME type	Suma, tamaño total, mínimo, máximo y promedio de todos los archivos del perfil, separados de acuerdo a sus MIME types.
File count and sizes by month last modified	Suma, tamaño total, mínimo, máximo y promedio de todos los archivos del perfil, separados de acuerdo al último mes en que fueron modificados. Los meses son representados por números 1 (enero) hasta 12 (diciembre).
File count and sizes by year last modified	Suma, tamaño total, mínimo, máximo y promedio de todos los archivos del perfil, separados por el último año en que fueron modificados.
File count and sizes by year and month last modified	Suma, tamaño total, mínimo, máximo y promedio de todos los archivos del perfil, separados por año y mes de última modificación.
Comprehensive breakdown	Reporte que combina todos los reportes precedentes.

JHOVE

Se ha utilizado como soporte adicional la herramienta JHOVE (JSTOR/Harvard Object Validation Environment) para la detección y validación de formatos. Se dice aquí "adicionalmente" porque si bien al inicio de este trabajo se pensó utilizar JHOVE de manera más integral, la versión instalada presenta algunos problemas (los cuales se explican al revisar los formatos en la sección dedicada al contenido y los formatos) y los archivos PDF, que constituyen el grueso del repositorio, fueron finalmente validados en su mayoría con la herramienta pdfaPilot, que también se expone en dicha sección. En particular, en relación al formato PDF, utiliza el módulo HUL. JHOVE puede ser integrado al flujo de trabajo de un repositorio en relación a la creación del SIP: por ejemplo, a partir del autoarchivo, JHOVE puede intervenir en el proceso y realizar la caracterización del formato.

La herramienta JHOVE2 realiza la validación de formatos, e incluso la extracción de metadatos técnicos, de manera automática. JHOVE2 realiza varias etapas para el procesamiento de los archivos. La última versión ha incorporado a DROID dentro de su estructura para extender los formatos que puede validar. Para una visión general de lo expuesto se presenta la figura 5.4 en la cual se puede observar la realización de identificación y validación de formatos en la pre-ingesta.

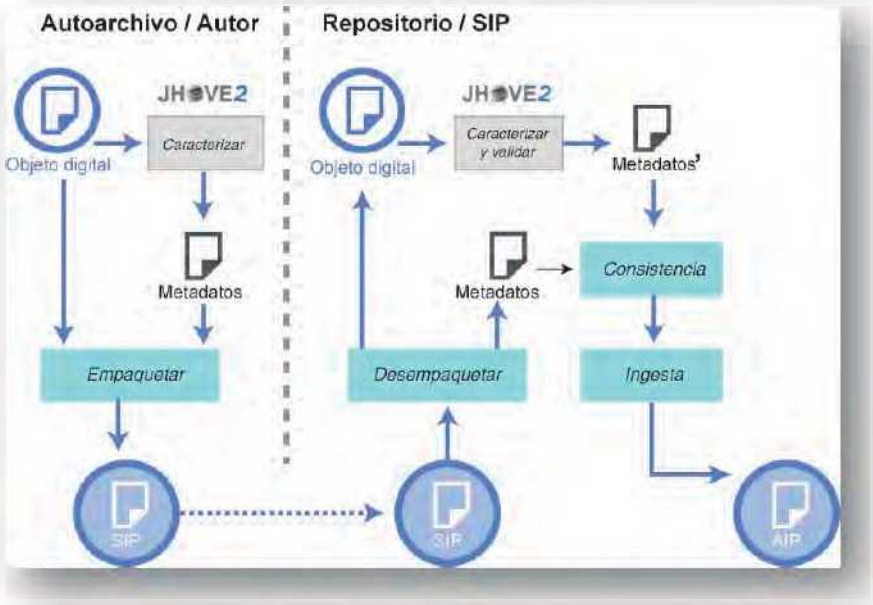


Figura 5.4: Intervención de JHOVE2 en la pre-ingesta y en la ingesta de OD

Fuente: JSTOR/Harvard Object Validation Environment (JHOVE), "Use cases".

Plato

Plato es una herramienta de planificación que sirve de apoyo a las decisiones de conservación de objetos digitales. Implementa un proceso sólido de planificación de la preservación e integra los servicios de caracterización de contenido, la acción de preservación misma y la comparación automática de objetos en una arquitectura orientada a servicios, para proporcionar el máximo apoyo a los esfuerzos de planificación de preservación de los repositorios y archivos.

Plato integra las siguientes herramientas y servicios:

- *DROID*, cuyo funcionamiento ya ha sido brevemente explicado en el apartado precedente.
- *JHOVE*, también explicado brevemente en los párrafos precedentes.
- *FITS* (File Information Tool Set) es un conjunto de herramientas de información de archivos que identifica, valida y extrae los metadatos técnicos de diferentes formatos de archivo. Fue creado por la Universidad de Harvard.
- *PRONOM*, registro técnico de formatos mantenido por The National Archives referenciado previamente.
- *Preserv2 (P2)* es el registro web semántico. Contiene información semántica útil en el proceso de preservación digital.
- *MyExperiment* es una plataforma de creación y compartición de flujos de trabajo, como Taverna.

Servicios que involucran acciones concretas de preservación:

- Los denominados flujos de trabajo (workflows) de Taverna, que sirven de base a las acciones de preservación.
- *MiniMEE*: servicio de migración de formatos incluido en Plato.

Si bien en la presente tesis no se plantea realizar un plan de preservación, el estudio de la herramienta Plato ha resultado de sumo interés para conocer cómo se diseña un plan de preservación, ver sus componentes, analizar otros planes, muchos de ellos con distintos fines generados por otras instituciones, pero muy importantes por los contenidos que mantienen y para, finalmente, pensar las futuras acciones previstas tras este trabajo.

Bibliografía del capítulo 5

Catalogue of Criteria for Trusted Digital Repositories (2006). Recuperado el 18 de junio de 2014, de http://www.dcc.ac.uk/resources/repository-audit-and-assessment/nestor.

Center for Research Libraries (CRL) (1949). Recuperado el 18 de junio de 2014, de http://www.crl.edu/

Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) (s/d). Recuperado el 18 de junio de 2014, de http://www.repositoryaudit.eu/.

Digital Curation Centre (DCC) (2005). Recuperado el 18 de junio de 2014, de http://www.dcc.ac.uk/.

Digital Repository Object Identification (DROID) (2006). Recuperado el 18 de junio de 2014, de http://droid.sourceforge.net.

Directory of Open Access Repositories (DOAR). Recuperado el 9 de junio de 2014, de www.opendoar.org

Directrices Driver 2.0. (2008). *Directrices para proveedores de contenidos. Exposición de recursos textuales con el protocolo OAI-PMH*. Recuperado el 23 de junio de 2014, de: http://www.driver-support.eu/documents/DRIVER_2_0_Guidelines_Spanish.pdf.

Dobratz, S.; Schoger, A. (2007). "The NESTOR Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification". *The Journal of Digital Information*, 8 (2).

File Information Tool Set (FITS) (s/d). Recuperado el 19 de junio de 2014, de http://code.google.com/p/fits/.

Hoeven, J. R.; Van Der Diessen, R. J.; Van En Meer, K. (2005). "Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects". *Journal of Information Science* 31(3), p. 196-208.

Investigating the Significant Properties of Electronic Content Over Time (InSPECT) (2007). Recuperado el 19 de junio de 2014, de http://www.significantproperties.org.uk/.

ISO 14721:2012. *Reference Model for an Open Archival Information System (OAIS)*. Junio de 2012.

ISO 16363:2012. *Space data and information transfer systems - Audit and certification of trustworthy digital repositories*. Recuperado el 19 de junio de 2014, de http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510.

JSTOR/Harvard Object Validation Environment (JHOVE). Recuperado el 15 de octubre de 2013, de http://sourceforge.net/projects/jhove/.

Lawrence, G. W.; Kehoe, W. R.; Rieger, O. Y.; Walters, W. H.; Kenney, A. R. (2000). *Risk management of digital information: a file format investigation. CLIR report 93*, Council on

Library and Information Resources. Recuperado el 15 de octubre de 2013, de http://www.clir.org/pubs/reports/pub93/reports/pub93/pub93.pdf.

MyExperiment. Recuperado el 19 de junio de 2014, de http://www.myexperiment.org/.

National Archives and Records Administration (NARA). Recuperado el 18 de junio de 2014, de http://www.archives.gov.

Nestor Working Group. Recuperado el 18 de junio de 2014, de http://www.dcc.ac.uk/resources/repository-audit-and-assessment/nestor#sthash.xuf56imV.dpuf.

PLANETS Project (2006-2010). Recuperado el 18 de junio de 2014, de http://www.planets-project.eu/.

Plato. Recuperado el 19 de junio de 2014, de http://plato.ifs.tuwien.ac.at/plato/index.jsf.

Plato. Sección de Documentación. Recuperado el 18 de junio de 2014, de http://www.ifs.tuwien.ac.at/dp/plato/documentation/.

Preserv2 (P2) (2007-2009). Recuperado el 19 de junio de 2014, de http://p2-registry.ecs.soton.ac.uk/.

PRONOM (2002). Recuperado el 19 de junio de 2014, de http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

Rothenberg, J. (1999). *Avoiding Technological Quicksand: Finding a Viable Technicaldation for Digital Preservation.* Council on Library and Information Resources. Recuperado el 15 de octubre de 2013, de http://www.clir.org/pubs/reports/rothenberg/contents.html.

Research Libraries Group (RLG) (1974). Recuperado el 18 de junio de 2014, de http://www.oclc.org/home.en.html.

Research Libraries Group (2002). *Trustworthy Repositories Audit & Certification (TRAC).* Recuperado el 18 de junio de 2014, de http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf?urlm=161690.

Taverna (2009). Recuperado el 19 de junio de 2014, de http://www.taverna.org.uk/.

United Nations Educational, Scientific and Cultural Organization (UNESCO) (2003). *Directrices para la preservación del patrimonio cultural.* Recuperado el 19 de junio de 2014, de http://unesdoc.unesco.org/images/0013/001300/130071s.pdf.

Capítulo 6 | Experimentación

Síntesis: Las acciones propuestas en este capítulo están vinculadas, de acuerdo a lo dicho hasta aquí, a los elementos constitutivos del paquete de información: 1) la información de contenido (CDO) y la información sobre la representación de ese contenido (RI); 2) la información descriptiva de preservación (PDI); 3) la información descriptiva (DI). Teniendo esto presente, es posible relatar y detallar a continuación el experimento realizado sobre el RI elegido.

Caso de estudio: SEDICI-DSpace

El experimento se realizó sobre el repositorio SEDICI, implementado con el software DSpace en su versión 1.8. En DSpace, cada objeto digital (archivos PDF, archivos de licencias, etc.) se almacena con un identificador de 38 dígitos, de forma tal que se puede obtener la localización exacta del archivo, pero no puede ser manipulado fuera del administrador de bitstreams debido a su complejidad. La carpeta de almacenamiento se define en el archivo de configuración [DSpace]/config/DSpace.cfg y esta carpeta, por defecto, se denomina *assetstore*. El assetstore en sí mismo puede verse como un conjunto de carpetas dentro del cual están los archivos.

La referencia de un ítem a sus archivos se encuentra en la tabla bitstream, en el campo *Internal_id*. Para este punto, es preciso observar el esquema de la figura 6.1.

Figura 6.1: Identificador de bitstream y organización de carpetas y archivos del *assetstore* en DSpace

Fuente propia.

Para identificar el bistream dentro del identificador puesto como ejemplo (157163341269448932463801798107206808 53), deben buscarse los seis primeros dígitos del identificador, que indican en qué subdirectorio de tercer nivel está el ítem (15 → 71 → 63) y el nombre real del archivo será 341269448932463801798107206808 53, pero ha desaparecido toda referencia al archivo de audio, que en este caso era xz.mp3.

Otro aspecto a tener en cuenta es el modelo de datos de DSpace (metadatos, *workflows*, estructura del repositorio, usuarios), que puede estar soportado por una base de datos Oracle o PostgreSQL. En el caso de SEDICI se tiene una base PostgreSQL. El modelo de datos se representa en la figura 6.2, a continuación:

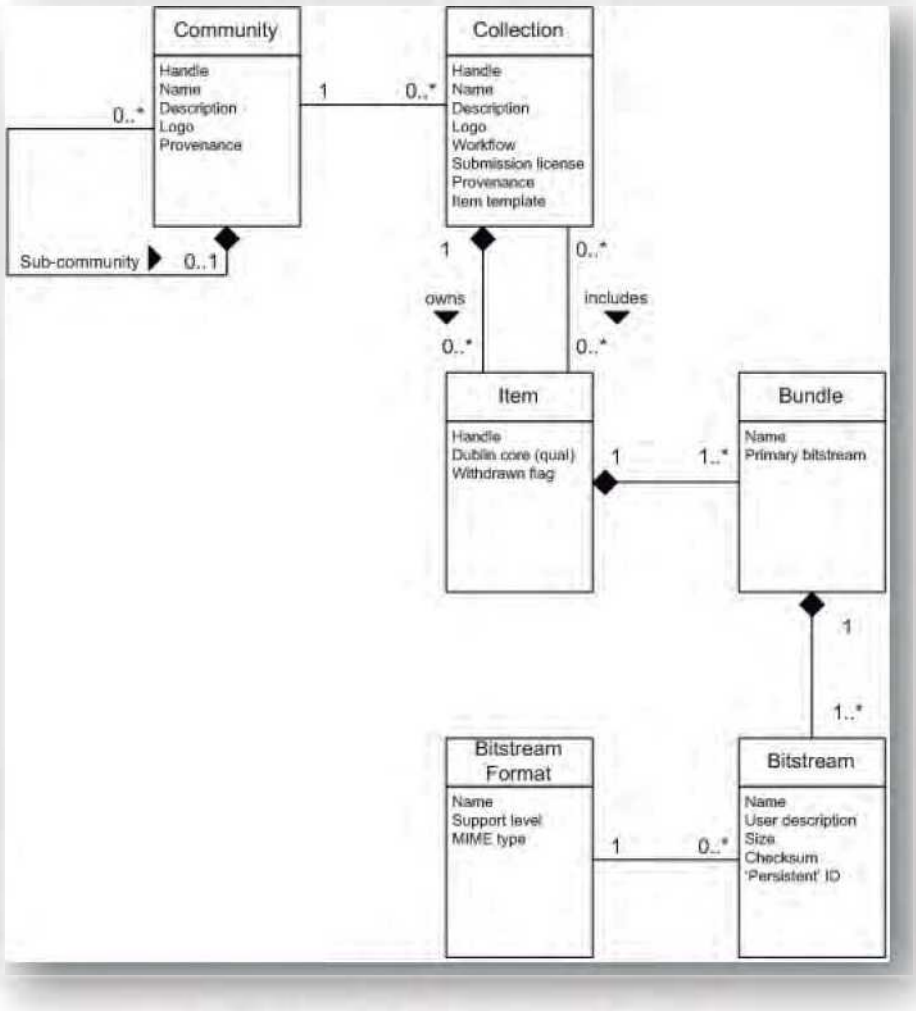


Figura 6.2: Modelo de datos de DSpace

Fuente propia.

En DSpace, las **comunidades** constituyen las jerarquías de más alto nivel y dentro de ellas pueden existir subcomunidades y/o colecciones, lo que permite segmentar la información de acuerdo a temas de interés, áreas de organización, unidades de investigación, etc. Las comunidades y subcomunidades no pueden contener ítems directamente. Cada comunidad o subcomunidad contiene **colecciones**, que a su vez contienen **ítems**. Un ítem en DSpace puede estar constituido por más de un bundle, y a su vez cada bundle tiene uno o más bitstreams (es decir, archivos), tal como se representa en la figura 6.3.

Figura 6.3: El ítem, sus bundles y bitstreams en DSpace

Fuente propia.

El ítem es la representación, en el modelo de datos, de cada elemento contenido en el repositorio. Cada ítem puede contener a uno o más bundles, como se dijo, mientras que un bundle puede pertenecer a un único ítem. Los bundles son agrupaciones de archivos dentro del ítem, que separan los diversos tipos de archivos de modo que DSpace pueda tratarlos en forma diferenciada. Las tablas de la base de datos vinculadas al ítem, al bundle y al bitstream dan cuenta de esto y, cuando se realiza la exportación de un ítem, DSpace genera un archivo denominado "Contents", que enumera los archivos que van en el ítem junto con la indicación del bundle en que van. He aquí un ejemplo:

license.txt	*bundle:LICENSE*
Archivo1.pdf	*bundle:ORIGINAL*
Archivo2.pdf	*bundle:ORIGINAL*

A su vez, los bundles se relacionan directamente con uno o varios bitstreams, y no son más que un conjunto de estos últimos agrupados bajo cierta lógica. Tal es así que, en la práctica, la mayoría de los ítems suelen tener asociados al menos alguno de estos bundles, como puede verse también en la figura 6.4:

- **Original**: es el bundle con el bitstream original depositado (los que se deben seleccionar para evitar listar varios bundle del mismo ítem).
- **Texto**: texto completo extraído del bitstream original, correspondiente al ítem. La extracción la realiza una aplicación de DSpace denominada Media Filter.
- **Licencia**: bundle que contiene la licencia que se subió en conjunto con el ítem al repositorio y especifica los derechos que se tienen sobre el mismo.
- **Licencia CC**: este bundle contiene la licencia de uso Creative Commons que especifica lo que el usuario final puede hacer con el ítem en cuestión, al acceder al mismo.

Figura 6.4: Un ítem, sus bundles y sus bitstreams visto desde la administración de DSpace

Fuente propia.

Un bitstream no es más que lo que su nombre indica, una secuencia lógica y ordenada de bits que representan al archivo propiamente dicho. Vale decir que el bitstream es el elemento digital en cuestión. Cada bitstream se asocia directamente con un formato de bitstream, y esto es así porque los sistemas de preservación implementados (muchas veces, como tareas de curación) requieren explícitamente que el usuario que sube cada ítem, defina el formato de archivo con el cual está trabajando, para poder conservarlo de mejor forma. Por último, cada formato de bitstream posee

un *MIME type* y un nivel de soporte únicos, que son los datos reales que suelen utilizar los sistemas de preservación.

1) Análisis del Contenido y la Representación

En este apartado, del paquete de información (particularizado en AIP para preservación) se analiza el contenido y su representación, como se resalta en verde en la figura 6.5:

Figura 6.5: Información de contenido del Paquete de Información, resaltada en verde

Fuente: Norma ISO 14721: 2012.

Los conceptos ya mencionados, vinculados a la organización del *assetstore* y al elemento ítem del modelo de datos de DSpace, son fundamentales para explicar cómo fue madurando este experimento. El primer experimento realizado contenía toda la estructura de carpetas del *assetstore*, además de los archivos en sí mismos. Esto trajo algunas dificultades al explorar los formatos con la herramienta DROID. El primer perfil obtenido se basó en un *assetstore* con una dimensión superior a 90.000 "objetos", de los cuales alrededor de la mitad eran carpetas y el resto bitstreams. Tras las primeras pruebas, se decidió cambiar las preferencias por defecto de DROID para

lograr la posibilidad de que evaluara el algoritmo MD5 de los archivos, como se ve en la figura 6.6:

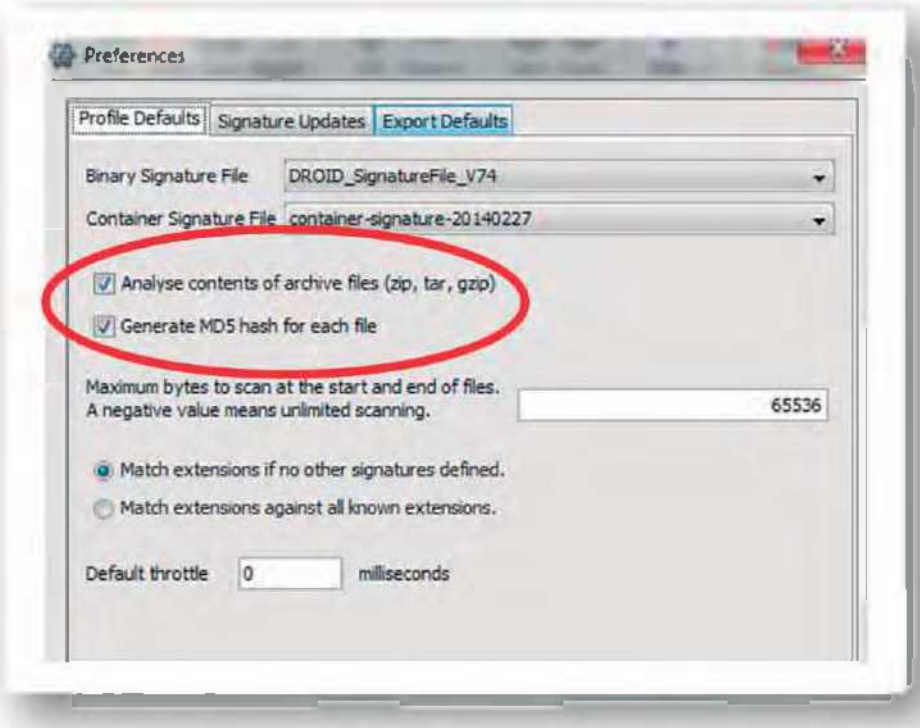


Figura 6.6: Captura de pantalla de la pestaña de preferencias de DROID

Fuente propia.

Se obtuvo un primer perfil que, desplegado, mostraba muchas filas sin evaluación: he aquí la primera dificultad derivada de estar acoplando en el análisis la estructura de carpetas del *assetstore*, sumada a la propia dificultad de un tamaño tan grande de archivos a evaluar, como puede verse en la figura 6.7:

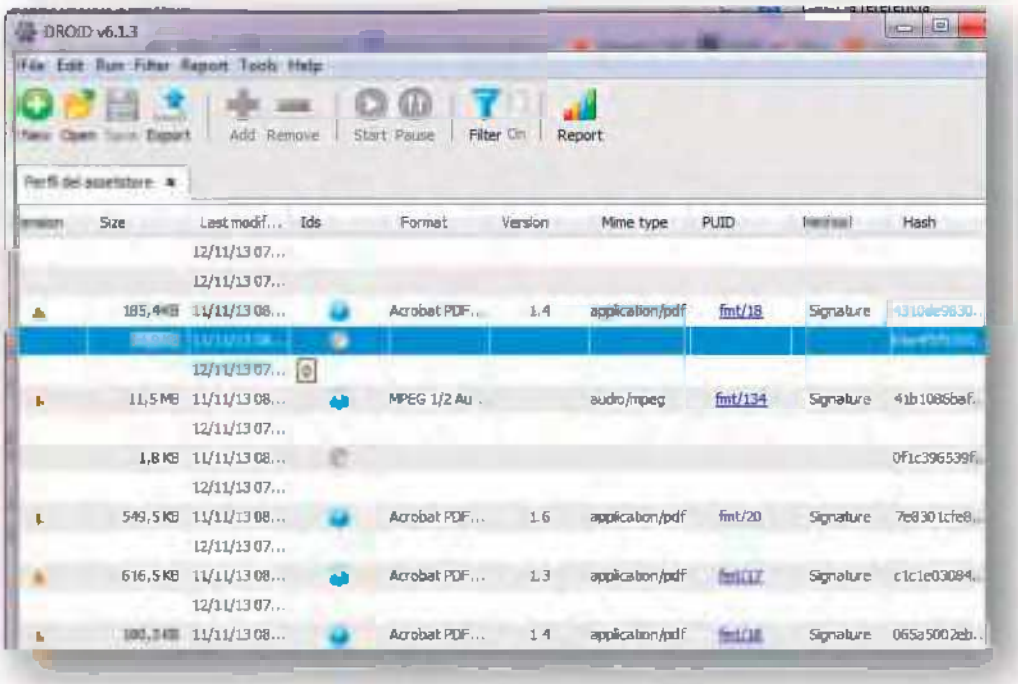


Figura 6.7: Captura de pantalla del archivo de perfiles generado por DROID, con elementos sin evaluación (resaltado en celeste)

Fuente propia.

A pesar de esto, el análisis de los MD5 permitió descubrir muchos valores de checksum repetidos, en algunos casos reales, es decir, archivos duplicados, pero en otros falsos duplicados, producto de tener en cuenta todos los bundles del ítem. Si se observan los bundles del ítem ejemplificado en la figura 6.4, se puede ver que hay dos bundles: *bundle text* y *bundle original*. El bundle text es un bitstream resultado de aplicar al bitstream del bundle original el citado componente de DSpace llamado Media Filter, que se encarga de extraer el texto completo del archivo que está en el bundle original. Si este componente se aplica sobre archivos PDF producto de una digitalización, pero éstos son mantenidos como imágenes (es decir, sin el reconocimiento óptico de caracteres u OCR) el media filter siempre extrae el mismo texto y por lo tanto el checksum de todos los archivos con esa misma condición, resultará igual, generando los falsos duplicados.

El perfil y el reporte de este experimento de prueba no resultaron inútiles, ya que se detectaron también duplicados reales de distinto tipo. Se complementó entonces el perfil de DROID con una consulta a la base de datos para identificar las correspondencias entre los MD5 y el handle y el ID interno de los bitstreams para

identificar de manera sencilla los ítems a depurar.

El caso de los falsos duplicados marcó la necesidad de, en los casos de materiales digitalizados, realizar el OCR, todas las veces que fuera posible, de modo de tener archivos que no fueran imágenes tras los procesos de digitalización. Esta determinación, a su vez, generó una importante tarea de priorización de los PDF para el reconocimiento de caracteres. Esta priorización se realiza actualmente de forma automática mediante una herramienta que recorre todo el *assetstore*, y para cada documento se fija si es un PDF o si está "corrupto". En caso de que el archivo esté corrupto, el script genera una notificación de error, y en el caso de que sí sea un PDF se comprueba que efectivamente tenga OCR. Para los PDF que no tienen OCR, se les hace entonces una detección de caracteres de máximo 10 páginas (se obtienen 10 páginas del medio del archivo en JPG, y en caso de no ser posible se toman las 10 primeras, o las que tenga el PDF) y se calcula un promedio (total de caracteres reconocidos/cantidad de páginas procesadas). En base a ese promedio, se ubicará el PDF en el archivo de prioridades (mientras mayor el número, más prioridad).

Las dificultades precedentes y los logros del experimento, hasta este punto, quedaron plasmados en tres reportes iniciales, sumamente extensos, cuyos respectivos archivos digitales llevan los nombres: "Perfil del assetstore completo"; "Perfil completo y prueba de duplicados" y "Casos de checksum repetidos". Debido a su complejidad y extensión no se incluye copia digital de los mismos, puesto que la experimentación continuó usando las mismas herramientas, y los reportes posteriores, de menor extensión, parecen suficientes a los efectos de este trabajo.

Este primer experimento planteó, por todo lo dicho, la realización de una nueva prueba descartando las carpetas y considerando sólo el bundle original de cada ítem. El nuevo *assetstore*, así, sólo tiene un nivel de carpetas y dentro de cada carpeta sólo archivos, y particularmente no todos los ítems del repositorio, sino aquellos que al momento de esta segunda experimentación, llevada a cabo en diciembre de 2013, contaban con al menos un bitstream en el bundle original. Generar un nuevo *assetstore* sin los tres niveles del anterior simplificó grandemente la tarea y la extensión del perfil y el reporte generados por DROID.

Metodología y resultados obtenidos con el experimento final

Se ingresó el nuevo *assetstore* completo en DROID, constituido ahora por 19.230 objetos, de los cuales 18.522 son archivos y el resto el primer nivel de carpeta de cada conjunto de archivos. En la figura 6.8 puede observarse una captura de pantalla del perfil: allí se destacan dos carpetas, rotuladas 100 y 1000, y sus archivos ya perfilados.

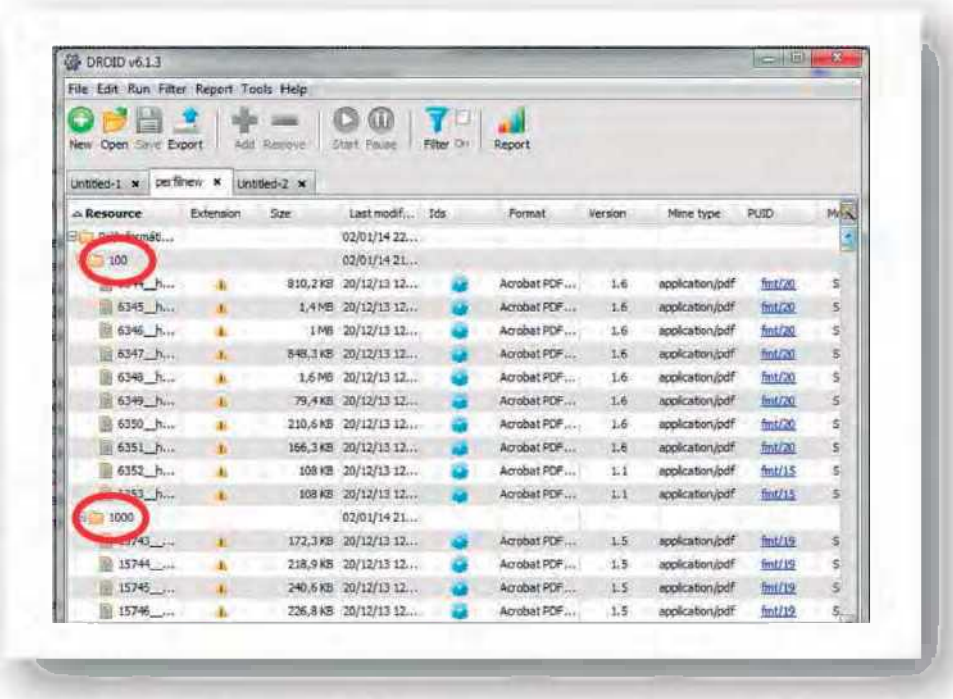


Figura 6.8: Captura de pantalla del archivo de perfiles generado por DROID con el nuevo *assetstore*

Fuente propia.

Los resultados obtenidos pueden observarse en los distintos anexos de esta tesis. Así, en el anexo bajo el título "PRONOM: Detailed Report", a modo de ejemplo, puede verse la descripción (File Format Information) del PUID fmt/20 (el primero de la lista que aparece en la figura 6.8), que se corresponde al archivo del siguiente ítem de SEDICI: http://sedici.unlp.edu.ar/handle/10915/6344. Este ítem es un artículo cuyo archivo digital se encuentra en formato Acrobat PDF 1.6, de un tamaño de 810.2 KB; más a la derecha de la vista, se encuentra el checksum de ese archivo calculado con el algoritmo MD5. El perfil completo elaborado por DROID fue exportado a una planilla Excel en la cual se filtraron las carpetas, dejando solamente los archivos y el MD5. Se incluye dicho reporte en el CD de esta tesis, bajo el título "Perfilsolofiles marcado MD5

v2.xls".

El perfil elaborado por DROID, a su vez, puede condensarse en el reporte denominado "Comprehensive Breakdown" que se agrega como parte de los anexos de esta tesis, y que puede sintetizarse en la siguiente lista de archivos:

- **47 archivos Macromedia Flash 5**, formato que se corresponde con el PUID fmt/108 de PRONOM.
- **4 archivos Windows Bitmap 3.0**, formato que se corresponde con el PUID fmt/116 de PRONOM.
- **5 archivos MS PPT 1997-2002**, correspondiente al PUID fmt/126 de PRONOM.
- **583 archivos MPEG ½ Audio Layer 3**, correspondiente al PUID fmt/134 de PRONOM.
- **17359 archivos PDF**, de los cuales:
  - **1 archivo Acrobat PDF 1.0**, formato que se corresponde con el PUID fmt/14 de PRONOM.
  - **137 archivos Acrobat PDF 1.1**, formato que se corresponde con el PUID fmt/15 de PRONOM.
  - **1021 archivos Acrobat PDF 1.2**, formato que se corresponde con el PUID fmt/16 de PRONOM.
  - **1468 archivos Acrobat PDF 1.3**, formato que se corresponde con el PUID fmt/17 de PRONOM.
  - **5599 archivos Acrobat PDF 1.4**, formato que se corresponde con el PUID fmt/18 de PRONOM.
  - **2426 archivos Acrobat PDF 1.5**, formato que se corresponde con el PUID fmt/19 de PRONOM.
  - **6707 archivos Acrobat PDF 1.6**, formato que se corresponde con el PUID fmt/20 de PRONOM.
  - **301 archivos Acrobat PDF 1.7**, formato que se corresponde con el PUID fmt/276 de PRONOM.
- **5 archivos Ogg Verbis Codec Compressed Multimedia**, formato que se corresponde con el PUID fmt/203 de PRONOM.
- **1 archivo PocketMobi**, formato que se corresponde con el PUID fmt/396 de PRONOM.

- **188 archivos JPEG**, de los cuales:
  - **2 archivos JPEG File Interchange Format 1.0**, formato que se corresponde con el PUID fmt/42 de PRONOM.
  - **32 archivos JPEG File Interchange Format 1.01**, formato que se corresponde con el PUID fmt/43 de PRONOM.
  - **154 archivos JPEG File Interchange Format 1.02**, formato que se corresponde con el PUID fmt/44 de PRONOM.
- **1 archivo ePub Format**, formato que se corresponde con el PUID fmt/483 de PRONOM.
- **4 Audio/Video Interleaved Format**, formato que se corresponde con el PUID fmt/5 de PRONOM.
- **1 archivo Adobe Illustrator 12.0**, formato que se corresponde con el PUID fmt/561 de PRONOM.
- **2 archivos Microsoft Excel 97 Workbook 8**, formato que se corresponde con el PUID fmt/61 de PRONOM.

Significado e interpretación de los datos

A partir del reporte generado por DROID se revisó el registro PRONOM correspondiente a cada formato, para saber si existían formatos que pusieran en riesgo la preservación de alguno de los ítems del repositorio. Afortunadamente, ninguno de los formatos existentes hace peligrar los contenidos; sin embargo, es importante destacar algunas cuestiones significativas:

- Existe una enorme preeminencia del MIME type Aplication/PDF, formato que está presente en el RI en versiones que van desde la 1.0 a la 1.7. Es el MIME type más importante, dada la cantidad de ítems que tiene el repositorio en este formato. El registro PRONOM muestra para cada versión las preeminencias (los mejores formatos dentro de ese MIME type) y si es un superset de otro formato dentro del mismo MIME type. La preeminencia de este formato determinó una gran dedicación de tiempo para su estudio. Este y el resto de los formatos son analizados en detalle en el apartado "Análisis de los formatos surgidos en el relevamiento sobre SEDICI'.
- El perfil de DROID de cada archivo brindó la posibilidad de contar con el MD5 y a partir de esto detectar numerosos errores de archivos repetidos.

- El análisis del MD5 llevó a la necesidad de familiarizarse con el sistema de gestión de incidencias de SEDICI, soportado por Redmine, para reportar los errores, indicar las acciones de mejoramiento y dar seguimiento continuo a los cambios. Se generaron 87 tickets (o tareas) dedicados al mejoramiento de la calidad, en su mayoría para depuración de ítems repetidos, siguiendo el orden de formatos ofrecido en el perfil de DROID. Además, en cada ticket se indicó la colección que contenía los ítems con problemas dentro de SEDICI para realizar el trabajo de manera más ordenada. La figura 5.13 es una vista del sistema de gestión de incidencias de SEDICI, al que se le ha aplicado un filtro para que muestre los tickets generados exclusivamente en este proceso.
- Del mismo modo, se hizo una primera determinación de ítems sin bitstream y sin localización electrónica. Al momento de realizarse este análisis, había **5652** ítems sin bitstream ni localización electrónica (enlace), de los cuales **5381** son recursos textuales (**2711** tesis y **2585** artículos en todos sus subtipos). Se generaron 16 tickets sobre este problema. La figura 6.9 da una vista de esos tickets y de su estado. El análisis de los ítems sin bitstream y sin localización (electrónica y/o física) es abordado con gran detalle en el punto de análisis de la PDI de este trabajo. Se incluye adjunto en CD de esta tesis este reporte bajo el título "Informe de ítems sin archivos con y sin links.xls".

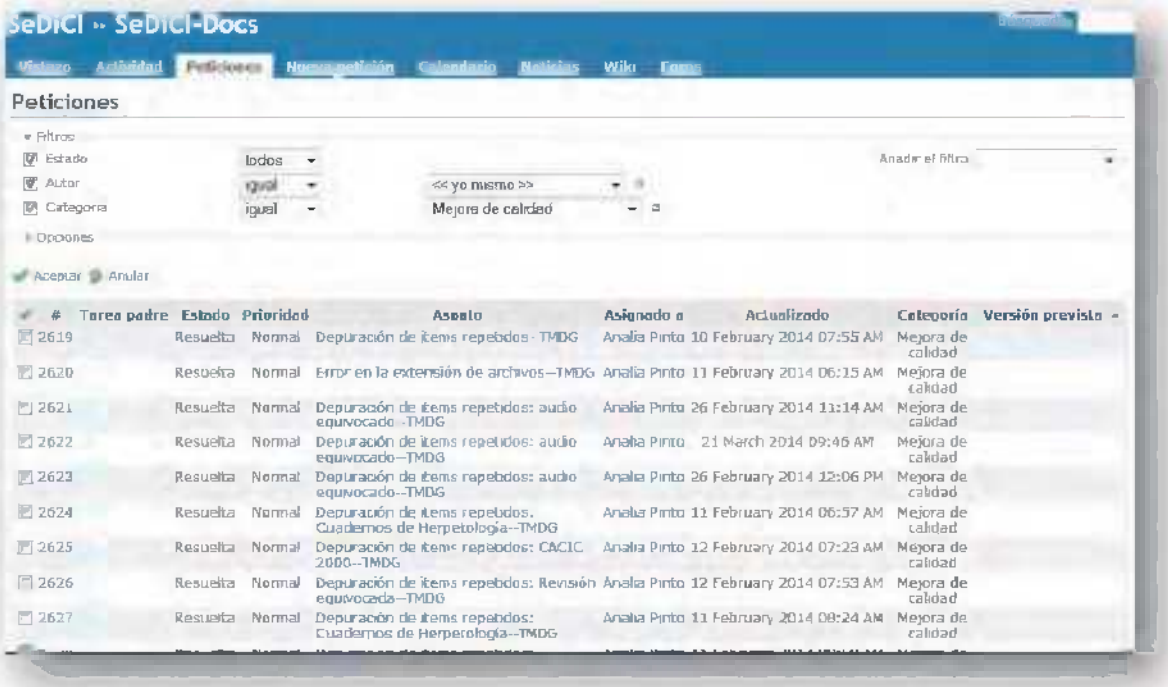


Figura 6.9: Captura de pantalla que muestra los tickets generados en el sistema de gestión de incidencias de SEDICI

Fuente propia.

Figura 6.10: Captura de pantalla de los tickets generados y su estado

Fuente propia.

Sobre este trabajo se elaboró un documento de respaldo, además del propio sistema de gestión de incidencias, para ir organizando el agregado de tareas. Dicho archivo se llama "Revisión de formatos y errores en sedici.docx", pero no se anexa a la tesis porque sólo ha servido para organización interna.

La figura 6.11 muestra, de manera simplificada, los formatos presentes en el repositorio, lo que permite observar la gran prevalencia del formato PDF. Está claro que, en el caso de SEDICI, el mayor problema, si es que lo hubiere, es el vinculado a los PDF, que constituyen un porcentaje muy elevado del total de los formatos presentes en el repositorio. Esto sugiere la necesidad de focalizar el análisis sobre este formato en particular.

Figura 6.11: Distribución de los formatos presentes en el repositorio SEDICI al momento de la experimentación

Fuente propia.

Para el análisis de los formatos se va a aludir, entre otros documentos, a la *Guía de buenas prácticas para la construcción de objetos digitales* realizada por la National Information Standards Organization (NISO). Esta obra lista los principios fundamentales que rigen la creación y gestión de objetos digitales. Para NISO, una colección digital está constituida por objetos digitales seleccionados y organizados de

modo tal de facilitar su localización, acceso y uso. Estos objetos, con sus metadatos y las interfaces de usuario en conjunto, posibilitan a los usuarios experimentar la colección. En este sentido, una *buena colección* está regida por una serie de principios:

1. La colección digital se crea siguiendo una política de desarrollo predeterminada.
2. La colección debe tener una descripción adecuada que permita que el usuario descubra sus características, alcance, formato, restricciones de acceso, derechohabientes y cualquier información significativa para asegurar su autenticidad, integridad e interpretación.
3. Una buena colección está sometida a tareas de curación, es decir que todos sus recursos son gestionados durante su ciclo de vida completo.
4. Una buena colección tiene una disponibilidad amplia y esquiva cualquier impedimento para su uso. Las colecciones deben ser accesibles para personas de diferentes capacidades y pasibles de ser usadas en tecnologías adaptadas a la diversidad de usuarios.
5. Una buena colección debe respetar los derechos de propiedad intelectual.
6. Debe contar con mecanismos y herramientas capaces de dar cuenta del uso de los datos, y permitir almacenar algunas medidas de utilidad.
7. Debe ser interoperable.
8. Debe integrar a los usuarios en el flujo de trabajo.
9. Debe ser sostenible a lo largo del tiempo.

Los nueve principios enunciados tienen una enorme importancia para un repositorio digital. En particular, en esta parte del trabajo dedicado a la perduración en el tiempo de los objetos digitales y su legibilidad, queda claro que el principio 3 es central y que el principio 4 se mantendrá como un objetivo más ambicioso y a realizar a futuro. El punto 3 implica que la colección deberá gestionarse adecuadamente para asegurar que la información original permanezca y pueda ser interpretada, lo que requiere diversas operaciones a lo largo de todo el ciclo de vida de los objetos digitales. Para ello, se deben repasar los formatos encontrados en el relevamiento, e indicar las acciones de migración, recomendando el uso de estándares, en lo posible abiertos.

Los formatos propietarios (o cerrados) son aquellos que tienen restricciones legales de uso, y no resulta posible intervenir en su implementación ya que, por lo general, sus

especificaciones no son públicas, y/o están sujetos al pago de licencias a empresas privadas. Por el contrario, los formatos libres (o abiertos) son aquellos que poseen una especificación de referencia bajo una licencia libre y pueden ser adaptados para su uso sin restricciones legales. Existen varios formatos que podrían llamarse "híbridos", pues se hacen públicas sus especificaciones (o sea, la estructura y la lógica interna de los archivos guardados con ese formato), pero los programas que leen esos formatos deben pagar una licencia para hacer uso de ellos en forma legal.

2) Análisis de los formatos surgidos del relevamiento sobre SEDICI

El presente análisis busca anticipar los riesgos para la preservación de los objetos y programar las acciones de migración y el futuro Plan de Preservación del Repositorio Institucional SEDICI. El análisis abarca los formatos presentes en el repositorio, según el detalle que se diera en páginas precedentes. Además de analizar el formato, ahora se elegirán los formatos que se consideran más útiles a los fines de la preservación y se presentarán las acciones y las herramientas para llevar adelante las acciones propuestas.

Portable Document Format (PDF)

Es el formato mayoritario en SEDICI y se trata de un formato de archivos usado para representar documentos en forma independiente del sistema operativo, del hardware y del software de aplicación. Cada archivo PDF encapsula una descripción completa de una distribución y organización fija y plana del documento, incluyendo el texto, las fuentes, gráficos y cualquier otra información necesaria para poder mostrarlo.

Si bien la especificación de PDF se hizo pública por Adobe Systems en 1993, siguió siendo un formato propietario controlado por Adobe hasta que se liberó como estándar abierto en julio de 2008, publicado como ISO 32000-1:2008. Con la versión 1.7 Adobe siguió extendiéndolo, pero dentro de la misma versión base. En la página web de Adobe pueden verse las especificaciones de las distintas versiones.

Cada nueva versión contiene todas las características de las anteriores e incorpora las nuevas características, excepto en los casos en que fueron eliminadas deliberadamente por Adobe. Muchos de los cambios introducidos generan mejoras en

la seguridad y encriptación de los documentos, o soporte para incrustar nuevos formatos de archivos[23].

La tabla 6.1, que se muestra a continuación, resume algunos agregados considerados importantes en cada versión.

Tabla 6.1. Características de las distintas versiones de PDF

Acrobat 1 1993, PDF 1.0	PDF 1.0 incorpora la mayoría de las funciones ofrecidas por el lenguaje de descripción de página PostScript de nivel 2: funciones básicas de texto, gráficos vectoriales y gráficos de mapa de bits.
Acrobat 2 1994, PDF 1.1	Esta versión admite el espacio de color Lab y CalRGB. También es compatible con las fuentes TrueType.
Acrobat 3 1996, PDF 1.2	Esta versión permite la separación de color y es compatible con las fuentes Unicode y CID (chino, japonés y coreano). También admite la compresión ZIP.
Acrobat 4 1999, PDF 1.3	PDF 1.3 contiene el modelo de gráficos de PostScript de nivel 3 completo. Permite espacios de color multicanal (DeviceN) y es compatible con perfiles ICC para la reproducción fidedigna de colores. Introduce sombras suaves y cuadros de geometría de página que son útiles para procesos de preimpresión (Trimbox, CropBox y Bleedbox).
Acrobat 5 2001, PDF 1.4	Desde esta versión, los archivos PDF pueden contener transparencias. Esta versión también presenta el "PDF etiquetado" (= PDF estructurado), que permite accesibilidad de contenido. Las opciones de seguridad mejoran con esta versión. Además, el tipo de compresión de imagen JBIG2 está soportado.
Acrobat 6 2003, PDF 1.5	Con esta versión, los documentos PDF pueden contener capas (también llamadas "contenido opcional"). Se admite la compresión de imágenes JPEG2000.
Acrobat 7 2004, PDF 1.6	Esta versión es compatible con las fuentes OpenType. Con esta versión, se pueden insertar contenidos 3D.
Acrobat 8 2006, PDF 1.7	El soporte completo de Unicode en esta versión simplifica la creación de vínculos con independencia del idioma. La nueva función "Paquete PDF" de Acrobat permite que varios documentos PDF independientes sean enviados como un solo archivo, si bien el destinatario necesita tener Acrobat o Reader 8.
Acrobat 9 2008, PDF 1.7 Extensión nivel 3	La especificación 1.7 extensión de nivel 3 proporciona entre otras cosas un nuevo tipo de comentario.

Fuente: Wikipedia.

Teóricamente, no debería haber ningún problema al convertir archivos PDF desde una versión anterior hacia una más nueva, e incluso podría verse una reducción del tamaño en el caso de que se optimicen los archivos multimedia. Sin embargo, la compatibilidad hacia versiones previas alcanza sólo hasta la versión 1.3, como puede

[23] Para una especificación completa de los cambios entre versiones puede consultarse en Wikipedia la tabla que expone y explica los agregados de cada nueva versión a partir de la especificación 1.0: Portable Document Format. Recuperado el 24 de junio de 2014, de http://en.wikipedia.org/wiki/Portable_Document_Format#Adobe_specifications.

verse en la propia página de Acrobat. También, a partir de la versión 1.4, se ha incorporado una capa especial para el texto extraído a través de procesos de digitalización que incorporan el reconocimiento óptico de caracteres (OCR).

El formato PDF, a pesar de sus ventajas, no garantiza la reproductibilidad a lo largo del tiempo, ni tampoco la independencia de software y hardware. Para garantizar ambos principios es necesario limitar y expandir el formato estándar PDF para permitir la reproducción exacta del contenido. Así nace el "nuevo" estándar PDF/A que define requisitos que son necesarios y otros que están prohibidos.

El comité de ISO TC171 utilizó el PDF 1.4 como base para el estándar PDF/A-1. La referencia ISO se utilizó en Adobe para el producto Acrobat 5, y por lo tanto esta versión cumple la totalidad de los requerimientos de referencia. No obstante, está claro que hay diferencias entre la referencia para la versión 1.4 y el estándar PDF/A-1. Por ejemplo, PDF 1.4 permite la integración de audio y video que no está permitida en PDF/A-1. La norma ISO 19005-1 especifica cómo usar el formato PDF 1.4 para la preservación a largo plazo y el cumplimiento del estándar PDF/A-1. Es decir, que el estándar especifica qué características del PDF 1.4 son necesarias, cuáles recomendadas y cuáles no permitidas. La versión 1.7 es la que se encuentra estandarizada.

Subconjuntos estandarizados de PDF

Los siguientes subconjuntos especializados de la especificación de PDF han sido estandarizados como estándares ISO:

- ***PDF/X*** (desde 2001, series de estándares ISO 15929 e ISO 15930). La X viene por Exchange (PDF for Exchange), por la tecnología Gráfica-Prepress digital data exchange. PDF 1.3, 1.4 y luego 1.6. Muy usado por la industria gráfica.
- ***PDF/A*** (desde 2005, estándares ISO 19005): "PDF para archivar". Gestión de Documentos - Formato de archivo de documento electrónico para preservación a largo plazo. Se basó en PDF 1.4 y luego también en ISO 32000-1, PDF 1.7.
- ***PDF/E*** (desde 2008, ISO 24517): "PDF para ingeniería", Gestión de documentos, formato de documento de ingeniería usando PDF, basado en PDF 1.6.
- ***PDF/VT*** (desde 2010, ISO 16612-2): "PDF para intercambio de datos variables e impresión transaccional".
- ***PDF/UA*** (desde 2012, ISO 14289-2): "PDF para Acceso Universal", Aplicaciones

de gestión de documentos, formato de archivo de documento electrónico mejorado para accesibilidad, basado en ISO 32000-1, PDF 1.7 formalizado en ISO 19005-2: 2011: Document management - Electronic document file format for long term preservation - Part 2: ISO 32000-1 (PDF/A-2).

PDF 1.7

La documentación final revisada para PDF 1.7 se aprobó por ISO en enero de 2008 y se publicó en julio de ese año. Los estándares anteriores ISO PDF (PDF/A, PDF/X, etc.) están pensados para usos más especializados. ISO 32000-1 incluye toda la funcionalidad documentada en las especificaciones PDF de Adobe para versiones desde la 1.0 a la 1.6. Algunas funcionalidades de versiones previas fueron removidas por Adobe, y tampoco se incluyeron en PDF 1.7. ISO 32000-1:2008 especifica un formato digital para representar documentos electrónicos para permitir a los usuarios intercambiar y visualizar documentos electrónicos independientemente del entorno en el cual fueron creados, en el entorno de visualización, o en el entorno de impresión.

Sobre PDF/A

PDF/A es un estándar para codificar documentos en un formato "impreso", que es portable entre sistemas y ampliamente usado para distribución y archivado de documentos. Sin embargo, la pertinencia de un archivo PDF para preservación depende de las opciones elegidas cuando el PDF fue creado: en particular, si se embebieron las fuentes necesarias para renderizar el documento, si se usa encriptación y si se preserva información adicional del documento original, más allá de lo que se precisa para imprimirlo.

El estándar PDF/A no define una estrategia de archivado o los objetivos de un sistema de archivado. Sí identifica un "perfil" para documentos electrónicos que asegura que los documentos pueden ser reproducidos exactamente de la misma manera durante años. Un elemento clave para esta reproductibilidad es que los documentos PDF/A deben ser 100% auto-contenidos: esto significa que toda la información necesaria para mostrar el documento de la misma manera cada vez, debe embeberse dentro del archivo. Esto incluye (pero no se limita a) todo el contenido (texto, imágenes rasterizadas, gráficos vectorizados), fuentes, información de color,

etc. Un documento PDF/A no puede jamás depender de información de fuentes externas (por ejemplo, programas de fuentes o *streams* de datos), aunque se permite que tengan anotaciones (como hipertextos) que enlacen a documentos externos.

Otros elementos de la compatibilidad PDF/A incluyen:

- El contenido de audio y video está prohibido.
- Java script y enlaces a archivos ejecutables están prohibidos.
- Todas las fuentes deben estar embebidas, y también deben ser legalmente embebibles para renderización ilimitada y universal. Esto también se aplica a las llamadas fuentes estándares PostScript, como Times o Helvetica. Esto significa para un usuario poder abrir el documento y que los caracteres se muestren de manera correcta (de aquí a X años) aunque no tenga esa tipografía en su computadora.
- Los espacios de colores deben ser especificados de una manera independiente del dispositivo.
- Se prohíbe la encriptación.
- El uso de metadatos basados en estándares se mantiene.
- Las referencias a contenidos externos están prohibidas.
- La compresión de imágenes LZW y JPEG2000 están prohibidas en PDF/A1, pero JPEG 2000 se permite en PDF/A2.
- Capas y objetos transparentes están prohibidos en PDF/A1 pero no en PDF/A2.
- Firmas digitales provisionales se permiten en PDF/A2.
- Los archivos embebidos están prohibidos en PDF/A1, pero PDF/A2 permite embeber archivos PDF/A, lo cual permite archivar múltiples documentos PDF/A en un solo archivo. PDF/A3 permite embeber cualquier formato como XML, CSV, CAD, archivos de Word, planilla de cálculo, otros PDF/A, etc. como objetos archivados completos.

PDF/A1 posee dos niveles de cumplimiento:

PDF/A-1a aplica corrección semántica y estructura. Cada carácter debe tener su equivalente Unicode. La estructura se expresa por medio de etiquetas.

PDF/A-1b aplica integridad visual.

Cualquier documento que cumple PDF/A1-a cumple todos los requisitos de PDF/A1-b, que es menos estricto. PDF/A1-a, como se especifica más arriba, requiere que el documento se encuentre estructurado y que utilice caracteres Unicode (lo que en el estándar A2 constituye el A2-u), que son requisitos de accesibilidad. Solicitar un PDF/A-1a ofrece más garantías de que sea un PDF accesible, pero no asegura que sea accesible. Todas las acciones que se deben llevar a cabo para hacer un PDF accesible se harán independientemente del tipo de PDF que sea.

Un PDF estructurado es aquél que especifica la secuencia exacta de su contenido. En el caso de diseños con columnas, sería imposible para el software determinar automáticamente el orden del contenido. Los autores de los documentos deben especificar esta secuencia. Cada elemento en un archivo PDF etiquetado tiene una etiqueta que contiene información sobre el tipo, la posición y el contenido. Estas etiquetas se utilizan para definir la estructura del documento como se muestra en la figura 6.12.

Figura 6.12: Captura de pantalla de las etiquetas que contiene un PDF etiquetado

Fuente propia.

En 2011, ISO liberó el estándar PDF/A-2 que, al estar basado en las versiones superiores a la 1.4 de PDF, incluye nuevas facilidades; es más, PDF/A-2 ya no está

basado en una versión de Adobe PDF sino que se basa en el estándar ISO 32000-1. Entre las facilidades que acopla, cabe citar:

1. Posibilidad de usar compresión JPEG2000, que resulta muy beneficiosa para documentos que son escaneados, particularmente cuando se trata de mapas, libros o cualquier otro contenido con color.
2. Posibilidad de embeber archivos PDF/A, armando así una colección.
3. Transparencia: posibilidad de uso de sombras, *cross fade* y resaltados.
4. Conformidad con el estándar PDF/A2-u (unicode) que simplifica la búsqueda de texto, su copia y reconocimiento en caso de digitalización.

Estado de situación en SEDICI de acuerdo al reporte de DROID

El reporte demuestra la existencia en SEDICI de una mayoría de archivos PDF, como se dijera (95% de acuerdo a la figura 5.15), entre las versiones 1.0 y 1.7. La mayoría de ellos no cumple los requisitos para ser *PDF/A-compliant* por varias razones, a saber:

1. Muchos de los PDF archivados en SEDICI fueron producidos antes de la publicación del estándar PDF/A y el desconocimiento de las tareas apropiadas de preservación, junto a la falta de herramientas para validación y conversión masiva de PDF a PDF/A, no colaboraron en la apropiación de este formato.
2. Las pruebas realizadas con JHOVE sobre toda la gama de versiones de PDF han mostrado una debilidad en la herramienta, en particular para este formato. El módulo HUL de JHOVE sólo detecta apropiadamente el archivo en los casos de las versiones de PDF 1.4, 1.5 y 1.6; en muchos casos, al no reconocer los formatos de PDF previos, no selecciona el módulo y deja la validación en manos del módulo Bitstream, que resulta incapaz de reconocer adecuadamente al PDF.

Por esta última razón se analizó la conversión de los archivos con otras herramientas y se obtuvieron pruebas muy satisfactorias con pdfaPilot, una herramienta de la empresa Callas, que realiza la conversión a PDF/A1, PDF/A2 y PDF/A3 en todas sus variantes. Esta aplicación encuentra los problemas en los PDF/A mal formados y genera un reporte completo acerca de ellos.

Se hace constar que se evaluaron PDF de todas las versiones presentes en SEDICI,

esto es: PDF nacidos digitales y PDF obtenidos como resultado de tareas de digitalización y procesamiento. También es necesario tener en cuenta que para crear archivos PDF/A la situación es distinta si se trata de documentos de papel que son sometidos a procesos de digitalización, o bien archivos PDF que deben ser creados desde aplicaciones como procesadores de texto, editores de imágenes, etc. Por esta razón, serán tratados en dos apartados diferentes.

Digitalización en SEDICI

En SEDICI se está utilizando la herramienta ABBYY FineReader 11 para generar un PDF con OCR.

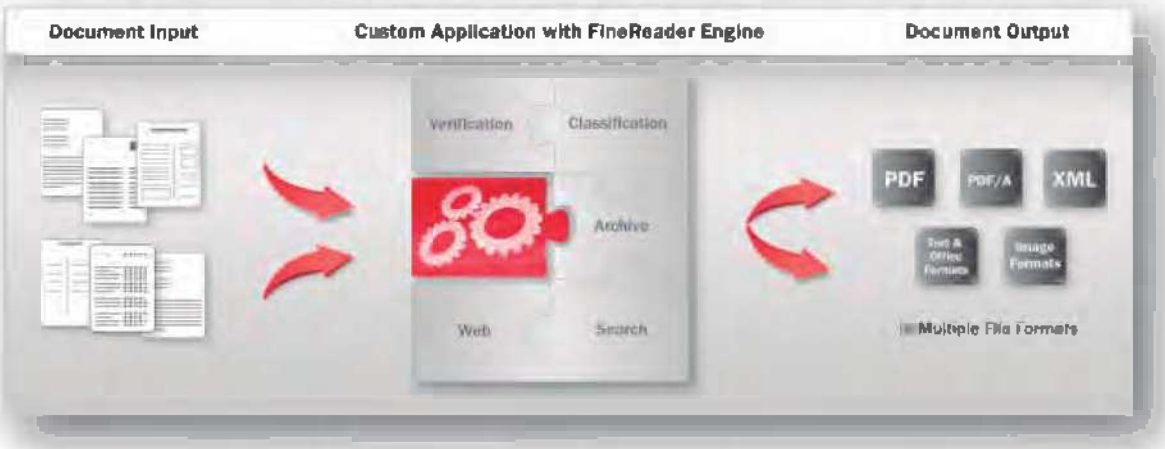


Figura 6.13: Captura de pantalla de los procesos y opciones de salida de ABBYY

Fuente propia.

La figura 6.13 muestra que las opciones de salida/guardado son, entre otras, PDF y PDF/A. Para los propósitos del repositorio, y especialmente de la preservación, se descartan las opciones de salida de Text & Office Formats, particularmente DOC y RTF, aunque no se descarta la opción de texto plano (TXT). La figura 6.14 particulariza las opciones de guardado de PDF/A.

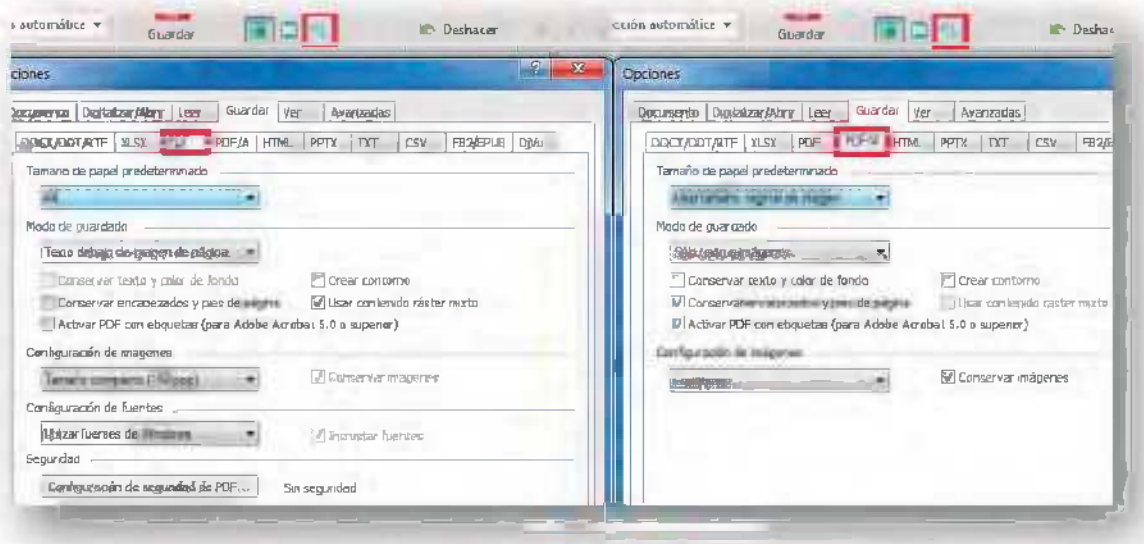


Figura 6.14: Captura de pantalla de las opciones de ABBYY para PDF/A

Fuente propia.

Al crear un PDF simple, la versión varía de PDF 1.4 a 1.6, pero el programa no explicita la versión hasta realizar el guardado. El programa tampoco explicita qué formato de PDF/A permite y esto es lo que se analizará a continuación, siguiendo los pasos del proceso de digitalización de documentos que se realiza en el repositorio.

En SEDICI, al momento de realizar la tarea de generación de un PDF con OCR, se selecciona la conversión a PDF/A (dado que es el formato adecuado para preservación). Luego, dentro del flujo de trabajo de la administración de SEDICI, es posible que los archivos PDF/A deban ser editados a través de Acrobat Writer, con el objetivo de realizar un control de calidad y mejoramiento de los documentos; por ejemplo, debido a la necesidad de revisión del texto extraído de manera automática por el proceso de OCR aplicado a los documentos digitalizados, o porque puede ser necesario incorporar a la obra una portada, imágenes, etc. Como se viera, los documentos PDF/A no pueden ser protegidos contra ediciones posteriores incluyendo medidas como la encriptación o las contraseñas. Al hacerlo, se contraviene la normativa PDF/A, ya que el contenido debe estar disponible en su totalidad, sin medidas de seguridad.

En cualquiera de estos casos descriptos, el resultado es que un archivo PDF/A que ya era compatible con el estándar, puede perder esa condición como resultado de cambios no intencionados o deliberados sin que sea evidente que ya no es compatible

con el estándar. Es decir, que un archivo puede estar rotulado como PDF/A y no tener sus características, y esto crea la necesidad de contar con alguna herramienta que valide que el documento sigue el estándar. Así, las herramientas como Comprobaciones de Adobe Acrobat y pdfaPilot están especialmente diseñadas para la validación de PDF/A y se puede descubrir de forma segura y fiable este tipo de problemas.

Caso 1: digitalización y OCR

Se realizaron tres conjuntos de pruebas denominados "prueba1", "prueba2" y "prueba3", en los cuales intervinieron ABBYY, Acrobat Writer y en algunos pasos pdfaPilot, que se muestran en la figura 6.15. Los archivos pueden verse también en la figura 6.16 que es una captura de pantalla del directorio llamado "Set de Lira".

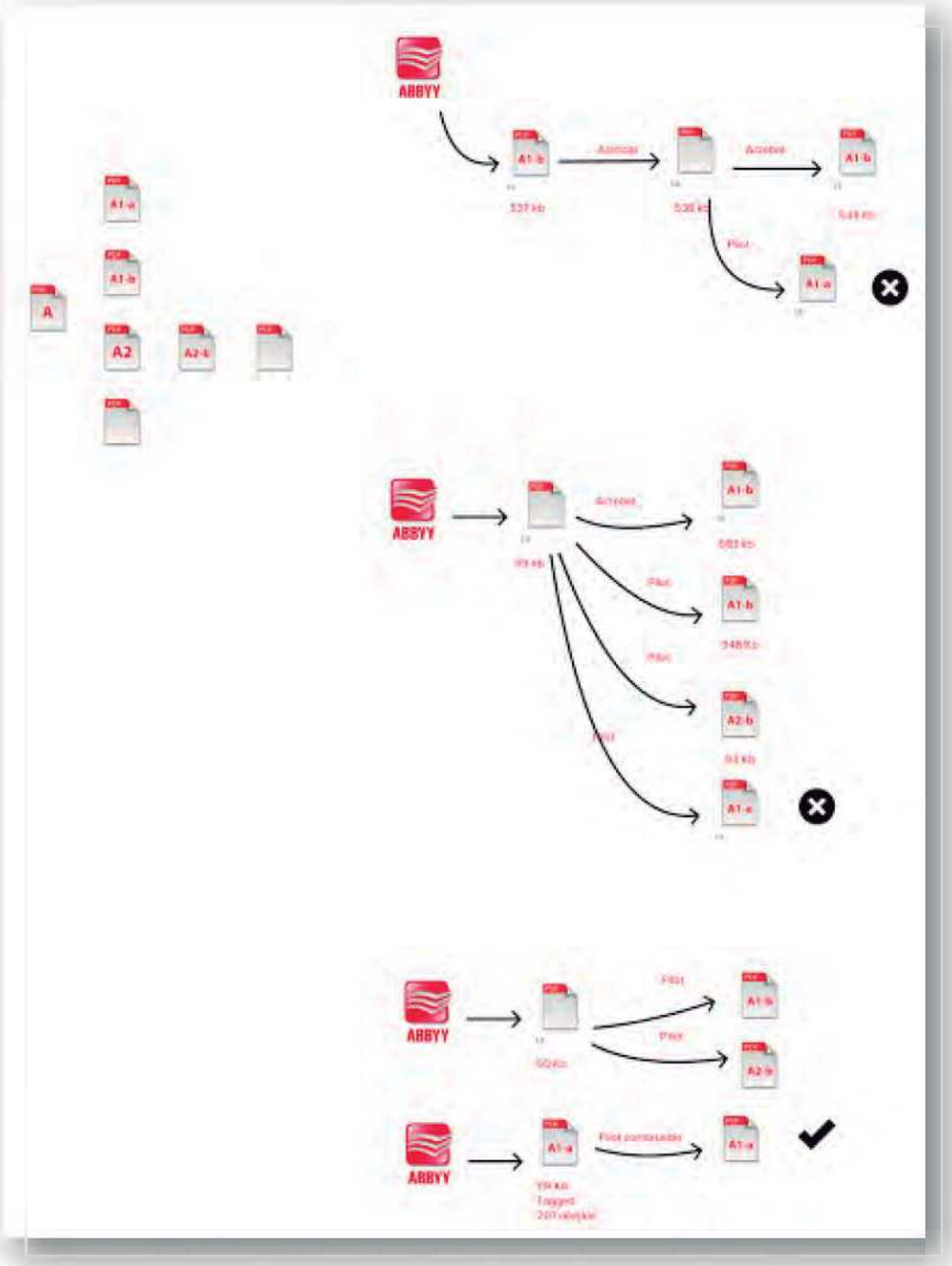


Figura 6.15: Pruebas realizadas sobre los archivos con distintas herramientas y formatos de archivos obtenidos

Fuente propia.

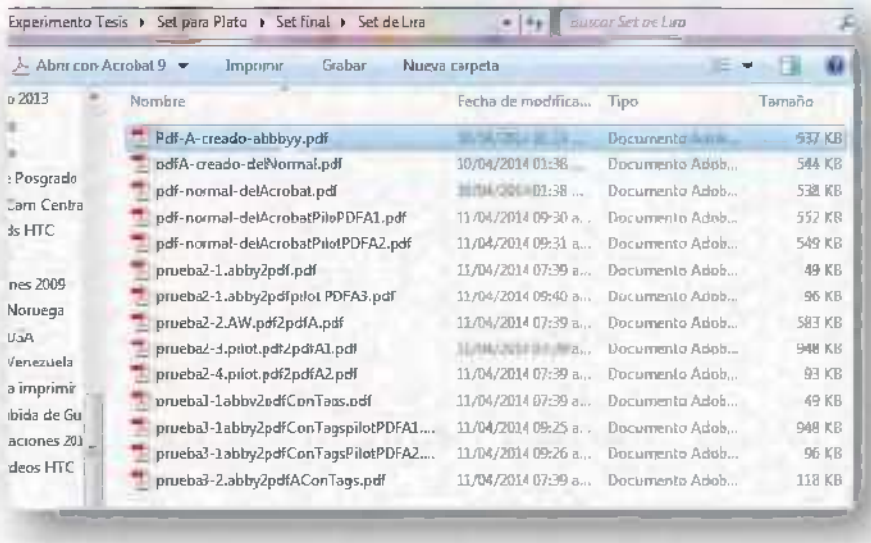


Figura 6.16: Archivos del directorio "Set de Lira" usados en las pruebas de la figura 6.15

Fuente propia.

La figura 6.15 muestra 9 archivos obtenidos tras el proceso de digitalización y luego trabajados con Acrobat Writer, a los que se suman los obtenidos por pdfaPilot en PDF/A1-b, A2-b y A3-b. Los primeros 5 archivos se corresponden con la salida PDF/A1-b de ABBYY, PDF de Acrobat Writer 1.6 (que, en el proceso de edición, pierde el formato PDF/A y vuelve a PDF), PDF/A generado a partir de Acrobat Writer y dos versiones A1-b y A2-b generadas desde pdfaPilot, todos con tamaños similares. Este conjunto de archivos constituyen la "prueba 1".

El primer paso de esta prueba fue la generación de un archivo PDF/A en ABBYY:

Productor de PDF:	ABBYY FineReader 11		
Versión PDF:	1.4 (Acrobat 5.x)		
Ubicación:	C:\Users\Administrador\Desktop\Prueba 1\		
Tamaño de archivo:	536,38 KB (549.254 bytes)		
Tamaño de página:	220,3 x 307,0 mm	Número de páginas:	3
PDF etiquetado:	No	Vista rápida en Web:	No

Este PDF/A-1b se genera en versión 1.4 y con un tamaño de 537 KB. Tras las operaciones de edición con Acrobat, la versión asciende a 1.6 pero se transforma en un

PDF normal (se pierde el pasaje a PDF/A) de aproximadamente el mismo peso (538 KB). Actualmente, en SEDICI, como se comentaba en párrafos precedentes, al finalizar la edición no se vuelve a convertir el archivo a PDF/A; en caso de hacerlo, el archivo volverá a la versión 1.4 y su tamaño se incrementará ligeramente a 544 KB (se toma como referencia el formato PDF/A1-b) pero la conversión desde el PDF normal de Acrobat a PDF/A1-a en pdfaPilot no puede realizarse. Al hacer clic sobre el botón "Convertir en PDF/A-1a" se inicia el proceso de conversión y la herramienta encuentra problemas que impiden la conversión a PDF/A-1a, e informa al usuario: "Arreglar problemas en estructura de etiquetado de PDF".

La salida de JHOVE para el PDF/A1-b generado desde ABBYY de la figura 6.17 muestra 54 objetos, en su mayoría bloques de texto e imágenes, en tanto que en la generada a partir de Acrobat (no se expone la vista) tan sólo 51 objetos y la de Pilot (no se expone la vista) 56 objetos. Esto deja en claro que las diferentes herramientas muestran discrepancias en la cantidad de objetos y la pregunta es entonces qué se pierde (o se gana) con cada conversión y qué conviene hacer en cada caso. El intento de generar un PDF/A1-a con pdfaPilot nuevamente no funciona. La generación de PDF/A1-a desde Acrobat no supera las comprobaciones como puede apreciarse luego en la figura 6.18.

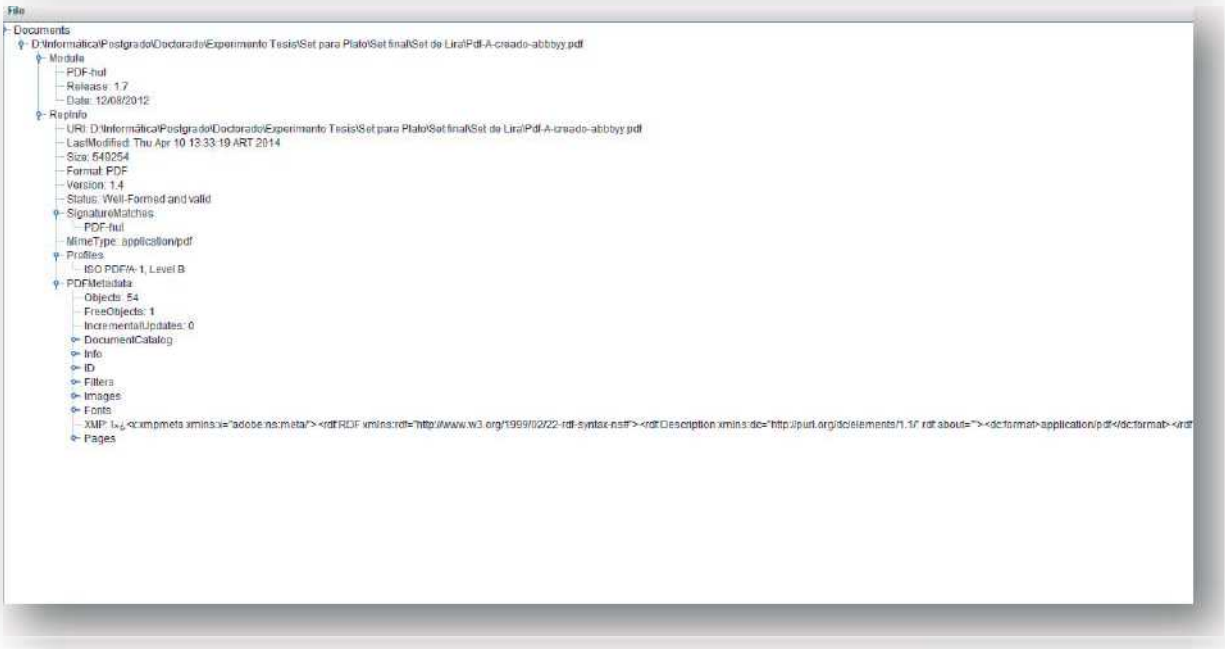


Figura 6.17: Captura de pantalla de la detección de formatos de JHOVE

Fuente propia.

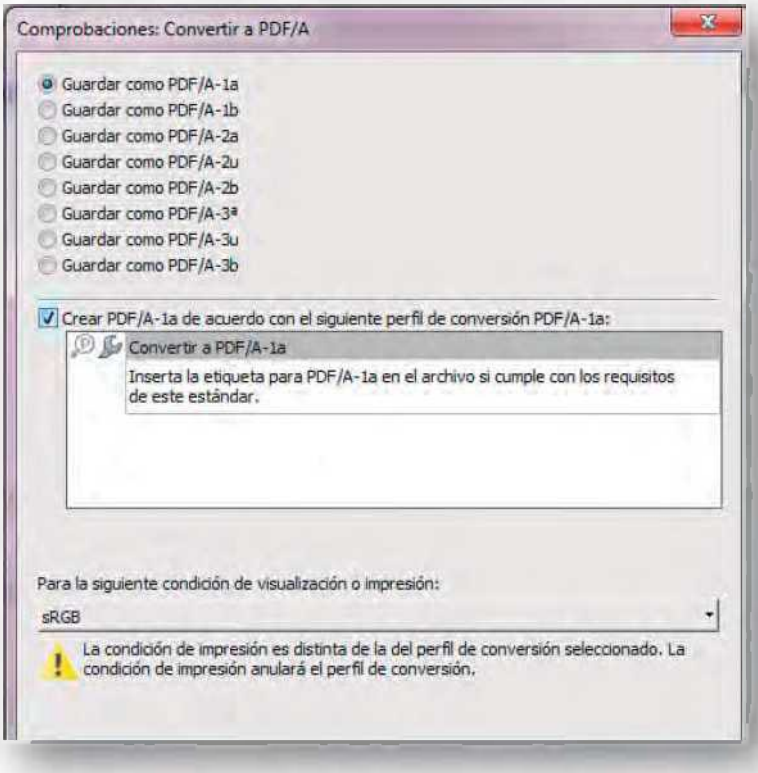


Figura 6.18. Captura de pantalla del intento de conversión a PDF/A1-a y detección de problemas con la herramienta de Comprobaciones de Acrobat

Fuente propia.

En el caso de la prueba 2, el primer paso fue la generación de un PDF desde ABBYY:

Productor de PDF:	ABBYY FineReader 11		
Versión PDF:	1.5 (Acrobat 6.x)		
Ubicación:	C:\Users\Administrador\Desktop\Prueba 2\		
Tamaño de archivo:	48,75 KB (49.920 bytes)		
Tamaño de página:	210 x 297 mm	Número de páginas:	3
PDF etiquetado:	No	Vista rápida en Web:	No

En este caso, el PDF generado es versión 1.5 y pesa aproximadamente 49 KB; al pasarlo a Acrobat y generar el PDF/A1-b a partir de la versión 1.5 de ABBYY, el tamaño se incrementa a 583 KB. Si desde el PDF generado por ABBYY se ingresa a pdfaPilot y se convierte a PDF/A1-b el archivo resultante pesa 948 KB; si, en cambio, la conversión elegida en pdfaPilot es PDF/A1-b, el archivo pesa 93 KB. Si se observa lo que sucede

con JHOVE, el primer PDF tiene 36 objetos, el PDF/A de Acrobat 55 objetos, el generado por Pilot en versión A1-b tiene 50 objetos, y en versión A2-b JHOVE no puede reportar la cantidad de objetos. Si desde Acrobat se observa el archivo con la herramienta Comprobaciones, la verificación resulta errónea como se aprecia en la figura 6.19 (puede acudirse a esta vista desde el propio pdfaPilot también).

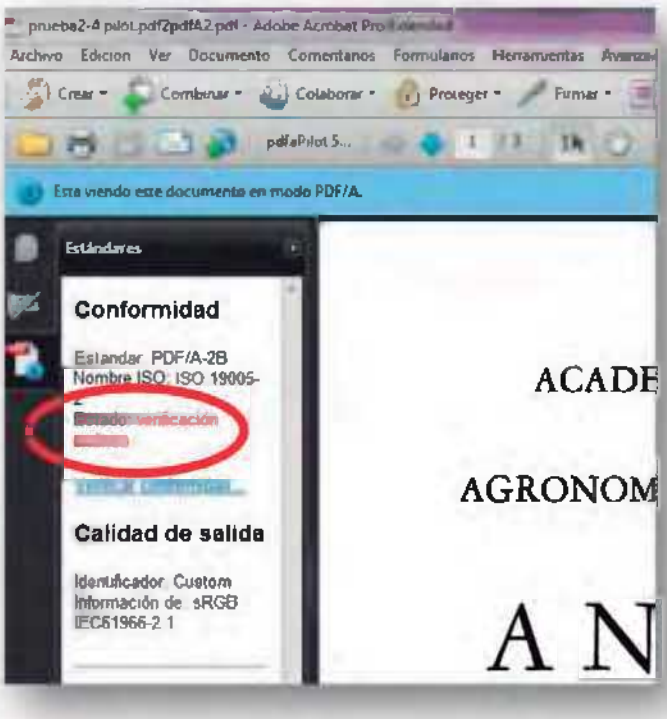


Figura 6.19: Captura de pantalla de la conversión a PDF/A2-b sin conformidad con el estándar

Fuente propia.

Si desde pdfaPilot se quiere crear, a partir del PDF normal de ABBYY, un PDF/A1-a no se logra porque falla la estructura; esto es, el PDF original no fue etiquetado y, por lo tanto, pdfaPilot no encuentra la organización de las partes del PDF, de acuerdo a lo que solicita el estándar de accesibilidad PDF/A1a.

PDF etiquetados

La prueba 3 puso de manifiesto un problema muy importante que está relacionado con los PDF etiquetados. Para que el PDF sea accesible debe ser un PDF etiquetado, lo que significa que incluye el contenido, la estructura y el orden de lectura para que los

lectores de pantalla puedan interpretarlo correctamente. Además, tal procedimiento facilitará también su exportación a otros formatos.

En la prueba 3 (que se corresponde a los archivos en la lista del directorio de la figura 5.20 indicados como "prueba3-x"), se parte de un PDF normal etiquetado con ABBYY cuyo peso es de 49 KB. Al analizarlo con JHOVE, muestra que cuenta con 36 objetos. La conversión con ABBYY a PDF/A etiquetado, eleva el número de objetos a 207 mientras que el archivo pesa 118 KB. Desde el archivo PDF normal de ABBYY se pasó con pdfaPilot a un archivo A1-b que pesa 948 KB con 50 objetos y a un A2-b de 96 KB y 50 objetos, pero lo interesante es comprobar que este archivo sí pasa la prueba de transformación a PDF/A1-a con pdfaPilot. Esto significa que ABBYY ha generado un archivo con la estructura adecuada y, por lo tanto, con condiciones de accesibilidad como las requeridas por el estándar a. El archivo tiene 207 objetos y pesa 118 KB, es decir que es una excelente opción. Cabe mencionar, por otro lado, que este es un caso de falla de JHOVE, que no puede reconocerlo como un PDF/A1-a. Si el archivo se abre con Acrobat, puede verificarse que se advierte al usuario que se trata de un archivo de sólo lectura, para evitar que éste altere lo que se ha logrado en cuanto a estructura, como puede verse en la figura 6.20:

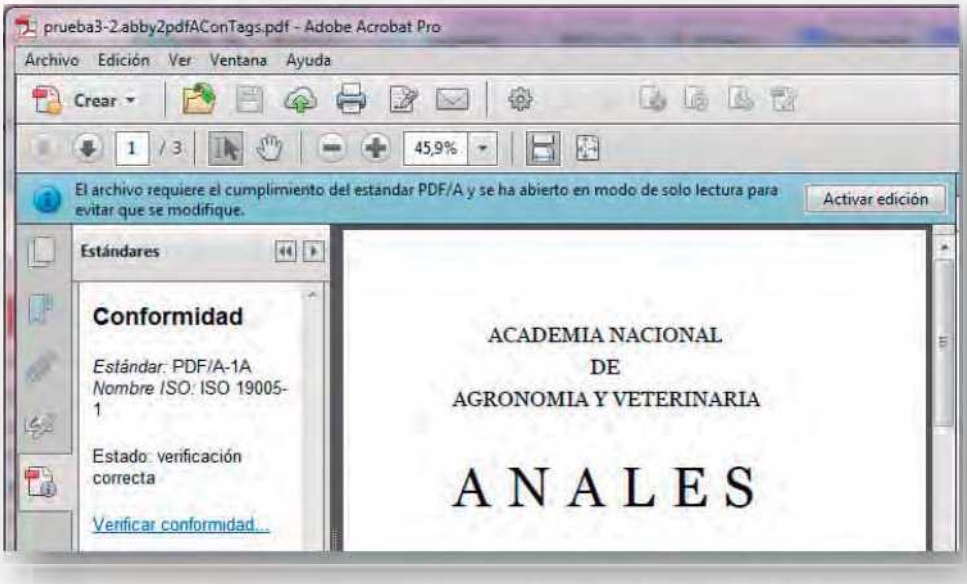


Figura 6.20: Captura de pantalla para verificar conformidad con la herramienta Comprobaciones de Acrobat

Fuente propia.

También se observa que supera la verificación del estándar PDF/A1-a. Sin embargo, queda un interrogante pendiente: ¿qué sucede si, además de optimizar el archivo, es necesario corregir errores tras el proceso de OCR cuando se lo procesa desde Acrobat? En este caso, no sería posible mantener el PDF/A1-a obtenido desde ABBYY (y que ya se comprobó que supera todas las pruebas). Pero sí es posible editar el archivo sin perder el etiquetado realizado por ABBYY y, tras esa operación, al realizar la comprobación nuevamente con Acrobat, se obtiene un PDF/A1-a etiquetado y optimizado. Para realizar la operación basta con elegir la opción de edición como se muestra en la figura 6.21.

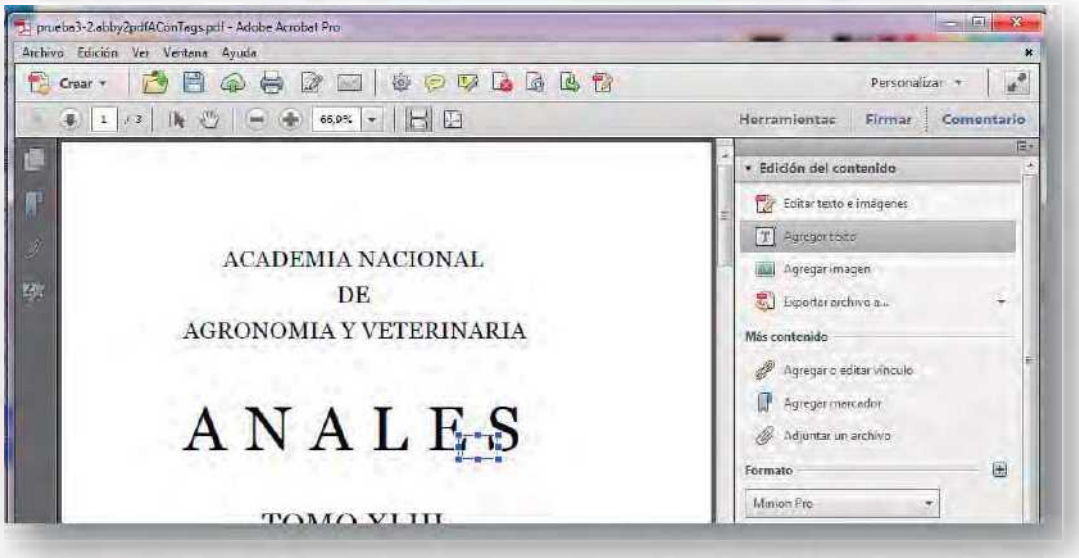


Figura 6.21: Captura de pantalla que muestra la selección de la opción de edición (a la derecha de la imagen)

Fuente propia.

Para completar el proceso, tras realizar los cambios, hay que volver a repetir las comprobaciones. Esto optimizará nuevamente el PDF/A1-a que, como se ve en la figura 6.22 muestra que el PDF está efectivamente optimizado.

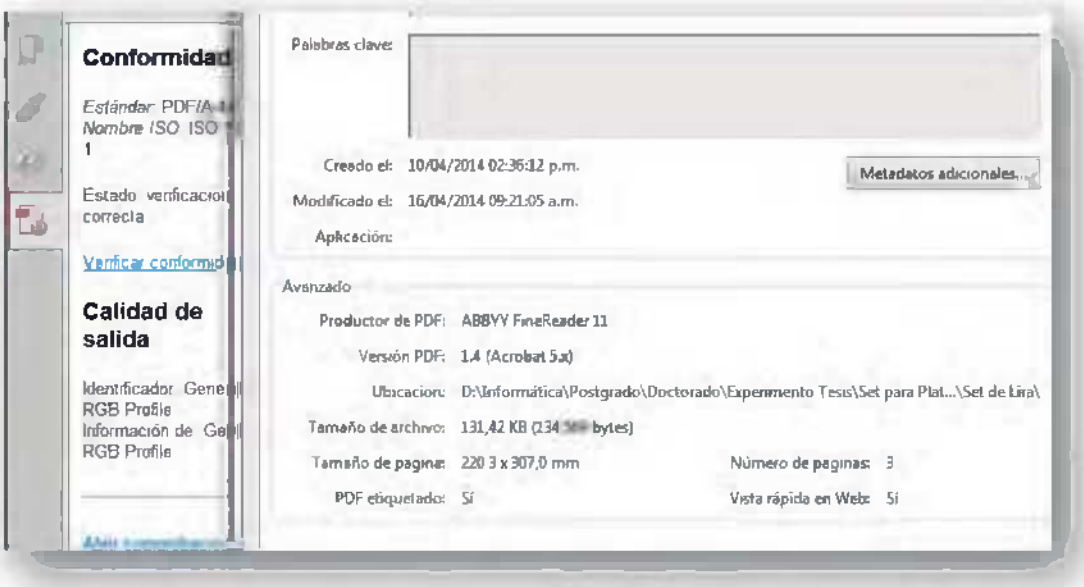


Figura 6.22: Captura de pantalla de la verificación del PDF

Fuente propia.

Todos los archivos de las tres pruebas fueron procesados por DROID. Se verificó que el archivo generado con ABBYY etiquetado y luego optimizado, efectivamente cumple con el estándar PDFA1-a, que en el registro PRONOM se corresponde con el formato fmt/95 como se muestra en la figura 6.23.

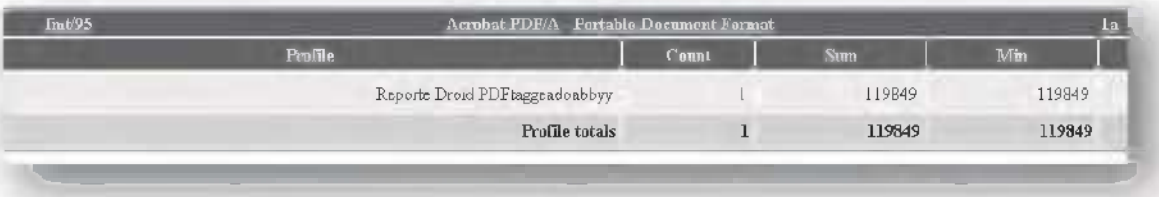


Figura 6.23: DROID reporta que fmt/95 se corresponde con PDFA1-a

Fuente propia.

Recomendaciones

Cubiertos los casos de digitalización que se presentan en SEDICI y vistos los resultados obtenidos, surge como recomendación principal, a la hora de la adecuada preservación de los OD, generar un PDF etiquetado desde ABBYY y no realizar modificaciones desde la administración; de tener que realizarlas, hay que utilizar luego

una herramienta de validación que asegure el regreso a PDF/A. La opción de optimizar el PDF e incluso editarlo con Acrobat, aunque pasa las verificaciones de Acrobat y de pdfaPilot, genera algunas dudas porque surgen incompatibilidades entre las versiones de Acrobat para reconocer ese PDF. Por ejemplo, un mismo archivo sin errores en Adobe Acrobat 11 Pro, visto desde Acrobat 9, reporta incumplimiento en el estándar. Vale aclarar que los que resultan validados en la versión 9, también lo son en la 11. En el caso de los libros digitalizados, la opción de optimización es sumamente importante, ya que incide en numerosos aspectos (correcta recuperación del texto en búsquedas a texto completo, óptima visualización en web, por ejemplo), por lo que debe encontrarse la mejor manera de cumplimentar ambos objetivos: que el archivo cumpla con el estándar y que quede correctamente optimizado.

Caso 2: Materiales nacidos digitales

En los procesos de ingesta de SEDICI puede haber materiales en distintos formatos, como PDF, DOC, JPEG, PPT, MP3, XML y hasta archivos ejecutables (aunque esto último es un caso bastante raro, en general). Así, si lo que se tiene es un PDF, en versiones de 1.0 a 1.7 inclusive, deberá ser transformado a PDF/A, para lograr etiquetarlo y optimizarlo, según se viera en el caso anterior. Sin embargo, tales tareas no son simples y a continuación se muestran algunas pruebas de ello.

En la figura 6.24 se puede observar un PDF procesado con Acrobat 11, que pertenece a un artículo de la revista *Auster* (es decir, un material que no fue digitalizado en SEDICI sino que llegó a SEDICI así). Como se puede apreciar, el PDF no está en ninguna versión de PDF/A. Para lograrlo se intenta guardar una copia, eligiendo PDF/A y específicamente PDF/A1-a.

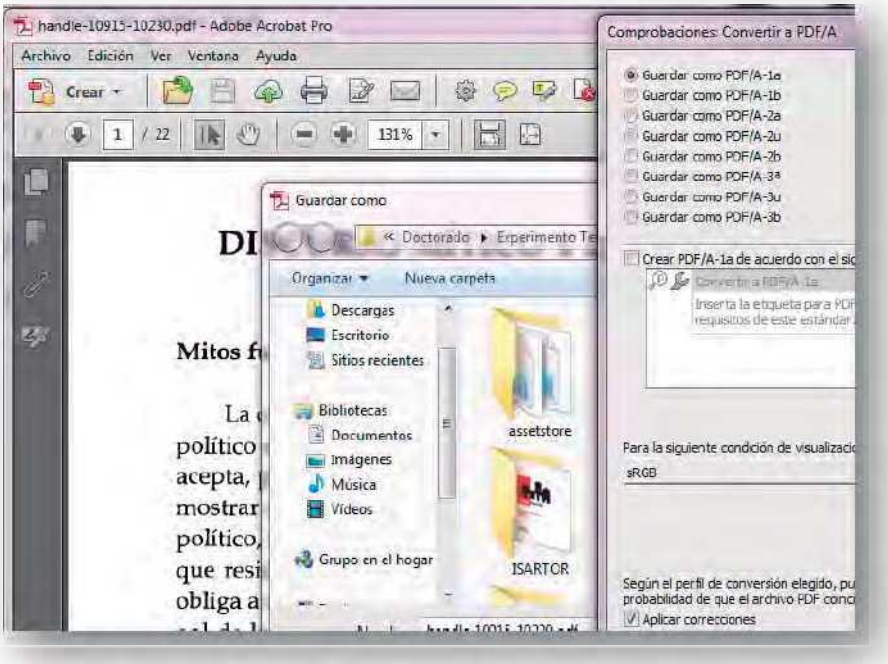


Figura 6.24: Captura de pantalla de selección de formato de archivo PDF/A-1a

Fuente propia.

Lamentablemente el intento de conversión no funciona, como lo muestra a continuación la figura 6.25.

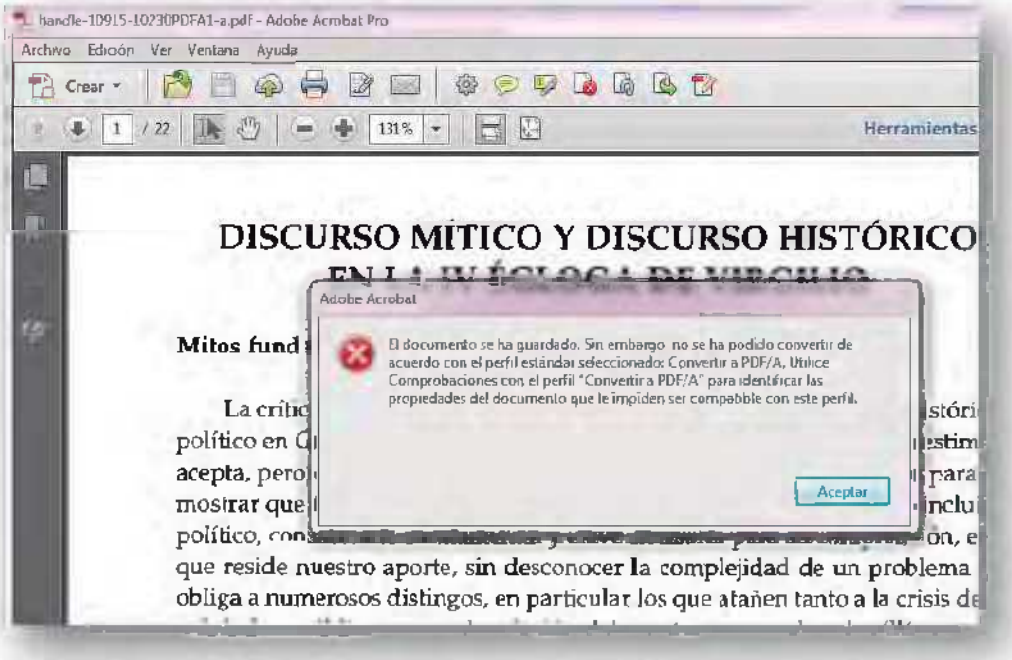


Figura 6.25: Captura de pantalla que reporta que el archivo no puede convertirse a PDF/A1-a

Fuente propia.

Ante una incidencia así, se debe abrir el panel de herramientas y elegir la opción "comprobar accesibilidad", dado que la accesibilidad es la característica primordial de un PDF/Ax-a. El reporte avisa que no supera la comprobación y lista los errores, algunos de los cuales se muestran desplegados, como puede observarse en la figura 6.26.

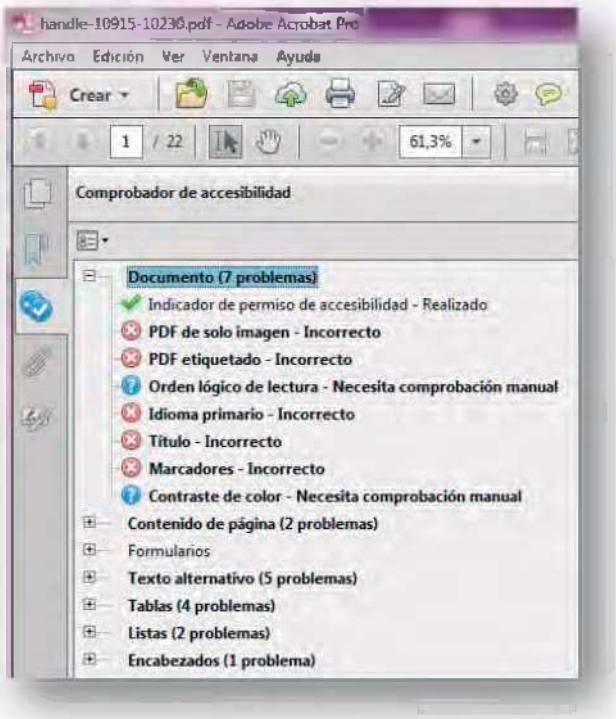


Figura 6.26: Lista de errores durante la validación del archivo a PDF/A1-a

Fuente propia.

Del listado de problemas que ofrece la figura 5.30, el problema del PDF etiquetado de manera incorrecta constituye una dificultad importante. Ni un PDF accesible ni un PDF/A-1a pueden habilitar una comprobación de PDF etiquetado para asegurar que las estructuras del documento son significativas (correctas). Ambos tipos de comprobaciones sólo determinan si la información estructural existe en las especificaciones del archivo PDF y no si cualquier información estructural tiene sentido. Por esta razón, el estándar estipula que la información estructural no puede agregarse más tarde. Debe ser importada durante la creación del PDF o añadida manualmente después. La creación automática de estructuras puede ser posible sin causar problemas para archivos PDF muy sencillos. Sin embargo, si un usuario utiliza

un proceso automatizado para reconstruir una estructura, debe asegurarse de que el proceso está validado.

Para un administrador de repositorio esta tarea insume un tiempo excesivo e incluso no es posible asegurar la validez de la estructura generada. Una posibilidad superadora de esta situación, aunque de trabajo a más largo plazo, es la de generar un manual orientativo o tutorial para los autores, de modo que generen archivos PDF con condiciones mínimas de calidad; este sería el proceder a futuro en relación a los documentos autoarchivados. Es del todo claro que, con más del 90% del repositorio en formato PDF en diferentes versiones, una tarea manual de adecuación al estándar que incluyera el etiquetado, sería imposible.

Finalmente, el mismo PDF del ejemplo se convierte con éxito a la versión PDF/A1-b, lo que resulta el procedimiento más adecuado para llevar adelante en esta primera etapa de las tareas de preservación.

Como se advirtió al comienzo del estudio del formato PDF, la compatibilidad hacia versiones previas alcanza a la versión 1.3. Sin embargo, SEDICI tiene archivos en las versiones 1.0, 1.1, 1.2 y 1.3 inclusive. Para estos casos se tomó una muestra pequeña de archivos en PDF, variando en su versión (1.0 a 1.3 inclusive), peso y año de creación, y para todos los casos se probó la posibilidad de generar PDF/A desde ABBYY. Las pruebas resultaron exitosas: ABBYY abre el archivo por imágenes (cada imagen es una página) y lo va analizando mientras lo carga (marca texto/imágenes). Sobre cada una de ellas, realiza un reconocimiento de caracteres e imágenes. Una vez terminada la carga y el reconocimiento automático del programa, se hace una revisión/modificación manual al archivo (puesto que el reconocimiento de caracteres, en algunos PDF, puede ser defectuoso) y se guarda como PDF/A etiquetado.

En la mayoría de los casos de estas pruebas, los archivos han pasado la verificación de Acrobat, por lo que se puede afirmar que es posible generar un PDF/A1-a con éxito y optimizado (al pasar la verificación de Acrobat, queda linealizado, es decir, optimizado). En los restantes casos se revisan y corrigen los problemas listados con Acrobat y se guarda el nuevo documento, en formato PDF/A-1a. Así, todos los archivos de esta muestra fueron guardados en formato PDF/A1-a etiquetado y optimizado. Esto representa una solución al problema de los formatos PDF de versiones no compatibles. El detalle total de ítems en versiones 1.0 hasta 1.3 inclusive se encuentran en el anexo

titulado "Archivos PDFs versiones viejas.xlsx".

Identificados los problemas y sus posibles soluciones, se generaron las tareas en el sistema de gestión de incidencias para iniciar la corrección de este problema, como se puede ver en la figura 6.27:

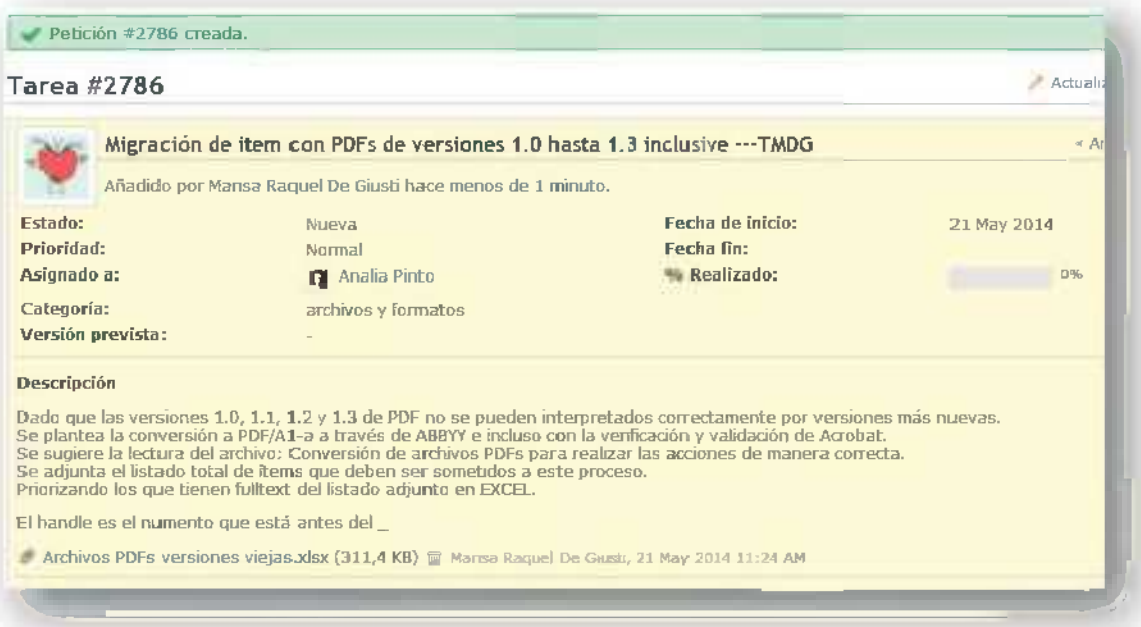


Figura 6.27: Tarea para la migración de PDF de versiones incompatibles

Fuente propia.

¿Qué hacer si el PDF debe obtenerse a partir de un documento de texto?

En este caso deben tenerse en cuenta otros parámetros, además de los ya vistos. Aquí se hace referencia a documentos con extensiones DOC, DOCX, ODT, RTF, entre otros. El primer punto a tener en cuenta debe ser la conveniencia de trabajar primero la accesibilidad desde el editor del documento (Microsoft Word, OpenOffice/LibreOffice Writer, etc.).

En este sentido, hay un documento muy interesante para seguir este tópico, producido por la Universidad Nacional de Educación a Distancia (UNED) en Madrid en 2012, denominado *Guía de accesibilidad de documentos electrónicos* (Sama Rojo, Sevillano Asensio, 2012), disponible en la web, en el que se aborda, de manera extensa, los postulados que hacen accesible un documento de texto en sus diferentes versiones. El concepto de accesibilidad que abarcan los autores excede las propuestas de este

trabajo debido a que apuntan a dar accesibilidad a personas con capacidades diferentes y, entre otros puntos, remarcan la necesidad de crear un documento cuya estructura se genere de manera automática, no con títulos o subtítulos destacados manualmente sino con las herramientas que ofrece el propio editor de texto. Esto se debe a que los productos de apoyo (un lector de pantalla, dispositivos braille) que utilizan las personas con discapacidades visuales requieren que el documento incorpore una serie de marcas que indiquen los elementos utilizados en la estructura. Más allá de esto, la guía ofrece pautas generales a seguir: elección de fuentes, espaciado, determinación clara de idioma principal, titulación de imágenes, tablas, generación de una tabla de contenidos y otras.

La ausencia de documentos de texto originales en SEDICI, en este momento, no amerita dar una mayor extensión a este apartado. Dicha ausencia se explicita en las políticas adoptadas por el repositorio hace ya algún tiempo, de no permitir la subida de archivos en formatos poco confiables para la preservación. Sin embargo y, a pesar de que algunos formatos como DOC y DOCX[24] no son recomendables como opciones de preservación de los archivos del repositorio, por ser parte de formatos propietarios, también es cierto que brindan posibilidades de transformación a PDF/A de manera sencilla. A diferencia de las versiones anteriores de Microsoft Office, en Office 2007 se permite la exportación de archivos a PDF/A sin el uso de Acrobat. Para realizar la conversión, el modo más sencillo se plantea a través de la opción "guardar como", donde se elige PDF y en el apartado de opciones del PDF se selecciona la casilla "Compatible con ISO 19005-1 (PDF/A)", tal cual se muestra en la figura 6.28.

[24] Por ejemplo, en la documentación de Microsoft, el formato del .doc se llama MS-DOC y el .docx se llama Office Open XML y tiene su especificación abierta.

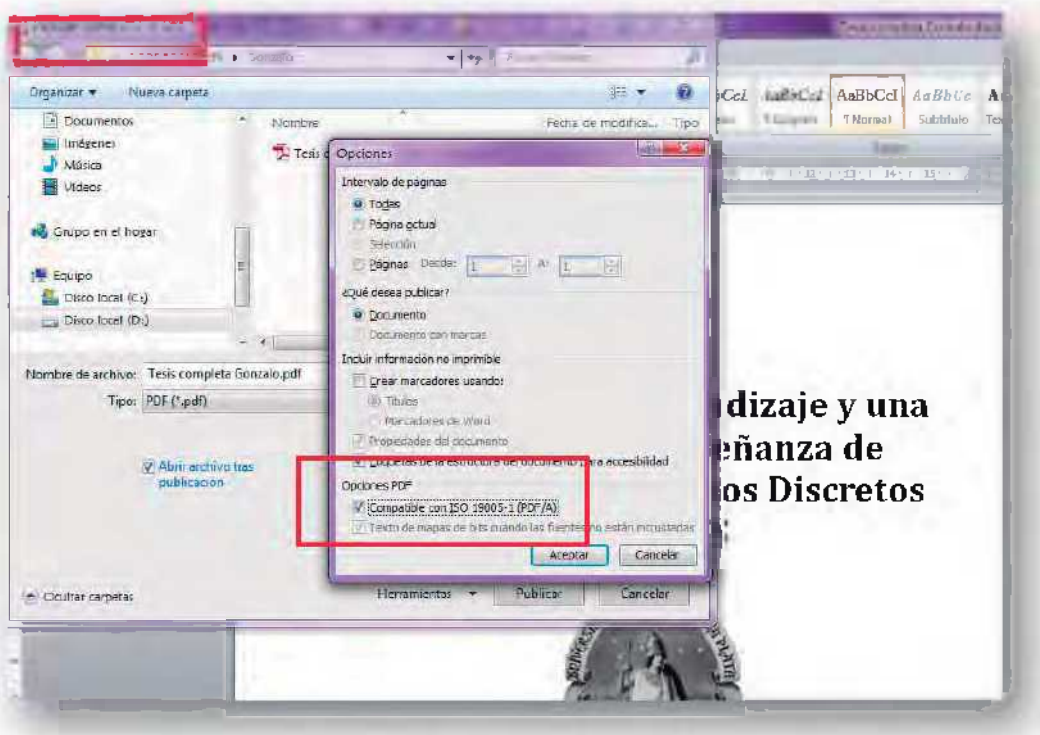


Figura 6.28: Opciones de guardado de PDF en Word sin utilizar Acrobat

Fuente propia.

Tras generar el PDF/A1-a es posible comprobar que cumple el estándar; vale destacar que siempre se tiene presente la necesidad de exponer en un formato abierto, como el de Open Document Text. Se estima que exponer en el repositorio los archivos en formatos abiertos puede crear conciencia en los usuarios acerca de las prácticas vinculadas al acceso abierto. OpenOffice y LibreOffice ofrecen también la posibilidad de guardar un archivo como PDF/A, como puede verse en la figura 6.29.

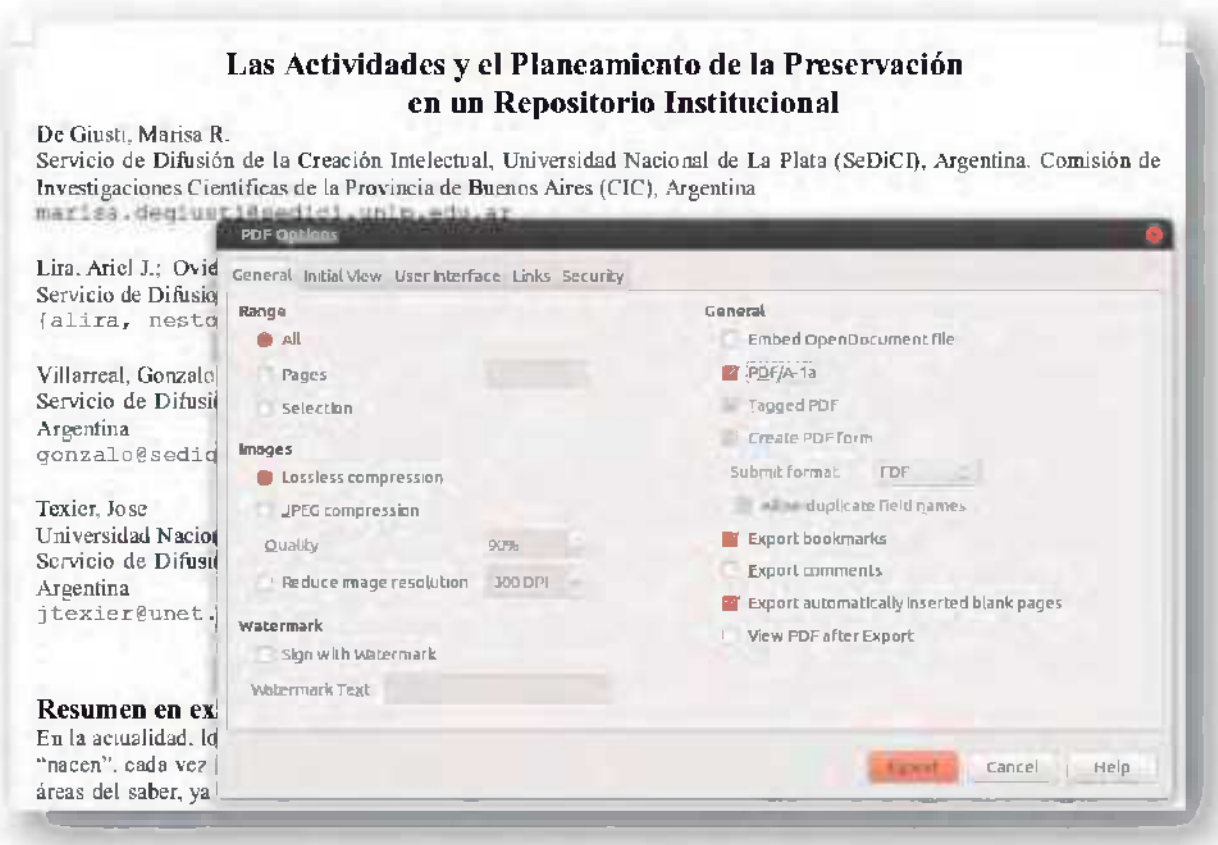


Figura 6.29: Conversión a PDF/A desde LibreOffice Writer

Fuente propia.

¿Qué hacer si el PDF debe obtenerse a partir de una presentación (MS PowerPoint, LibreOffice Impress, etc.)?

PowerPoint, como parte de Office 2007 también permite la exportación a PDF/A1-b. El proceso a realizar es idéntico al aplicado para documentos de texto, y resulta en un PDF/A sin optimizar. Al momento de este análisis, en SEDICI había sólo 5 PPT versión 1997-2002 que se corresponden con el formato fmt/126 del registro PRONOM. Se aconseja en este caso también exponer en un formato libre como ODP (Open Document Presentation), pues de este modo el usuario podrá acceder a los dos formatos posibles para visualizar la presentación, con la indicación expresa de que ODP es un estándar abierto; esta es una propuesta de trabajo a largo plazo y busca crear conciencia al respecto, como se dijera en el caso anterior.

Cabe resaltar que, cuando se trate de texto, siempre se va a guardar PDF-A: lo deseable es siempre PDF/A1-a (PDF/A1-b si no se puede -a), pero con vistas ya a

guardar PDF/A2-a.

Otros formatos en SEDICI

Como se desprende de la figura 5.15, tan sólo un 5% de los ítems de SEDICI se encuentra en un formato distinto a PDF. Con esos formatos se han hecho algunas pruebas mínimas para ver cómo pueden realizarse acciones de migración y con qué herramientas. Claramente, no forman parte de una primera etapa en el futuro plan de preservación del repositorio.

Del resto de los formatos existentes, un 3% (el segundo en importancia) de los archivos se encuentra en MPEG ½ Audio Layer 3 (MP3) y la alternativa es proponer como formato de preservación algún otro que sea abierto, como se verá a continuación.

En el caso de los formatos de imágenes, que tan sólo constituyen el 1% de los contenidos del repositorio, como en el caso precedente, se propone analizar alternativas de migración a estándares abiertos.

Formatos alternativos y herramientas de conversión propuestas

Documentos en formato SWF

SWF fue inicialmente la abreviación de *Shockwave Flash* y posteriormente de *Small Web Format* (formato web pequeño). Es un formato de archivo de gráficos vectoriales creado originalmente por la empresa Macromedia, actualmente Adobe Systems. Los archivos SWF pueden ser creados por el programa Adobe Flash, aunque hay otras aplicaciones que también lo permiten, entre ellas muchos softwares libres. Por lo general, estos archivos son ejecutados sobre el navegador mediante la extensión Adobe Flash Player, aunque también pueden ser encapsulados para ejecutarse de forma autónoma. Es un formato vectorial aunque también admite bitmaps, con posibilidades de animación y este es el caso que se presenta en SEDICI.

Los archivos SWF suelen ser lo suficientemente pequeños como para ser publicados en la web en forma de animaciones o *applets* con diversas funciones y grados de interactividad, ya que ésta es una de sus principales funciones y objetivos.

Con respecto a su licencia, la especificación completa del formato está disponible. Hasta el 1 de mayo de 2008 el formato no era totalmente abierto: reproducirlo no estaba permitido por la especificación de la licencia. En esa fecha, como parte de su Open Screen Project, Adobe eliminó tales restricciones sobre los formatos SWF y FLV.

El formato es bastante simple, pero es binario y por lo tanto no es de lectura accesible, como el SVG, que es un estándar abierto basado en XML, recomendado, por otra parte, por el W3C (World Wide Consortium) para imágenes fijas.

SVG: la alternativa estándar

Siempre que sea posible, lo óptimo para fomentar la accesibilidad de los contenidos es ofrecer tecnologías estándares. En el campo de los gráficos vectorizados, el W3C recomienda, como se dijera, la tecnología SVG6, que presenta —entre otras— las siguientes ventajas:

- Fácilmente editable (el código fuente es XML y CSS).
- Pueden hacerse búsquedas en el código del gráfico.
- Los textos del gráfico pueden presentarse en el idioma preferido del usuario, de manera sencilla.
- Puede reutilizarse una hoja de estilos CSS para varios gráficos.
- Es un estándar abierto con implementaciones distintas y extensibles.
- Se le pueden aplicar efectos típicos de las imágenes bitmap a imágenes vectoriales (rellenos degradados, efectos, etc.).
- Pueden generarse gráficos automáticamente, transformando el código XML.

Las limitaciones actuales de la tecnología SVG son su aún reducido soporte en algunos agentes de usuario (sigue siendo necesario el uso de extensiones en los navegadores) y su incapacidad para incluir directamente elementos multimedia como videos o sonido.

Como alternativa para dar mayor accesibilidad puede usarse Swiffy, un servicio provisto por Google, que convierte archivos SWF a HTML5 y también a SVG. Las pruebas realizadas con este conversor demuestran que convierte muy bien y cubre una buena gama de navegadores. Como una solución de este tipo (Swiffy) queda sujeta a que la empresa (Google) decida mantener el servicio, se estudiarán otras alternativas

abiertas como SWF2SVG.

SEDICI tiene actualmente 47 archivos en Flash en versión 5 (SWF5), todos ellos imágenes en movimiento muy pequeñas, las cuales podrían convertirse con Swiffy fácilmente y resultar de un tamaño similar al de flash original. La propuesta sería, entonces, para exponer recursos en SWF, exponer en la última versión de Flash (versión 13 y sucesivas) y en HTML5. En el bundle de preservación, se almacenaría el archivo original y el SVG.

Planillas de cálculo (MS EXCEL, OpenOffice/LibreOffice Calc, etc.)

En todo SEDICI sólo hay un archivo con extensión XLS, que se corresponde con el handle 26277, que expone una presentación en PDF y un archivo MS Excel con un ejercicio. En este caso, el formato abierto alternativo es el ODS (Open Document Sheet) al cual es posible exportar el archivo sin dificultad aunque con algún ligero aumento en el tamaño. El archivo original pesaba 78 KB, mientras que el ODS exportado desde Libre Office pesa 114 KB. Nuevamente, la propuesta es exponer en ambos formatos para así diseminar también los formatos abiertos. Excel, como parte de MS Office 2007 también permite la exportación a PDF/A1-b. Tanto en este formato como en cualquier otro existente en el repositorio, se recomienda siempre tener versiones actualizadas.

Archivos de audio en formato MP3

Los archivos en MP3 sirven de ejemplo para mostrar un formato de especificación abierta, pero cuyo uso legal requiere de una licencia. En SEDICI, como se indicara, hay más de 500 archivos con este formato. La propuesta es, como siempre, guardar también en un formato abierto, por ejemplo OGG. Se ha utilizado un conversor en línea a OGG: online-convert.

Como una solución de este tipo queda también sujeta a que la empresa (Google, igual que en el caso de Swiffy) decida mantener el servicio, se estudiarán otras alternativas abiertas como Gstreamer, que además tiene la ventaja de presentar algunas plataformas de interacción con los usuarios, sumamente cómodas, como SoundConverter.

Archivos de imágenes en formato JPEG

Respecto al JPEG, es un estándar que, si bien expone de manera abierta su especificación, respecto de su licencia de uso existen fuertes divergencias. En SEDICI se cuenta con 125 archivos en este formato en las versiones 1.0 (fmt/42), 1.0.1 (fmt/43) y 1.0.2 (fmt/44); la versión actual es la JPEG 1.0.2. Los archivos que se encuentren en versiones previas deberían migrarse a la versión 1.0.2 para mantener actualizados los archivos y a la vez también debería analizarse la migración a un estándar libre como Portable Network Graphics (PNG), que al presente es el formato abierto que soporta compresión sin pérdida de mayor uso en Internet.

Otros formatos en SEDICI

Además de los mencionados, existen otros formatos, que aparecen en muy pocos casos, por lo cual un análisis más detallado no se corresponde a los fines de este trabajo. Sin embargo, sí cabe mencionar algunas reglas generales, como la necesidad de tener siempre versiones actualizadas y especialmente de alternativas abiertas. Por ejemplo, en SEDICI hay sólo 1 archivo en formato Adobe Illustrator, el cual debería contar con alguna alternativa abierta como INKScape, por citar una, que también es aplicable para el caso de CorelDraw.

Propuesta de trabajo a futuro y tareas más inmediatas

La propuesta es analizar el formato de cada uno de los archivos que forman el paquete de información (SIP) del proceso de ingesta de manera que se ajusten a los requerimientos de preservación de SEDICI. Según el tipo de archivo recibido, el sistema debería ser capaz de generar nuevos formatos derivados, mejor preparados para la preservación a largo plazo, como PDF/A.

A futuro, la propuesta debiera aproximarse lo más posible a lo que se hace en desarrollos específicamente implementados para dar cuenta de los desafíos de la

preservación digital. Por ejemplo, DAITSS[25] implementa diferentes estrategias de preservación basadas en la transformación de los archivos con el fin de normalizarlos, y en la migración hacia nuevos formatos en caso de ser necesario. En el primero de los casos, los archivos propietarios se convierten a formatos abiertos y normalizados (por ejemplo, una hoja de cálculo propietaria pasa a otro formato abierto basado en XML); en el segundo de los casos, se procura migrar los archivos a formatos más actuales (por ejemplo de una versión 2007 a 2014).

Las estrategias basadas en la transformación deberían aplicarse básicamente a archivos de imagen, texto, audio y video, y esto es lo que se ha tratado de mostrar en este apartado con los ejemplos y pruebas precedentes.

El resto de archivos, como pueden ser los ejecutables, programas, etc., son candidatos a otro tipo de estrategias como las de emulación; en este caso, puede asegurarse la preservación de este tipo de archivos, pero no necesariamente su funcionamiento en futuros entornos si los mismos no son emulables. Estos casos no constituyen por el momento un objetivo en SEDICI, puesto que aún no se cuenta con este tipo de contenidos.

El sistema debería procurar almacenar y preservar por lo menos tres versiones de cada uno de los archivos ingresados al repositorio: la versión original tal y como ha sido subida, un nuevo formato normalizado y una posible migración a nuevas versiones, o a formatos abiertos (si los precedentes no lo fueran).

Durante mucho tiempo, las organizaciones responsables de conservar grandes cantidades de información utilizaron el formato TIFF (Tagged Image File Format) para la preservación. Este formato convierte a píxeles texto e imágenes. Por un lado, por guardar píxeles, se tiene como ventaja que no hay pérdida de gráficos, ni problemas con las fuentes, y siempre se mantiene la apariencia del original. Sin embargo, los textos pasan a imágenes y por lo tanto no pueden realizarse búsquedas (excepto realizando OCR), y los tamaños de archivo son relativamente grandes. TIFF ha sido un estándar de hecho, pero no lo es a partir de ninguna norma o recomendación, con lo cual, a pesar de sus ventajas, no se propone la inclusión de un archivo extra en TIFF

[25] DAITSS es una aplicación de software para la preservación digital desarrollado por el Florida Center for Library Automation (FCLA). Es utilizada por el Archivo Digital de Florida (FDA), un servicio de repositorio de preservación a largo plazo proporcionado por el Campus Virtual de Florida para el uso de las bibliotecas de las once universidades que reciben fondos públicos en Florida.

para archivos de origen con formato textual o mixto.

Qué preservar y qué mostrar

En el repositorio SEDICI-DSpace aún no se cuenta, dentro del registro del ítem, con un bundle específico dedicado a la preservación (el bundle de preservación), pero el objetivo es que este bundle exista y que en el mismo se almacene el archivo original, las migraciones que van ocurriendo durante el ciclo de vida del objeto y un formato abierto, como se dijera anteriormente. Este bundle de preservación indicaría qué es lo que habría que preservar del objeto digital.

Por otra parte, en cuanto a qué exponer a los usuarios, esto puede variar de acuerdo al tipo de archivos de que se trate; sin embargo, no hay dudas de que cuando se trate de PDF, se mostrará PDF/A, y cuando sea posible PDF/A1-a y si esto no puede lograrse, PDF/A1-b.

Existen, sin embargo, contenidos como los guardados en hojas de cálculo que pueden muy bien guardarse en el futuro bundle de preservación, por ejemplo en PDF/A, pero que serían expuestos a los usuarios en su formato original para que no pierdan características que dificulten su comprensión. Sería muy deseable exponer también en un formato estándar abierto, pero existen restricciones en cuanto a la exposición de contenidos duplicados, por lo cual esta opción deberá analizarse de manera cuidadosa para que el repositorio no sufra penalizaciones.

Gran parte del AIP está formado por metadatos. DAITSS, por ejemplo, implementa otros estándares de metadatos específicos entre los que encontramos MIX para imágenes digitales, AES para audio, textMD para texto y docMD para documentos; tales estándares podrían ser analizados para ver si pueden auxiliar en el proceso de preservación.

Otra recomendación adicional es revisar el vínculo entre un ítem y sus versiones según se ha planteado aquí, de modo de mantener la trazabilidad a lo largo del tiempo y prever las acciones de migración apropiadas para mantener la accesibilidad para todos los objetos del repositorio.

¿Qué hacer con lo que ya existe en el repositorio?

El único formato sobre el cual es necesario realizar acciones masivas es PDF; de

hecho, es muy importante, como ya se ha indicado, pasar a versiones más actuales de PDF, por lo menos los archivos que se encuentran en las versiones 1.0, 1.1, 1.2 y 1.3 debido a que no pueden ser interpretados correctamente por las versiones más nuevas. Estas primeras acciones se están llevando adelante con ABBYY, a través del cual es posible convertir estos archivos a PDF/A, como se mostrara en los apartados precedentes.

Alternativas y problemas de una migración masiva

En la actualidad existen herramientas con las cuales es posible realizar una migración masiva de versiones viejas de PDF al formato estándar buscado (PDF/A-1a o 1b). Una de ellas es Ghostscript: se trata de un paquete de software basado en un intérprete de PostScript y Portable Document Format (PDF) de Adobe Systems, que entre sus muchas funcionalidades permite realizar la migración de formatos.

Para la migración masiva puede usarse Ghostscript desde un intérprete de comandos (cmd en Windows, una terminal en Linux), especificando la versión de destino. Esta opción resulta muy conveniente, pues se puede implementar un script para transformar todo un árbol de directorios, en lugar de ir convirtiendo archivo por archivo. Hay numerosos sitios en Internet (Superuser, por ejemplo) que explican cómo realizar estas operaciones de migración masiva a PDF/A. Una de las recomendaciones encontradas es la utilización de Ghostscript para convertir a PDF/A utilizando el comando *gs* y especificando, entre otras variables, a qué PDF/A se quiere realizar la transformación (1, 2, 3).

Fueron realizadas tres pruebas, a modo de ejemplo, con archivos de distinto peso y contenido (con imágenes o sin ellas), verificando el peso resultante de los archivos, la calidad de las imágenes y la posibilidad de selección de texto. Del mismo modo, se hicieron pruebas de exportación a PDF/A con Libre Office. Pero es necesario aclarar que estas pruebas se realizaron en etapas iniciales de este trabajo y que, tras haber seguido distintos experimentos (como las mencionadas pruebas 1, 2 y 3), entre otros, surge la preocupación ante el gran riesgo que implica una migración masiva (errores, pérdida de archivos, pérdida de calidad de imágenes, por ejemplo), por lo cual esta parte del experimento no se ha continuado.

3) Análisis de la información descriptiva de preservación (PDI)

Figura 6.30: Paquete de información donde se resalta la información descriptiva de la preservación (en naranja)

Fuente: Norma ISO 14721: 2012.

Metodología y resultados

La información descriptiva de la preservación (PDI), como queda expuesta en la figura 6.30, del paquete de información de la norma OAIS, está compuesta por información de *procedencia*, de *contexto*, de *referencia*, de *integridad* y de *derechos*. Esta información, de estar presente, es agregada (de manera automática) por el software que sustenta el repositorio (como en el caso de procedencia), o se incorpora a través de tareas que se corresponden con el flujo de trabajo de la administración, en los propios metadatos del objeto a preservar. Para verificar la corrección de la PDI, es preciso observar si estos metadatos están presentes en los ítems del repositorio y tienen los valores adecuados. Dada la cantidad de ítems presentes en el repositorio, surge de manera inmediata la necesidad de realizar esta tarea de manera automática porque, de lo contrario, el tiempo consumido en tareas de revisión por parte de la administración sería enorme.

Así, la metodología propuesta es el desarrollo de un validador que tome la forma de una tarea de curación. Las tareas de curación (*curation tasks*) son programas desarrollados en Java para añadir una funcionalidad adicional a las que otorga la instalación base de DSpace. Estas tareas están relacionadas con la gestión de los objetos del repositorio, de ahí el término "curación" utilizado, homologable a "preservación", en tanto se busca evitar alteraciones en los metadatos que representan al OD. El propósito de las tareas de curación aquí planteado es ejecutar tareas de preservación (en otras palabras, efectuar el mantenimiento) de ítems en el tiempo y a lo largo de todo su ciclo de vida. Una tarea de curación en DSpace puede aplicarse a nivel ítem, y pueden definirse tareas que se apliquen a un conjunto determinado de ítems, colecciones o comunidades, incluso al repositorio completo.

Además, desde el marco teórico que conlleva el concepto de curación, puede verse claramente que es posible evaluar y reportar el "grado de preservación" o el desgaste que cada ítem va sufriendo con el tiempo, por lo que su ejecución resulta muy útil para todo gestor de un repositorio. Una ventaja adicional es la posibilidad de expansión de esta metodología, por ejemplo desarrollando otras tareas que lleven a cabo las acciones de reparación de aquellos ítems que obtuvieron malos resultados en las validaciones anteriores. Por todo esto, el modelo de desarrollo fue enfocado a nivel ítem de DSpace, y la tarea de curation correspondiente que puede tratar con él.

Debido a la necesidad de que la solución sea incremental y fácil de expandir, se pensó en un modelo orientado a objetos, en donde se hace énfasis en la ductilidad y atomicidad de cada regla que el validador puede aplicar a un ítem determinado. Esto permite que el agregado de nuevas reglas pueda realizarse de manera simple y deductiva, reutilizando cada elemento desarrollado con anterioridad. El modelo diseñado cuenta con una clase abstracta *Rule* (figura 6.31), que posee una sola cualidad: la de evaluarse sobre un ítem determinado y devolver un resultado. Dicha clase puede ser expandida de manera sencilla, para implementar todas las reglas que sean necesarias; esto se puede ver, por ejemplo, en la codificación de la regla creada para validar la existencia de un identificador persistente como el handle.

Esta forma de implementar las reglas resulta muy útil por el hecho de que cada ítem se puede querer evaluar no sólo para una regla, sino para un conjunto arbitrario de ellas y, como una extensión, se puede agregar una serie ordenada de pasos, con "pesos"

distintos, que podrían aplicar varias validaciones seguidas a un ítem determinado, producto de un agrupamiento necesario para el usuario. Por ejemplo, idealmente podría pensarse en evaluar la PDI completa con sus 5 elementos: precisamente, el objeto *Recipe*, puede verse como una colección de Pasos (*RuleStep*), en donde cada paso consiste en la aplicación de una regla, en un orden y con un peso específico, que da cuenta de la importancia, definida por el usuario, de cada regla aplicada. De nuevo, se pensó no sólo en el individuo (ítem), sino en el conjunto, y se otorgó en el validador la posibilidad de predefinir recetas a aplicar para una colección o para una comunidad.

Finalmente, para extender las posibilidades se dejaron dos puntos a tener en cuenta para desarrollos futuros:

- Que la tarea de curation genere una salida o *output*, en forma de un objeto Reporte, que podría ser expandido si se necesitaran distintos tipos de reportes o distintos muestrarios de datos.
- Se creó un constructor (*RecipeBuilder*) de las recetas, para que se puedan definir distintas recetas, con distintos tipos y distintas reglas, de manera que el usuario esté ajeno al desarrollo del validador y solo tenga que escribir las recetas en los archivos correspondientes para ejecutar distintas validaciones en distintas ocasiones.

El modelo del validador se puede apreciar en la figura 6.31.

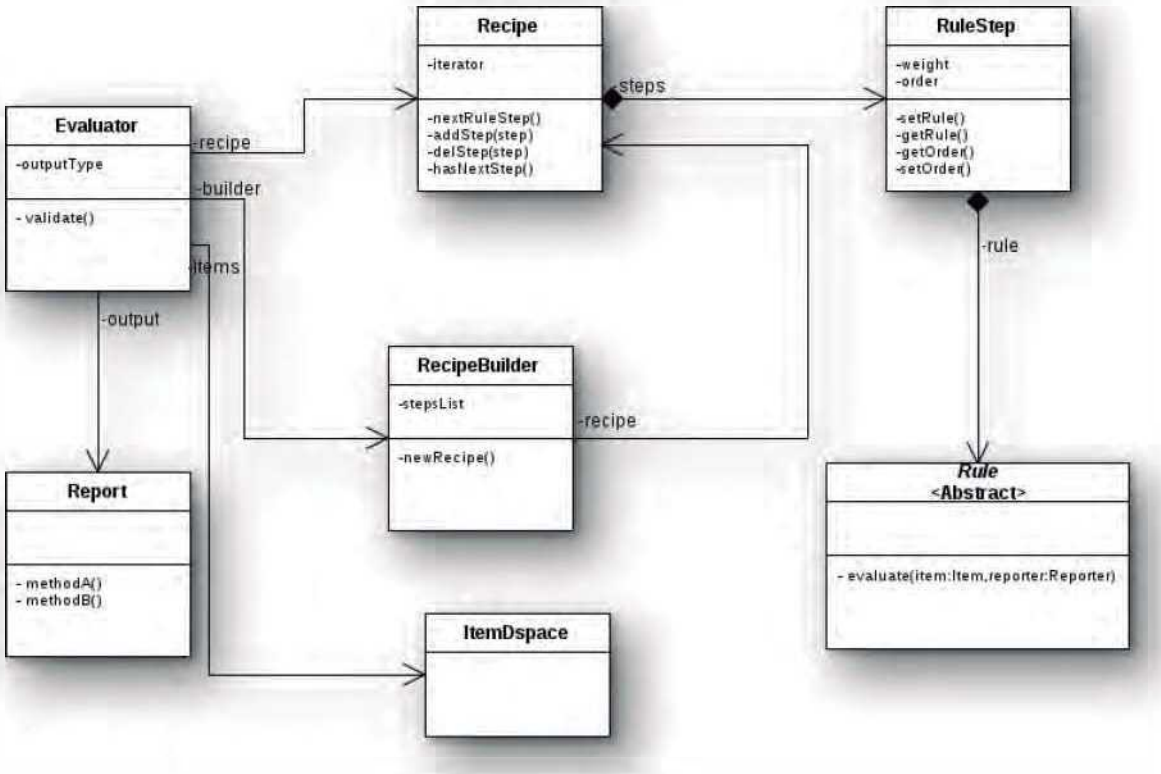


Figura 6.31: Modelo de validador en UML

Fuente propia.

La propuesta es la generación de reglas que evalúen de manera independiente y también global (que evalúen todos) los elementos que integran la PDI, a saber:

Referencia: se evalúa validando los identificadores persistentes; para el caso de DSpace se evalúa el handle. Esta regla es la que ya ha sido realizada y probada como manera de justificar la estructura propuesta para el módulo *Evaluate preservation*. El resto de las reglas se irán desarrollando durante 2014.

Integridad: se evalúa utilizando el checksum; para el caso de DSpace debe validarse que el algoritmo MD5 sea correcto y que no sea el que está por defecto.

Procedencia: se evaluará el metadato *provenance*, comenzando por ejecutar una consulta que muestre el contenido de ese metadato.

Contexto: se evaluará el contexto según OAIS. La información de contexto tiene como objetivo principal documentar las relaciones de la información de contenido con su medioambiente (por qué fue creada esa información de contenido y su relación con otra información de contenido). Esta es la regla que aparece como más compleja.

Acceso: se evaluará a partir de las licencias. En primer lugar se realizará una consulta que liste las licencias de cada ítem, y luego se evaluará cómo construir una regla que valide esto. La consulta y sus resultados se pueden observar en el próximo apartado.

Aclaración de algunos de los elementos de la PDI OAIS y la implementación en SEDICI-DSpace: *Provenance* y *Context*

Desde el punto de vista de OAIS, la información de referencia (*Reference Information*) es utilizada para la identificación unívoca del contenido: puede ser un identificador persistente (handle/DOI) y también, en algunas tipologías, un número único (por ejemplo, el ISBN para libros). La información sobre la integridad (*Fixity*) sirve para validar que no ha sido alterado el contenido. La información de derechos de acceso (*Rights*) identifica las restricciones de acceso al contenido, vinculadas a las licencias de distribución, de transformación, de uso y el embargo, en caso de existir. Ninguno de estos tres elementos de la PDI requieren mayores aclaraciones, pero sí es preciso hacerlas para procedencia (*Provenance*) y contexto (*Context*), que resultan más complejas de entender en términos de la norma y muy especialmente en su traslación al escenario particular SEDICI-DSpace.

Provenance

Provenance, en la norma OAIS, está dedicada a documentar la historia de la información de contenido. Por ejemplo, comienza con la fuente de origen de esa información, los cambios que se han sucedido y quienes han tenido custodia de la información desde su origen hasta el momento actual. Esto es útil fundamentalmente desde el punto de vista de los usuarios, porque aporta confiabilidad a la información a la que acceden, pero también sirve de soporte para asegurar la autenticidad del contenido. Como se verá en el decurso de estas aclaraciones, surgirán algunas dudas sobre estos metadatos, de las cuales ha dado cuenta la bibliografía y que tienen su origen en la dificultad que siempre existe al pasar de un modo abstracto a una implementación. En este trabajo se planteará una solución concreta.

La consulta específica sobre el metadato provenance se realizó a través de Solr,

generando una salida en CSV, que fue importada desde MS Excel. Se expondrá aquí un resumen de los resultados y su interpretación, para mayor facilidad de comprensión.

En SEDICI, este metadato lleva como nombre amigable el de "Registro de Origen" y su código en el esquema de metadatos de Dublin Core es *dc.description.provenance*. Este metadato se encuentra oculto por default en DSpace, debido a que contiene datos sensibles como el email del autor, del administrador, etc. Sólo puede verse desde la edición avanzada en la administración. Sin embargo, este metadato se carga cuando se recolecta por OAI o se ingresa al documento por algún otro medio automatizado, así como se genera en la instalación del ítem y en el proceso de administración, por lo cual pueden aparecer varias instancias de *dc.description.provenance* para un mismo ítem.

Provenance, como se ve, no es un metadato que se coloque manualmente en SEDICI, sino que es el propio DSpace el que, en las distintas operaciones del *workflow*, lo va adicionando. Así, DSpace genera un metadato provenance que declara tanto el evento como el agente, dicho en términos de las entidades de PREMIS. En las operaciones donde se presupone que puede haber posibilidad de alteración de contenido, calcula también el checksum (MD5).

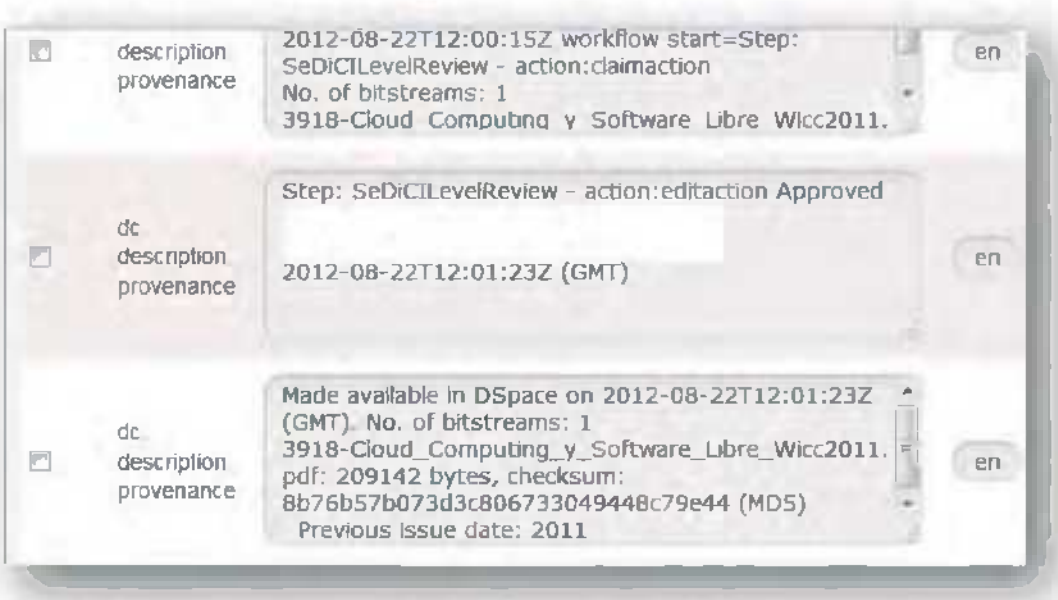


Figura 6.32: Metadato *provenance* en el flujo normal de la administración en DSpace

Fuente propia.

En el caso de la figura 6.32, puede verse que el ítem ha sido trabajado en la

administración (el nombre y email del agente se han borrado). La primera ocurrencia del metadato se corresponde con el ingreso del ítem al repositorio (depósito mediado por administración); la segunda ocurrencia es un paso de revisión en el que sólo se han agregado metadatos y se ha aprobado para su inmediata publicación; el tercer provenance alude a la instalación (*issue date*), es decir, a la publicación en sí del ítem (como es un proceso del propio DSpace no hay un agente que intervenga). En los dos casos en que se produjo algún cambio en los bitstreams del ítem, se ha calculado el checksum; como en el segundo paso, el agente no ha cambiado nada el cálculo no se ha realizado.

El segundo caso, que se muestra en la figura 6.33, corresponde a un ítem que ya estaba instalado en el repositorio SEDICI cuando el mismo se gestionaba con Celsius-DL, software propio utilizado desde la creación del repositorio en 2003. En el mayo de 2012, cuando se realizó la migración hacia DSpace, ese ítem se exportó por medio de un proceso automático; como puede observarse, el metadato provenance entre otras cosas registra dicha exportación.

Figura 6.33: Metadato *provenance* para un archivo exportado desde Celsius-DL hacia DSpace

Fuente propia.

Un tercer caso es el que se muestra en la figura 6.34, donde puede observarse un autor (cuyos datos se han borrado) que ha hecho autoarchivo del ítem y ese es el

primer provenance que se genera. Luego, se genera el provenance del proceso de revisión de la administración y finalmente el de la instalación del ítem.

Figura 6.34: Metadato *provenance* generado durante el proceso de autoarchivo

Fuente propia.

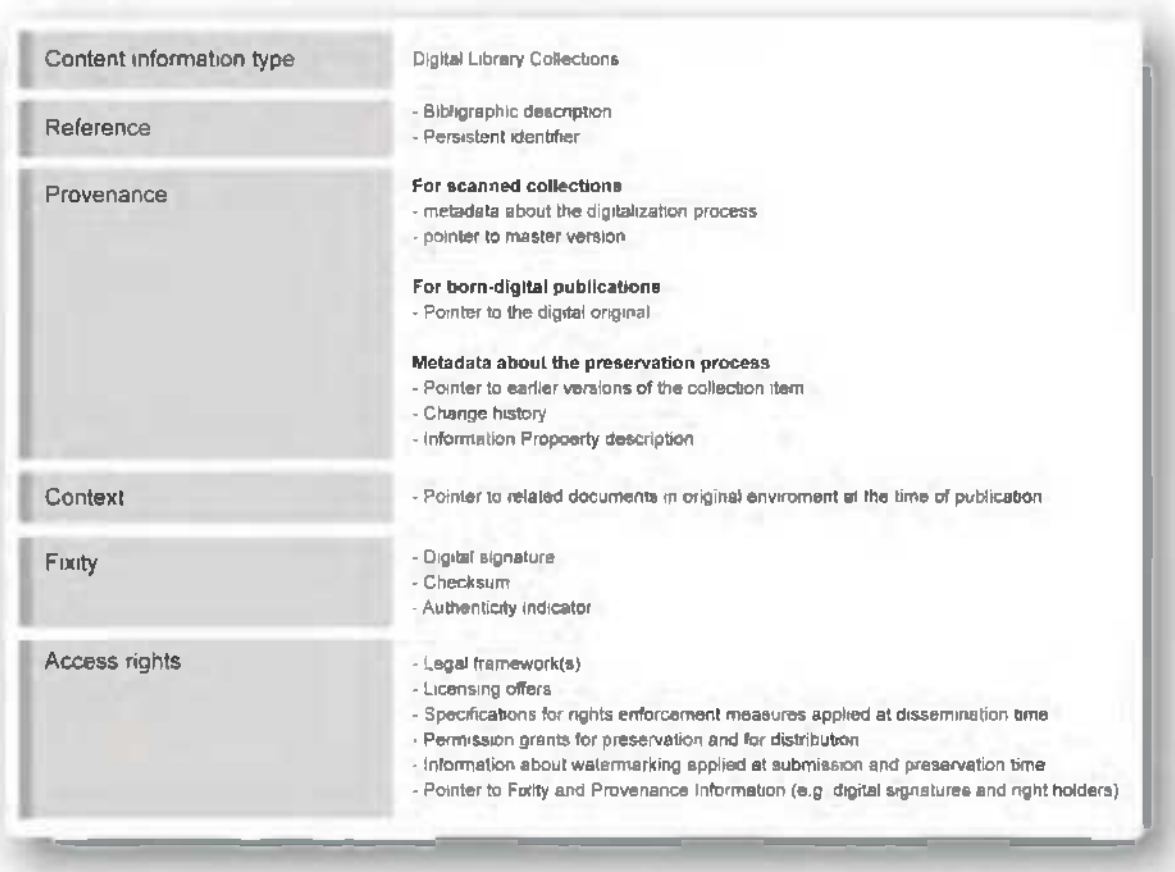
Context

OAIS describe el elemento Contexto en la PDI como el modo en que la información de contenido está relacionada con cualquier otra información por fuera del paquete de información. Por ejemplo (y esto puede resultar algo oscuro) por qué se produjo esa información de contenido, o una descripción de cómo se relaciona con otros objetos de contenido del repositorio. La norma excluye el contexto técnico, ya que esto pasa a la información de representación del objeto, e incluso parte de la información técnica que relaciona la información lógica con el medio físico que se integra en la información de empaquetado.

Es preciso hacer notar que la distinción que hace OAIS entre *Provenance* y *Context* no parece muy satisfactoria, en opinión de autores como Lupovici y Masanés (2000). La norma dice incluso que *provenance* puede verse como un tipo especial de *context*. Debido a estas ambigüedades, aquí se considera la noción de contexto de la PDI como todo aquello que se podría denominar "contexto externo", en el sentido de las relaciones internas a los objetos, como por ejemplo con qué otros objetos del

repositorio están vinculados o, en el caso de algunas tipologías documentales, por caso los artículos, si uno es una revisión de uno previo será parte del denominado "contexto interno", y las revisiones y validaciones se harán en acciones separadas para uno y otro.

Lo que se sugiere desde aquí es tratar de utilizar aquellas secciones de la norma donde se dan ejemplos, lo cual puede clarificar estas y otras cuestiones. La figura 6.35 es una adaptación propia de la tabla 4.1 de la ISO 14721, referida a posibles contenidos de un repositorio (digitales y digitalizados) con ejemplos de las partes de la PDI.



Content information type	Digital Library Collections
Reference	- Bibliographic description - Persistent identifier
Provenance	**For scanned collections** - metadata about the digitalization process - pointer to master version **For born-digital publications** - Pointer to the digital original **Metadata about the preservation process** - Pointer to earlier versions of the collection item - Change history - Information Propoerty description
Context	- Pointer to related documents in original enviroment at the time of publication
Fixity	- Digital signature - Checksum - Authenticity indicator
Access rights	- Legal framework(s) - Licensing offers - Specifications for rights enforcement measures applied at dissemination time - Permission grants for preservation and for distribution - Information about watermarking applied at submission and preservation time - Pointer to Fixity and Provenance Information (e.g. digital signatures and right holders)

Figura 6.35: Ejemplos de los elementos de la PDI para una colección de una biblioteca digital según OAIS

Fuente: ISO 14721.

Planteamiento de las reglas de validación

En este punto, ya se está en condiciones de afirmar que las validaciones se deben realizar tomando cuatro capas o niveles distintos, tal y como lo muestra la figura 6.36:

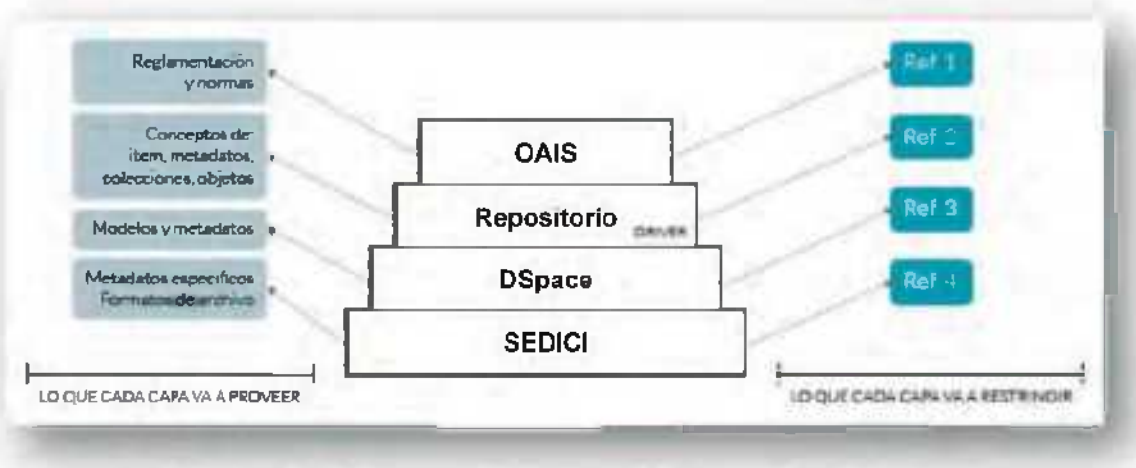


Figura 6.36: Capas posibles para las reglas de validación

Fuente propia.

Del lado izquierdo del gráfico, se muestran los recursos o las herramientas en las que se va a basar el validador para definir las reglas de validación. Del lado derecho, se muestran las restricciones que propone cada capa, que se deben tener en cuenta a la hora de plantear una regla para ese nivel.

Nivel OAIS (Ref. 1 en la figura)

Se debe evaluar:

- *Fixity* (de acuerdo a la documentación planteada)
- *Autenticidad* (de acuerdo a la forma propuesta por OAIS)
- *Contexto* (de acuerdo a la forma propuesta por OAIS)
- *Accesibilidad* (lo necesario para asegurar que siempre se pueda acceder al ítem). Al momento, evaluado en relación a la representación de los objetos del repositorio.
- *Legibilidad* (lo necesario para asegurar que los formatos de ítems pueden leerse y reconocerse). Al momento, evaluado en relación a la representación de los objetos del repositorio.

Nivel Repositorio (Ref. 2 en la figura)

Se debe tener en cuenta que deben cumplirse ciertas políticas, como:

- *Reglamentación de metadatos*: directrices DRIVER 2 u OpenAire.

- *Contenido*: tipologías y formatos permitidos.
- *Licencias*: licencias requeridas con las que debe contar el ítem (por ejemplo, de distribución y de uso).
- *Preservación*: se engloban en la fase tecnológica:
  - Conservación del archivo original y conservación en algún formato apto para la preservación.
  - Existencia de un identificador persistente que haga ubicuo al objeto y permita su acceso "para siempre".
  - Reglas de accesibilidad (que pueda verse siempre): incluso si se realiza la migración del objeto para mantenerlo en el tiempo.
  - Validaciones sobre el objeto digital.

Nivel DSpace (Ref. 3 en la figura)

Restricciones impuestas para DSpace:

- Cálculo y validación del checksum con MD5.
- Handle (siempre existe en DSpace), validación del handle (que sea válido) y que no sea el que trae la instalación por defecto.
- Formatos: que los formatos de archivo sean los que DSpace considera válidos.

Nivel SEDICI (Ref 4 en la figura)

Las restricciones impuestas por SEDICI están vinculadas fundamentalmente a las políticas del repositorio: políticas de contenido, de preservación, de metadatos, etc.

Construcción de las Reglas de Validación

Teniendo en cuenta lo que provee y restringe cada capa, se plantean a continuación las reglas que el validador debe corroborar:

Regla #1

Enfoque: DSpace

Nombre: "Verificar validez de handle"

Restricciones: el handle siempre existe en DSpace, puede ser el handle predefinido

Regla: El ítem contiene un handle válido y no se trata del handle por defecto de (123456789)

Metadato(s) asociado(s): *dc.identifier.uri* (sedici.identifier.handle)

Respuesta esperada: True → Válido

False → Inválido

La regla verifica que el handle que fue asignado para el ítem sea válido; este valor figura en el medatato *dc.identifier.uri*. Se considera un handle válido cuando no es el que viene predefinido por DSpace (12345678), como se dijo, y cuando el contenido alojado en el metadato *dc.identifier.uri* es igual al guardado en la tabla Handle de la base de datos de DSpace. Para realizar la validación correspondiente, se verifican las tablas del modelo de datos de SEDICI (figura 6.37).

Figura 6.37: Tablas del modelo de datos de SEDICI a consultar por la Regla #1

Fuente propia.

Se obtiene entonces el valor del handle utilizando como referencia el ID interno del ítem (*item_id*) y se constata contra la información contenida en la tabla *metadatavalue* que almacena el valor del metadato y *metadatafieldregistry* que almacena el nombre del metadato correspondiente al metadato *dc.indentifier.uri*. Si el handle en este metadato es igual al almacenado en la tabla *item_id*, entonces es válido.

Regla #2

Enfoque: DSpace

Nombre: "Verificar validez de MD5"

Restricciones: el checksum no debe ser el que viene por defecto en DSpace y debe ser correcto según el uso del algoritmo correspondiente, MD5 en este caso.

Metadato(s) asociado(s): - (aunque podría verse al checksum como un metadato puesto que se está trabajando en contexto DSpace).

Ejemplo: [DSpace]/bin/DSpace checker -a 123456/999).

Respuesta esperada: True → Válido

False → Inválido

Esta regla de validación verifica que se mantenga la integridad de cada bitstream de cada bundle del ítem correspondiente. Para esto recurre al módulo *checker* que trae DSpace, que verifica el checksum correspondiente a cada ítem, comparándolo contra un checksum esperado almacenado en la base de datos. El ítem sólo es válido cuando todos sus bitstreams han arrojado "CHECKSUM_MATCH" como resultado de la verificación. Para el caso de DSpace, el checksum se verifica y calcula usando el algoritmo MD5. La regla consulta las tablas de la figura 6.38.

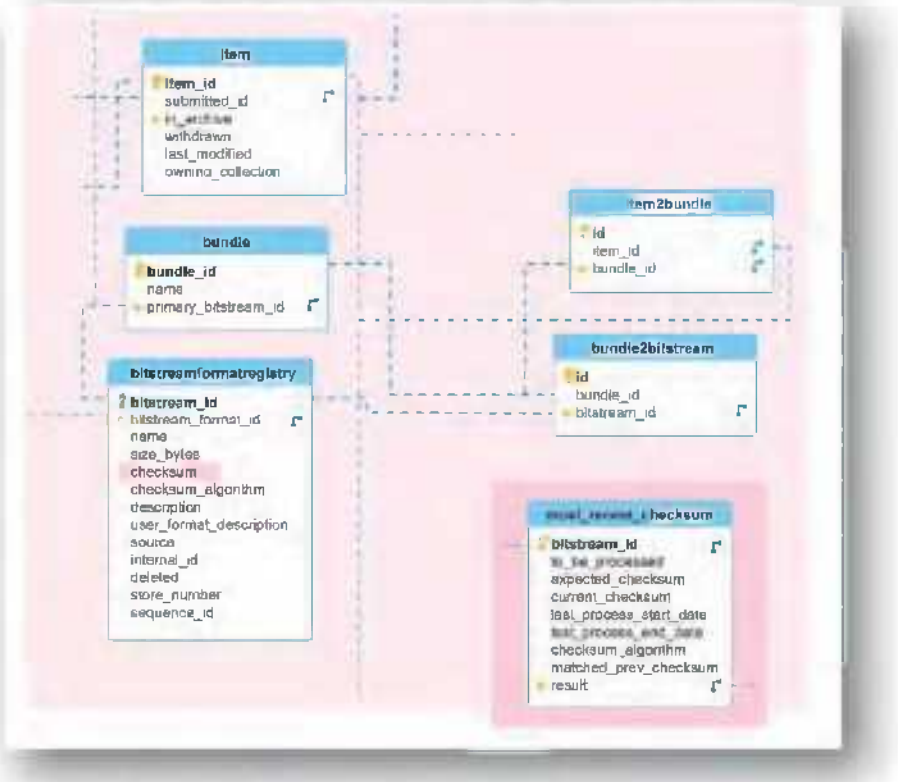


Figura 6.38: Tablas del modelo de datos de SEDICI a consultar por la Regla #2

Fuente propia.

Se utilizaron las tablas *item2bundle* y *bundle2bitstream* para detectar cuáles

bitstreams pertenecían al ítem correspondiente. Luego, para cada uno de esos bitstreams, se calculó el MD5 y se comparó con el asignado para su id (*bitstream_id*) en la tabla *most_recent_checksum*, de modo que el bitstream es considerado válido y arroja un "CHECKSUM_MATCH" cuando el *expected_checksum* y el *current_checksum* son iguales. (Notar que también se guarda en la tabla el checksum recalculado <<*current_checksum*>>, con el objetivo de ejecutar una tarea de corrección de ítems inválidos a futuro).

Regla #3

Enfoque: DSpace

Nombre: "Verificar validez de las licencias" (verificar *rights*)

Restricciones: todo ítem debe tener una licencia asignada, por lo que el metadato correspondiente a la licencia CC debe existir siempre. Además, la URI de la licencia debe también existir y debe ser la correspondiente a la elegida. Se añaden también las validaciones correspondientes sobre los embargos aplicados. Si la fecha de fin del embargo es mayor a la actual, significa que hay una restricción en el acceso al ítem, por lo que el mismo no debería validar la regla porque dará false.

Metadato(s) asociado(s): *sedici.rights.license*, *sedici.rights.uri*, *sedici.embargo.liftDate*

Respuesta esperada: True → Válido

False → Inválido

Esta regla verifica la validez de las licencias asociadas a los ítems. Para cada ítem se chequea que tenga una licencia asignada y que dicha licencia tenga asignada su URI correspondiente. Si cumple con los dos requisitos, el ítem es válido. Además, se chequea que exista o no un embargo y que la fecha de fin del embargo sea mayor a la fecha actual. Si cumple con los dos requisitos, el ítem es válido. Para realizar esto, se consultan las tablas de la figura 6.39.

Figura 6.39: Tablas del modelo de datos de SEDICI a consultar por la Regla #3

Fuente propia.

Para cada ítem, se obtiene la licencia y la URI que tiene asociada, utilizando las tablas *metadatavalue* y *metadatafieldregistry* para buscarlas. También se obtiene el metadato *sedici.embargo.liftDate*, que corresponde a la existencia o no de un embargo; todo referenciado con el ID interno (*item_id*) del ítem. Para cada tipo de licencia, se conoce cuál es la URI que debe asignarse, de modo que si la licencia es válida y tiene su URL correspondiente, entonces el ítem es válido para la regla. Por ejemplo un ítem sin embargo y con el metadato licencia:

sedici.rights.license: Attribution-NonCommercial 3.0 Unported (CC BY-NC 3.0)

debe tener necesariamente el campo:

sedici.rights.uri: https://creativecommons.org/licenses/by-nc/3.0/

Regla #4

Enfoque: DSpace

Nombre: "Verificar validez del contexto externo" (verificar *context*)

Restricciones: todo ítem debe tener una localización física y electrónica válida. En DSpace esto se refleja en que los metadatos *mods.location* y *sedici.identifier.uri* existan y sean válidos.

Metadato(s) asociado(s): *mods.location*, *sedici.identifier.uri*

Respuesta esperada: True → Válido

False → Inválido

Esta regla verifica el contexto externo del ítem, validando que la información de contexto exista y esté correctamente asociada al ítem. Para hacer esto, consulta los metadatos correspondientes a la localización física (*mods.location*) y electrónica (*sedici.identifier.uri*) asignados al ítem y verifica su existencia y validez. Se usan las tablas de la figura 6.40 del modelo de DSpace. Se hace notar que en las últimas tres reglas las tablas consultadas son las mismas, lo que varía es el nombre del metadato que se identifica a partir de la tabla *metadatafieldregistry*.

Figura 6.40: Tablas del modelo de datos de SEDICI a consultar por la Regla #4

Fuente propia.

Para cada ítem, se buscan, con las tablas *metadatavalue* y *metadatafieldregistry*, los metadatos *mods.location* y *sedici.indentifier.uri* y se chequea que existan y sean válidos, utilizando el id del ítem como referencia (*item_id*).

Regla #5

Enfoque: DSpace

Nombre: "Verificar validez del contexto interno para artículos"

Restricciones: el contexto interno a verificar es aquel que varía de acuerdo con el tipo de ítem que se está analizando (sus metadatos asociados son distintos).

Metadato(s) asociado(s): *sedici.relation.isReviewedBy*, *sedici.relation.isRelatedWith*, *sedici.relation.isPartOfSeries* (aunque a nivel DSpace sólo se debería verificar *dc.relation* y *dcterms.isPartOf*)

Respuesta esperada: True → Válido

False → Inválido

Ya que el contexto interno es dependiente del ítem que se está procesando, esta regla se aplica solo a ítems cuyo tipo documental es "Article", y se valida entonces que ciertos metadatos relacionados con el contexto interno de los artículos existan y estén relacionados. Para esto se utilizan las tablas de la figura 6.41.

Figura 6.41: Tablas del modelo de datos de SEDICI a consultar por la Regla #5

Fuente propia.

Para cada ítem, lo primero que se hace es consultar su *dc.type*, que representa a su tipo documental principal. Si su tipo es "article" puede aplicarse la regla, en caso contrario el ítem es omitido. De los artículos se toman los metadatos correspondientes al contexto interno utilizando las tablas *metadatavalue* y *metadatafieldregistry*, referenciadas con el ID interno del ítem (*item_id*). Cuando se tienen estos datos, se valida que los mismos existan y referencien elementos válidos. En en el caso de *dcterms.isPartOf* se sabe que posee el handle de la colección que representa a la revista, por lo que se valida que dicho handle exista y pertenezca a una colección válida en el repositorio.

De manera similar a lo realizado en esta regla, podrían desprenderse otras. Por ejemplo, un artículo puede tener un *sedici.relation.isReviewOf*, una tesis debe tener un *thesis.degree.name*, y en ambos casos reglas como la precedente se ejecutan si el tipo de contenido es aquél para el cual corresponde el metadato. El contexto interno permitiría entonces crear muchas nuevas reglas; por ejemplo, verificar si un artículo está correctamente vinculado a una revista y la revista a una serie.

Regla #6

Enfoque: DSpace

Nombre: "Verificar información de procedencia" (verificar *provenance*)

Restricciones: la información de procedencia es información que DSpace completa de manera predeterminada para cada ítem al momento de instanciarlo. Se debe tener en cuenta que esta información debe existir y debe ser válida.

Metadato(s) asociado(s): *dc.description.provenance*

Respuesta esperada: True → Válido

False → Inválido

Esta regla, por el momento, valida que el metadato *dc.description.provenance* exista para cada ítem analizado. Para esto usa las tablas de la figura 6.42.

Figura 6.42: Tablas del modelo de datos de SEDICI a consultar por la Regla #6

Fuente propia.

Usando las tablas *metadatafieldregistry* y *metadatavalue* se verifica la existencia del metadato *dc.description.provenance*, referenciado con el ID interno del ítem analizado (*item_id*). En un futuro, esta regla podría realizar validaciones más complejas, analizando la validez de la información que DSpace carga por defecto en el campo *provenance*.

Ejecución de las tareas de curation y resultados obtenidos

Regla #1

Se ejecutó la regla desde consola para validar el handle de los ítems del repositorio. El resultado fue satisfactorio (como puede verse en el anexo titulado "reporte.txt") y su ejecución se muestra en la figura 6.43.

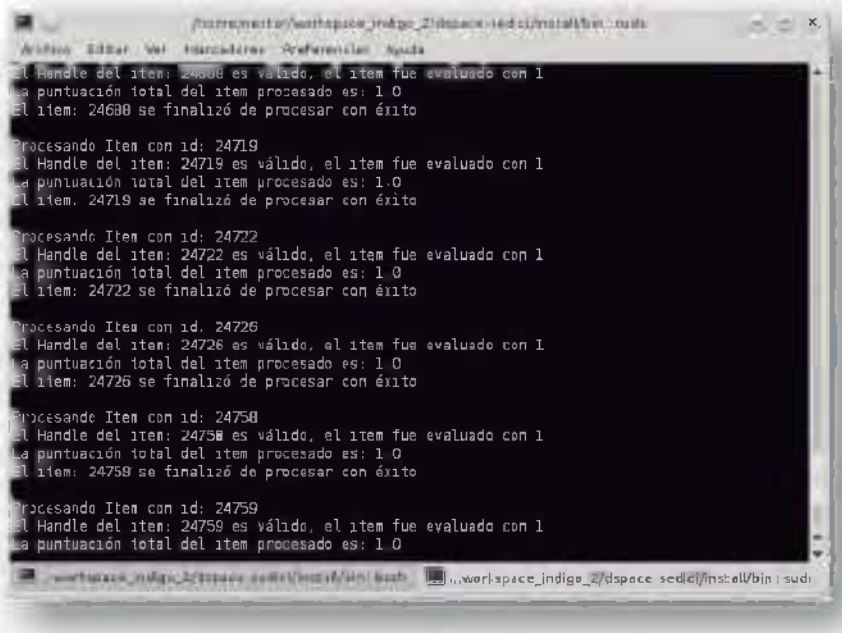


Figura 6.43: Ejecución de la Regla #1

Fuente propia.

Regla #2

Se ejecutó la validación con MD5 desde consola, el resultado fue satisfactorio y su procesamiento puede verse en la figura 6.44.

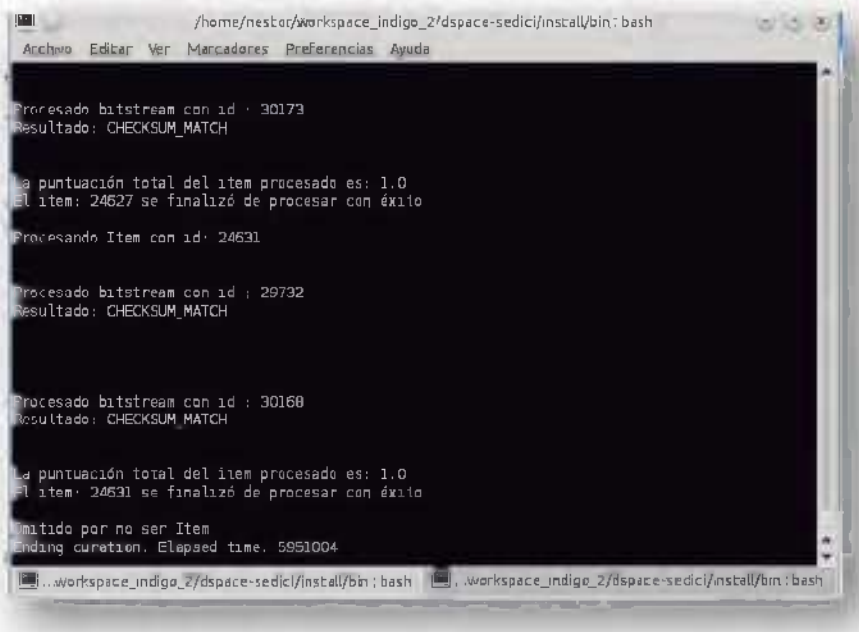


Figura 6.44: Ejecución de la Regla #2

Fuente propia.

Trabajos futuros

De la observación de la figura 6.36, referida a las posibles "capas" de las reglas, se plantean como trabajos futuros, a llevar adelante por otros integrantes de SEDICI, reglas más abstractas, pensadas más allá de que el repositorio se encuentre implementado en DSpace o en cualquier otro software para repositorios. Es decir, que se estarían considerando reglas a nivel repositorio. A continuación, se esbozan dos reglas más abstractas, la primera de ellas equivalente a la Regla #2 (DSpace), pero a nivel OAIS, que muestra la verificación de un identificador persistente en general. Esta regla habilitaría la verificación de un Digital Object Identifier (DOI) similar al handle e incluso otros identificadores vistos como persistentes desde el enfoque de la norma, como puede ser el caso del International Standard Serial Number (ISSN) de las publicaciones periódicas o el International Standard Book Number (ISBN) para libros. La segunda regla esbozada aquí, se correspondería con la Regla #3 (DSpace) pero habilitaría que el checksum se calculara con algún algoritmo diferente del MD5 que utiliza DSpace (ej. DES, AES, alguna variante de la familia SHA, etc.), con tal de que arrojara como resultado un valor comparable.

Reglas abstractas (enfoque OAIS-PDI, sin implementación)

Ejemplo de Regla #2

Enfoque: OAIS-PDI

Nombre: "Verificar existencia de Identificador persistente"

Restricciones: Según la reglamentación para almacenamiento se debe contar con algún tipo de identificador persistente.

Regla: El ítem contiene un identificador persistente.

Metadato(s) asociado(s): *dc.identifier.uri* (*sedici.identifier.handle*)

Respuesta esperada: True → Existe

False → No existe

Ejemplo de Regla #3

Enfoque: OAIS-PDI

Nombre: "Verificar validez de checksum" (*Fixity*: se usa para probar la autenticidad del AIP)

Restricciones: encontrar un checksum cuyos subelementos sean un valor y un algoritmo.

Metadato(s) asociado(s): -

Respuesta esperada: True → Válido

False → Inválido

Comprobaciones adicionales

En los casos en que resultó de interés, los componentes de la PDI fueron analizados a través de consultas directas a la base de datos para ver el estado de los ítems en relación a los metadatos vinculados a cada parte de la PDI. Como ya se habían validado *Fixity* e *Identifier*, y todos los ítems tienen ambos metadatos y son correctos, las consultas se realizaron sobre *Provenance*, *Rights* y *Context*.

Resultados obtenidos

Provenance

Como se explicó detalladamente en el apartado "Aclaración de algunos de los elementos...", todos los ítems del repositorio tienen por lo menos una instancia del metadato *provenance*, puesto que DSpace se encarga de su adjudicación en el momento en que el ítem ingresa al repositorio. Dada la extensión del reporte obtenido tras la consulta al indexador Solr referida a *provenance*, y ya que todos los ítems contienen el metadato (lo que implica que no debe realizarse ninguna acción), no se adjunta en anexos el archivo correspondiente, que sí se encuentra en el CD adjunto bajo el nombre "dc_description_provenance.xls". La consulta en Solr es sumamente sencilla al filtrar exclusivamente por ese metadato.

Rights

El caso del metadato sobre derechos representa una situación bastante diferente. Por un lado, SEDICI almacena (como ya se viera) el dato de la licencia de uso del ítem (cuál licencia Creative Commons ha elegido el autor) en un metadato y la dirección web del sitio oficial de esa licencia en otro. Sin embargo, desde la migración del repositorio hacia DSpace, muchos ítems cargados anteriormente tienen sólo una licencia en papel que no ha sido incorporada, y de la que no se incluyen aún sus datos. Esto hace que existan muchos ítems en los cuales no está explícito el metadato licencia. Por ello, se han realizado las consultas que incluyen a los metadatos que, en SEDICI, mapean al campo *dc:rights*. Estos son *sedici.rights.license* y *sedici.rights.uri*. Con esta información, se generó un reporte cuyo archivo se encuentra en anexos bajo el nombre "dc_rights.xlsx", una de cuyas pestañas es para los ítems que contienen los metadatos *rights* y otra para aquellos que carecen de ellos. En el archivo, además del handle que identifica al ítem, se han agregado el título y el subtipo. Dada la extensión de dicho reporte, no se anexa copia en papel.

De las consultas surge que hay 19.426 ítems que tienen licencia, es decir en los cuales se encuentran los metadatos *sedici.rights.license* y *sedici.rights.uri*. La figura 6.45 muestra la relación de ítems con y sin licencia.

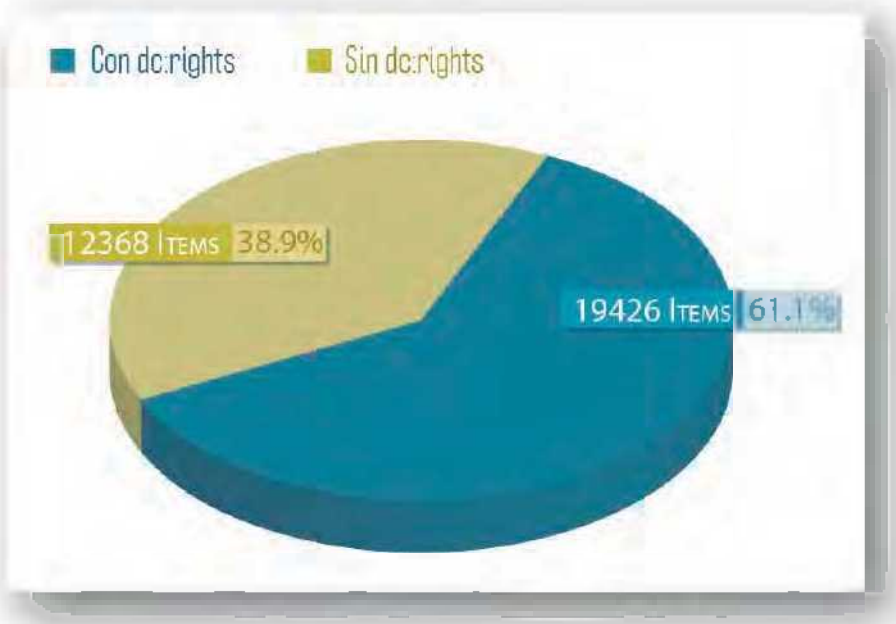


Figura 6.45: Porcentaje de ítems con y sin metadato *dc:rights* (licencia)

Fuente propia.

Sin embargo, de esos 19.426 ítems, no todos tienen el nombre correcto de la licencia, es decir, el metadato *sedici.rights.license*. Se encontraron 788 ítems con errores en el nombre de la licencia, problema que no se resuelve sólo con corregir esos ítems. Como la licencia se elige explícitamente, se descubrió que es en el flujo de trabajo donde las licencias propuestas tenían un error en el *string* que presentan debido a un error de configuración de ese dato en DSpace. En otras palabras, este trabajo sirvió para detectar un problema más importante (surgido por algún cambio reciente) que iba a continuar a lo largo del tiempo. Se generó asimismo una tarea en el gestor de incidencias para corregir el problema que se visualiza de manera abreviada en la figura 5.50. El reporte que expone esos ítems con el error en sus licencias se adjunta en un archivo que lleva el nombre: "Metadato Licencia para corregir.xlsx".

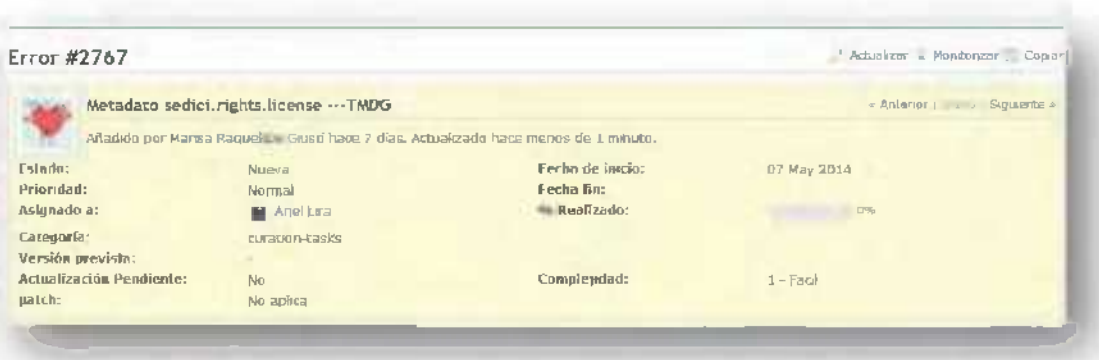


Figura 6.46: Captura de pantalla del reporte de error del metadato licencia

Fuente propia.

En relación a los 12.368 ítems sin licencia, el problema está vinculado a razones diferentes:

- Existen ítems que tienen asociada una licencia en papel (el archivo digital de éstas aún no fue realizado y/o no se ha incorporado al flujo de trabajo).
- Las licencias de distribución de SEDICI, previas a la versión 1.4 (de mayo de 2012), no obligaban a los autores a elegir licencias de uso Creative Commons, sino que solamente incluían la autorización del autor para la transformación de su obra con fines de preservación y el permiso de difusión.

Como hoy en día se considera que todos los ítems deben tener una licencia de uso, se ha generado una tarea para corregir el problema, que se muestra en la figura 6.47.

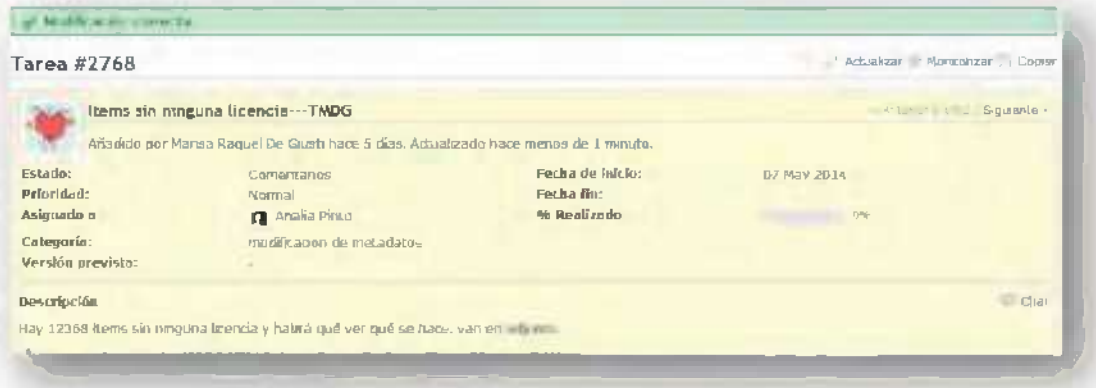


Figura 6.47: Reporte de error: ítems sin licencia

Fuente propia.

El tema de los derechos llevó a hacer varias consultas extra en Solr. La primera de

ellas estuvo dedicada a completar los registros con los subtipos para establecer prioridades para la tarea precedente. La figura 6.48 muestra un panorama amplio sobre los tipos y subtipos de documentos a los que le falta la licencia de uso.

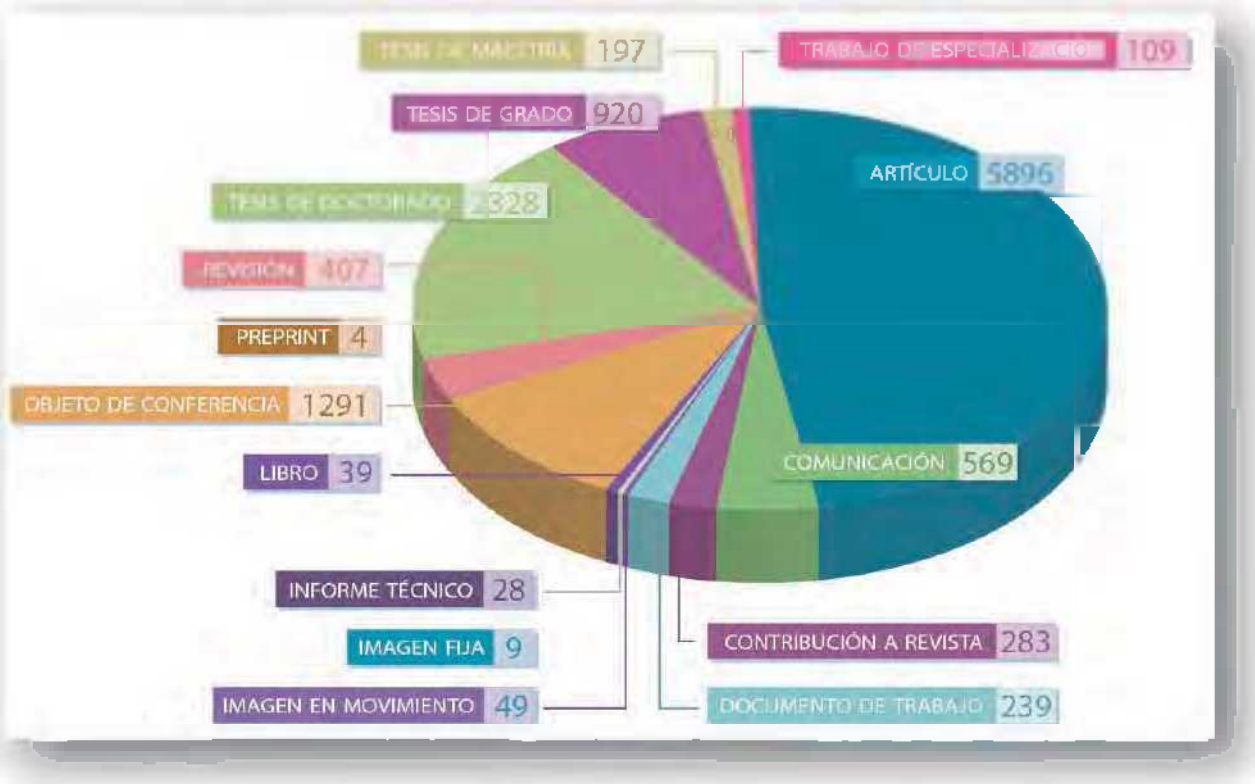


Figura 6.48: Ítems sin licencia CC clasificados por tipo documental

Fuente propia.

El procedimiento a seguir es distinto según la tipología documental de que se trate y, por ejemplo en el caso de las tesis, también según las fechas. En este sentido, existen 2771 ítems que son tesis en sus variantes de grado, maestría y doctorado que no cuentan con licencia. Se encuentran en el reporte cuyo archivo digital se adjunta bajo el nombre "Informe_completo_tesis-1.xlsx", donde también se tienen, en otra pestaña, los ítems tipo tesis que sí tienen licencia. El reporte identifica la unidad académica y el título de la tesis. En todos los casos, la posibilidad de agregar una licencia de uso es muy recomendable si existe el archivo (el bitstream) en el repositorio, ya que como se ha visto existen también muchos ítems sin bitstream por distintas razones.

Luego, se consultó a qué revistas pertenecían todos los subtipos del tipo artículo:

artículo, comunicación, revisión y contribución a revista, de modo de determinar las prioridades de las acciones según el estado y situación de las revistas:

- revistas discontinuadas;
- revistas que están en el Portal de Revistas Científicas de la UNLP y que, por tanto, tienen la obligación de preservar sus contenidos en SEDICI y que ya tienen puesta allí una licencia de uso;
- revistas que pertenecen a editoriales como Elsevier, lo que amerita realizar un análisis de las licencias (vale hacer notar que en estos casos no se tiene la revista completa en el repositorio, sino sólo artículos sueltos).

El caso de los artículos también ha arrojado algunos datos adicionales, como por ejemplo artículos sueltos que sí pertenecen a alguna revista, pero cuyo metadato *sedici.relation.journalTitle* (nombre amigable: título de la serie) no estaba completo. Este error en particular se produjo durante la migración desde Celsius-DL a DSpace. En este caso se ha generado la tarea correspondiente, que se ve en la captura de pantalla de la figura 6.49, para solucionarlo.

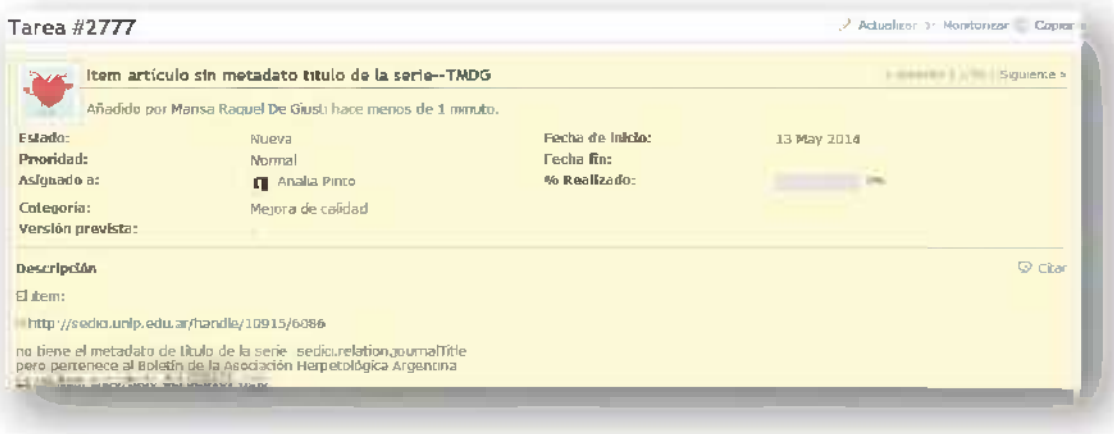


Figura 6.49: Reporte de tarea: falta metadato "Título de la serie" en artículos pertenecientes a revistas

Fuente propia.

A la tarea se ha adjuntado, además de los ítems revisados, otro archivo con 65 ítems que pertenecen al tipo artículo y que no tenían ningún título de revista incorporado.

Volviendo a la situación de los artículos y de los distintos casos que se plantean para

las revistas (discontinuadas, presentes en el portal o bien artículos sueltos), se han encontrado 4210 ítems sin licencia que pertenecen al tipo artículo en algunos de sus subtipos y que están a texto completo (metadato *sedici.description.fulltext=true*). Se ha particularizado esta búsqueda porque se considera prioritario completar los datos de licencia de los ítems que están a texto completo. Estos pertenecen a 54 revistas diferentes, muchas de ellas presentes en el Portal de Revistas Científicas de la UNLP. Se ha realizado un reporte que no se adjunta en papel, pero sí su archivo bajo el título "Artículos sin licencia y nombre de la revista y fulltext.xlsx".

Se ha generado otra tarea (figura 6.50) para analizar e incorporar las licencias faltantes, y se ha compartido un archivo para completar con los datos ya obtenidos.

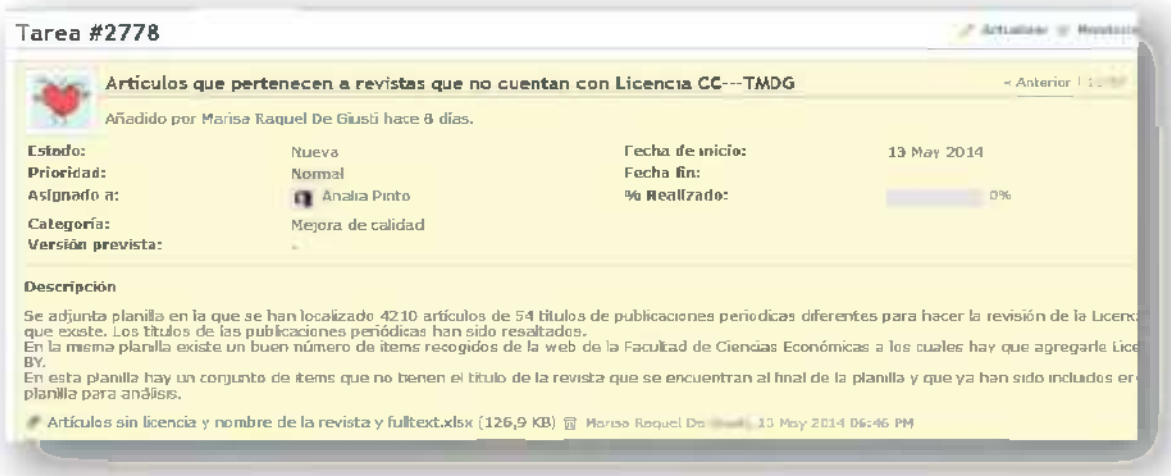


Figura 6.50: Reporte de tarea: revisar las licencias de las revistas

Fuente propia.

Una acción inmediata que puede remarcarse y que resultará muy efectiva es la adición de licencia CC (en la versión que corresponda) a los artículos que pertenecen a revistas que están en el Portal de Revistas Científicas de la UNLP y que por lo tanto ya tienen una licencia.

Contexto

Se ha realizado la consulta sobre todos los ítems del repositorio con el fin de verificar los siguientes metadatos:

- *sedici.identifier.uri*: denominado localización electrónica, es un metadato obligatorio cuando el documento se encuentra alojado en otra base o repositorio. En el mismo se consigna la URL de la revista o del sitio que contenga el ítem en cuestión, o bien el enlace directo al archivo.
- *mods.location*: es el metadato que indica la localización física del recurso. Este metadato debe ingresarse de manera obligatoria cuando no se tiene el recurso, dada la necesidad de proveer a los usuarios de algún modo de localizar la obra.
- *sedici.relation.isReviewOf*, *sedici.relation.isReviewedBy* y *sedici.relation.isRelatedWith*. Los dos primeros sólo se instancian cuando se trata de artículos. El último es sumamente importante para establecer relaciones entre los ítems del repositorio. Estas relaciones pueden ser muy variadas, por ejemplo un libro puede estar vinculado a un congreso donde se hizo la presentación del libro.

Todo el reporte se adjunta en el archivo denominado "campos relation.xlsx".

El análisis ha permitido detectar algunas prácticas discutibles, o variables (lo que no es deseable) en cuanto a los contenidos de los metadatos *sedici.identifier.uri* y *mods.location*. Como es usual en estos casos se ha generado el ticket correspondiente (figura 6.51).

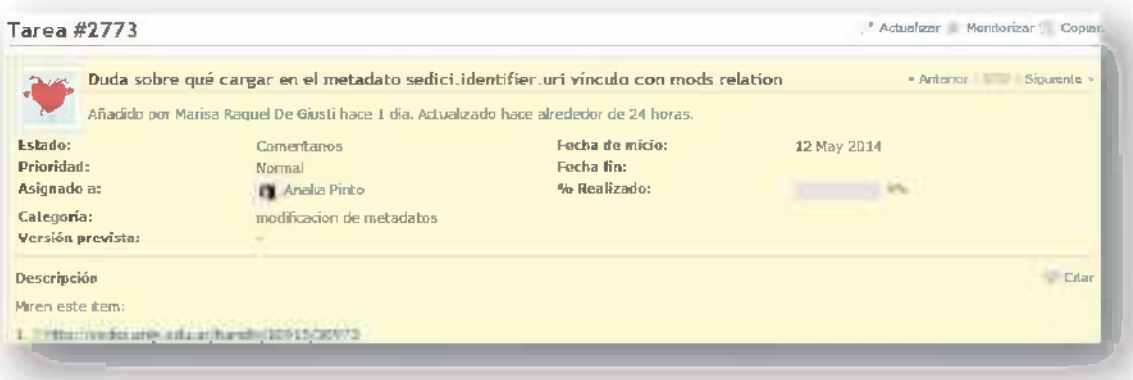


Figura 6.51: Reporte de tarea en el que se plantean las discrepancias encontradas en los metadatos localización electrónica y localización física

Fuente propia.

Un aspecto muy destacado de estos análisis es que se han detectado 958 ítems que no tienen ni fulltext, ni localización electrónica ni física. Con ellos se realizó un reporte

inicial, que luego fue mejorado, incorporando además los campos de tipo y fecha de publicación para, a partir de estos datos, priorizar las tareas. Este reporte se adjunta bajo el nombre "Ítems sin fulltext ni LE ni LF y fechas nuevo.xlsx".

En casos como este, se propone comenzar con los documentos más importantes, esto es las tesis, y en función de las fechas decidir su digitalización o localización de un documento relacionado que permita a los usuarios apreciar de algún modo el contenido de la obra original. Se sugiere indicar siempre, en el caso de las tesis, cuando menos su localización física (al menos la biblioteca de la unidad académica donde la tesis fue realizada), además de revisar si no existe en la biblioteca que corresponda algún dato de localización electrónica. Se generó el correspondiente ticket que se muestra en la figura 6.52.

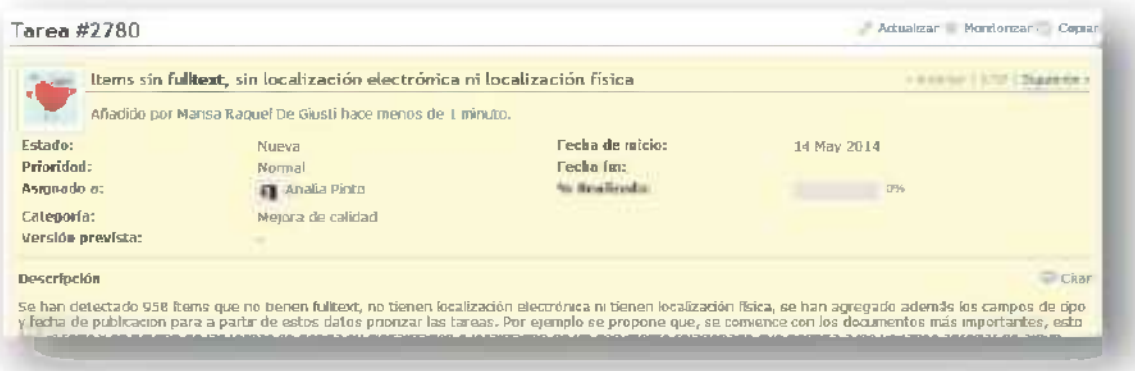


Figura 6.52: Reporte de tarea para ítems sin bitstream, sin localización electrónica ni física

Fuente propia.

4) Análisis de la Información Descriptiva

Los metadatos descriptivos existentes en el repositorio serán contrastados con las directrices DRIVER, ya mencionadas en otros puntos de este trabajo.

Figura 6.53: Paquete de información donde se resalta la información descriptiva (en lila)

Fuente: Norma ISO 14721: 2012.

Si bien las directrices DRIVER han sido establecidas pensando en recursos textuales, en esta verificación se extenderán para ser utilizadas como contraste sobre recursos no textuales también (de hecho, sobre todos los tipos del repositorio). En otras palabras, se verificará la totalidad de los ítems del repositorio para buscar si entre sus metadatos se encuentran todos los metadatos que, bajo las directrices DRIVER, se consideran obligatorios para la interoperabilidad entre repositorios.

Hay que destacar que en DRIVER el uso de los elementos puede ser:

- ***Obligatorio*** (M, por su sigla en inglés): el elemento debe estar siempre presente en el registro de metadatos.
- ***Obligatorio cuando aplica*** (MA): cuando el elemento se pueda obtener, debe estar presente en el registro de metadatos.
- ***Recomendado*** (R): el uso del elemento está recomendado.
- ***Opcional*** (O): no es importante si el elemento se usa o no.

La tabla 6.2 reproduce el estado propuesto de los metadatos según DRIVER:

Tabla 6.2. Metadatos descriptivos DRIVER

Basic element	Status	Encoding schemes
Title	M	None, free text
Creator	M	APA bibliographic writing style as in a reference list. Syntax: surname, initials (first name)
Subject	MA	Choice of keywords and classifications can be free text (preferably in English) and defined by an URI scheme (preferably info:eu-repo/classification)
Description	MA	None, free text. Recommended practice is to include an abstract in English. "Abstract" is the default interpretation to the value for dc:description
Publisher	R	None
Contributor	O	APA bibliographic writing style as in a reference list. Syntax: surname, initials (first name)
Date	M	Date	ISO 8601 W3C-DTF - "Published" is the default interpretation to the value for dc:date
Type	M	Publication type and Version type can be free text (preferably in English) and defined by an URI scheme (preferably info:eu-repo/semantics).
Format	R	IANA registered list of Internet Media Types (MIME types)
Identifier	M	URI scheme, linking to persistent identifier (URN, handle, DOI), full text document or human start page.
Source	O	Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata as in dcterms.bibliographicCitation
Language	R	ISO 639-3
Relation	O	None
Coverage	O	"Period" is the default interpretation to the value for dc:coverage Encoding: DCMI Period (http://dublincore.org/documents/2000/07/28/dcmi-period/) For more encoding schemas see Chapter 5 Use of vocabularies and semantics
Rights	R	None
Audience	O	None "Eduction level" is the default value for dc:audience.

De los metadatos descriptivos, según la sugerencia de DRIVER, son obligatorios: *Título*, *Creador*, *Fecha*, *Tipo* e *Identificador*. Los elementos *Description* que aparecen en el archivo dublin_core.xml y *Subject* en el archivo metadata_sedici.xml, en una prueba de exportación de un ítem (ver figuras 6.54, 6.55 y 6.56), también serán verificados para conocer qué porcentaje de los objetos del repositorio cuentan con esta información.

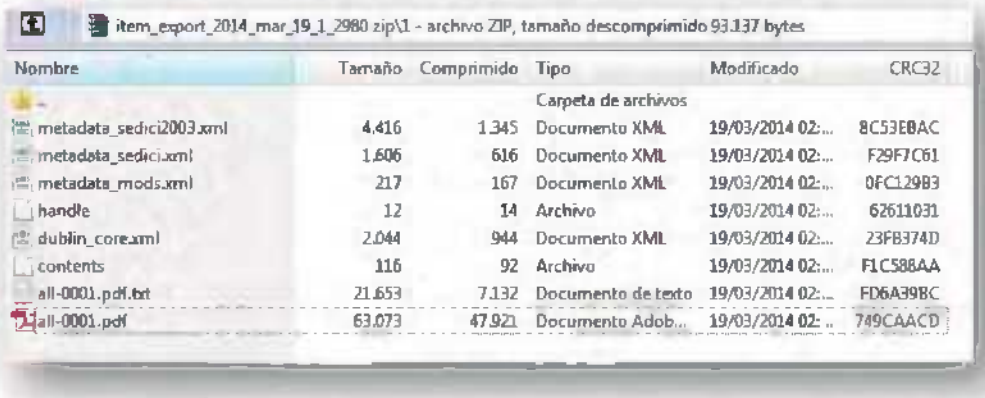


Figura 6.54: Archivos presentes en la exportación de un ítem

Fuente propia.

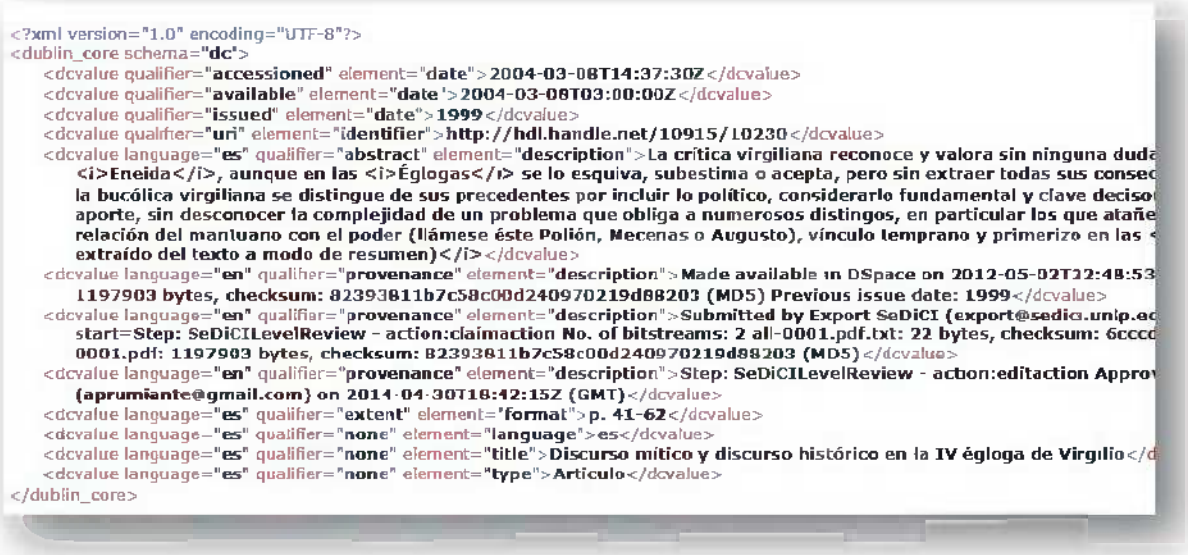


Figura 6.55: Archivo dublin_core.xml con elemento "description"

Fuente propia.

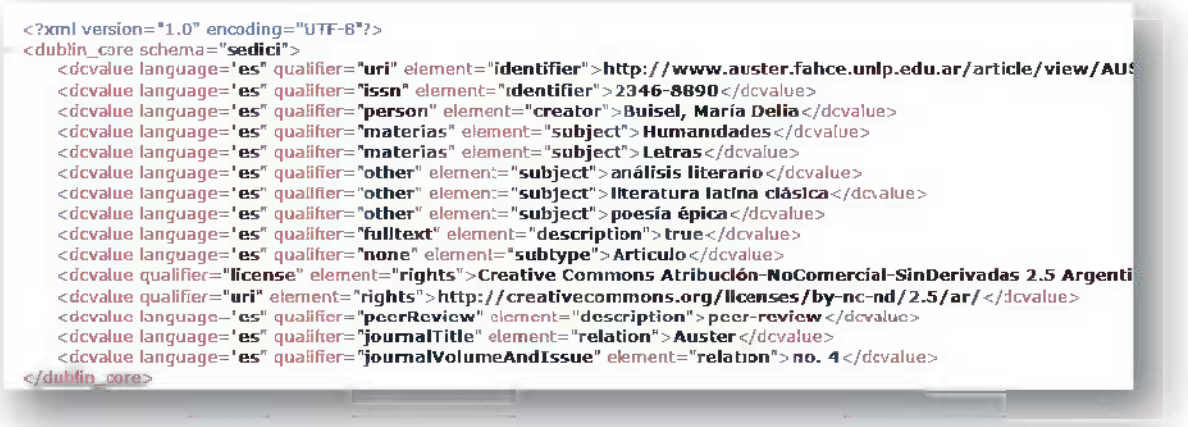


Figura 6.56: Archivo metadata_sedici.xml con elemento "subject"

Fuente propia.

Para verificar en el repositorio cuáles ítems cumplen con los requerimientos DRIVER, se necesitaron identificar los metadatos *unqualified* del esquema Dublin Core arriba especificados. Esto se realizó sobre los metadatos existentes para cada objeto digital, mediante el siguiente mapeo:

- Título (*dc.title=dc.title, dc.title.alternative*)
- Creador (*dc.creator=sedici.contributor.compiler, sedici.contributor.editor,*

sedici.contributor.translator, *sedici.creator.person*, *sedici.creator.corporate*, *sedici.creator.interprete*) (puede corresponder a cualquiera de estos, e incluso puede ser más de uno)

- Fecha de publicación (*dc.date=dc.date.issued*)
- Tipo (*dc.type=sedici.subtype*, *dc.type*)
- Identificador único (*dc.identifier=dc.identifier.uri*)

¿Cuáles objetos del repositorio cumplen con DRIVER?

Para corroborar esto, se realizó una serie de consultas sobre Solr, en particular, al motor XOAI, el cual está íntimamente relacionado en el proceso de *metadata harvesting* mediante OAI-PMH. Dicha consulta fue realizada teniendo en cuenta los parámetros arriba mencionados.

Una consulta previa permitió averiguar cuáles objetos en el repositorio cumplen con las directrices, pero ¿cuáles son los que no cumplen con DRIVER? Para averiguarlo, se realizó una diferencia de conjuntos entre el conjunto previo de todos los *objetos que cumplen con DRIVER* y el conjunto del total de *objetos que hay en el repositorio.* La diferencia da cero porque ambas consultas arrojan el mismo número, lo cual verifica que todos los ítems del repositorio cumplen con DRIVER.

Además de los metadatos considerados obligatorios en DRIVER, el experimento incluía la verificación de los elementos *Description* y *Subject*. Si se observa el archivo del anexo "Copy of Metadata y types sedici-DSpace.xls", puede verse que existen 8 metadatos vinculados a *Subject*:

1. *metadata.sedici.subject.materias*
2. *metadata.sedici.subject.lcsh*
3. *metadata.sedici.subject.decs*
4. *metadata.sedici.subject.eurovoc*
5. *metadata.sedici.subject.descriptores*
6. *metadata.sedici.subject.other*
7. *metadata.sedici.subject.keyword*
8. *metadata.sedici.subject.acmcss98*

De todos estos metadatos, el único obligatorio en el repositorio es Materias.

Inicialmente se realizó una consulta que abarcara todos los posibles metadatos SEDICI vinculados a *Subject*, y como resultado se detectó que sólo un archivo no contaba con ninguno. Se trata del ítem http://hdl.handle.net/10915/30365, el cual ha sido corregido. Por alguna razón que puede estar vinculada con problemas con la conexión a Internet al momento de ser cargado, el ítem había pasado por todas las etapas de revisión sin que se advirtiera la falta de un metadato que, en caso de no estar, impide que se siga adelante con la carga, dada su importancia.

En relación al metadato *Description*, existen dos metadatos en SEDICI vinculados a la descripción del contenido en sentido lato, los cuales pueden verificarse en el archivo anexo mencionado para el caso de *Subject*. Los dos metadatos de SEDICI toman los nombres amigables de "Notas" y "Resumen", y se corresponden con *sedici.description.note* y *dc.description.abstract*.

Ninguno de los dos metadatos es obligatorio. En el caso del resumen, el metadato es obligatorio cuando corresponde (MA) y opcional tanto en el autoarchivo (a menos que se trate de autoarchivo de tesis, donde es obligatorio) cuanto en el flujo de trabajo. El campo de notas se usa sólo ocasionalmente, mientras que el resumen puede ser muy importante en caso de no tener acceso al documento completo. Es este el metadato seleccionado para la consulta, en la que se identificaron 4887 ítems sin resumen, y se generó, en consecuencia, la tarea para la mejora de los ítems como se muestra en la figura 6.57.

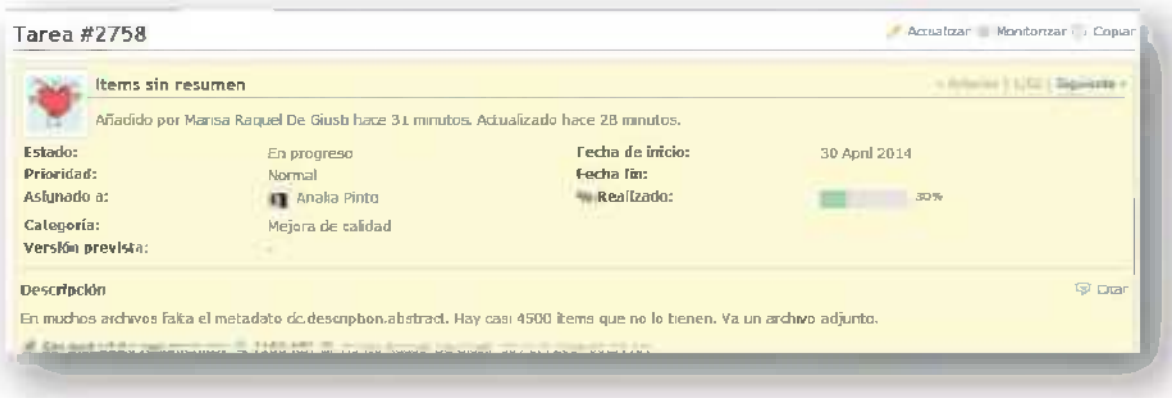


Figura 6.57: Reporte de tarea para los ítems sin resumen

Fuente propia.

Existen, además, 27.187 ítems que sí tienen resumen, pero en en cierta cantidad de casos, el resumen está repetido. El archivo con el total de ítems con resumen se anexa bajo el título: "Consulta sobre resumen v2.xlsx". Ese archivo se organizó para localizar los repetidos. Se armó un primer listado de resúmenes repetidos y a revisar que se anexa en el archivo "Resúmenes repetidos y para revisar.docx", que se incorporó luego al ticket #2758 de la figura 6.57.

Conclusiones

A lo largo de este capítulo se planteó el posible modelo de evaluación para el repositorio institucional SEDICI, elegido como caso de estudio. Se establecieron los antecedentes y proyectos que sirvieron de base para esta investigación. Se explicó la metodología y las herramientas a utilizar para llevar adelante la evaluación, la cual estuvo dedicada a determinar el cumplimiento del repositorio en relación a los elementos constitutivos del paquete de información del modelo OAIS de la norma ISO 14721, que es el modelo a seguir para demostrar las capacidades del repositorio y sus falencias.

Cada una de las partes del modelo abstracto del paquete de información (información de contenido y su representación; información descriptiva de la preservación; información descriptiva) fue contrastada siguiendo el método propuesto en este trabajo.

Para *la información del contenido y su representación*, se realizó el perfilamiento de todos los objetos del repositorio, determinando su riesgo de preservación en función de los formatos existentes y estipulando asimismo acciones que ya han sido iniciadas para mejorar el estado de los ítems (acciones de conversión, migraciones, presentación de estándares y elección de formatos para exposición y preservación de los ítems).

En relación a *la información descriptiva de la preservación*, además de trasladar el modelo abstracto de la PDI a la realidad de los metadatos en el repositorio, que dan cuenta de los cinco elementos (integridad, fijeza, procedencia, contexto y derechos), se desarrolló un validador y se establecieron las reglas con las que cotejar el cumplimiento con la PDI. También se realizaron numerosas consultas con el motor de indexación Solr para establecer el estado actual de los ítems en relación a los

metadatos implementados en SEDICI para cumplir con las funciones de los elementos de la PDI.

Finalmente, lo referido a *la información descriptiva* fue cotejado en relación a las directrices DRIVER e incluso dichas directrices fueron extendidas a otros metadatos descriptivos a la hora de la catalogación de los contenidos en SEDICI.

La labor de análisis llevó a la generación de numerosas tareas y trabajos para la administración del repositorio, de modo de llevar a la práctica las conclusiones del experimento y generar mejoras en la calidad de los contenidos. Esta tarea en la actualidad se encuentra avanzada y tendrá continuidad con nuevas propuestas.

Listado de anexos a los capítulos 5 y 6 (en papel y en CD)

Anexos en papel (se encuentran al final de la tesis, en el apartado "Anexos")

1) PRONOM: Detailed Report para PUID fmt/20
2) Comprehensive Breakdown
3) Metadatos y *types* en SEDICI

Anexos presentes en el CD que acompaña a esta tesis y localizables en el repositorio SEDICI en http://hdl.handle.net/10915/43157

1) Perfilsolofilesmarcado MD5 v2.xls
2) Informe de ítems sin bitstreams con y sin Links y agregados mios.xls
3) dc_ description_ provenance.xlsx
4) dc_rights.xls
5) Metadato licencia para corregir.xlsx
6) Artículos sin licencia y nombre de revista y fulltext.xlsx
7) Informe_completo_tesis-1.xls
8) Items sin fulltext sin LE ni LF y fechas nuevo.xlsx
9) Consulta sobre resumen v2.xlsx
10) Resúmenes repetidos y para revisar.docx
11) Copy of metadatos y types sedici-dspace.pdf
12) Campos relation.xls
13) Archivos PDFs versiones viejas.xlsx

Bibliografía del capítulo 6

ABBYY (2014). ABBYY FineReader 11. Recuperado el 24 de junio de 2014, de http://www.ABBYY.com/ocr_sdk_linux/.

Adobe Systems Incorporated (s/d). Niveles de Compatibilidad de PDF. Recuperado el 18 de mayo de 2014, de http://help.adobe.com/es_ES/acrobat/pro/using/index.html.

Adobe Systems Incorporated (2004). PDF Reference and Adobe Extensions to the PDF Specification. Recuperado el 24 de junio de 2014, de http://www.adobe.com/devnet/pdf/pdf_reference.html.

Adobe Systems Incorporated (2004). Adobe PDF Reference Archives. Recuperado el 24 de junio de 2014, de http://www.adobe.com/devnet/pdf/pdf_reference.html.

Callas Software (2008). pdfaPilot. Recuperado el 24 de junio de 2014, de http://www.callassoftware.com/callas/doku.php/en:products:pdfapilot

Deutsch, L. P., Artifex Software (1988). Ghostscript. Recuperado el 19 de junio de 2014, de http://www.ghostscript.com/.

Florida Center for Library Automation (FCLA). (2011). DAITSS Digital Preservation Repository Software http://daitss.fcla.edu/.

"Flying-sheep" (2012). SWF2SVG. En: GitHub. Recuperado el 25 de junio de 2014, de https://github.com/flying-sheep/SWF2SVG.

Gautier, P.; Wirzenius, L. (2004). SoundConverter. Recuperado el 25 de junio de 2014, de http://soundconverter.org/.

Google Inc. (s/d). Swiffy. Recuperado el 25 de junio de 2014, de https://www.google.com/doubleclick/studio/swiffy/.

GStreamer team (1999). GStreamer. Recuperado el 19 de junio de 2014, de http://gstreamer.freedesktop.org/.

ISO 32000-1:2008. *Document management — Portable document format — Part 1: PDF 1.7*. Recuperado el 26 de junio de 2014, de http://www.adobe.com/content/dam/Adobe/en/devnet/acrobat/pdfs/PDF32000_2008.pdf.

Lupovici, C.; Masanés, J. (2000). *Metadata for long term preservation*. Biblioteca Nacional de Francia. Nedlib Report series 2. Recuperado el 25 de junio de 2014, de http://www.kb.nl/sites/default/files/docs/NEDLIBmetadata.pdf.

Microsoft (2014) [MS-DOC]: Word (.doc) Binary File Format. Recuperado el 19 de junio de 2014, de http://msdn.microsoft.com/en-

us/library/office/cc313153%28v=office.12%29.aspx.

National Information Standards Organization (NISO) (2007). "A Framework of Guidance for Building Good Digital Collections". *NISO Recommended practices*. 3rd edition. Recuperado el 24 de junio de 2014, de http://www.niso.org/publications/rp/framework3.pdf.

Portable Document Format (s/d). En Wikipedia. Recuperado el 24 de junio de 2014, de http://en.wikipedia.org/wiki/Portable_Document_Format#Adobe_specifications.

Portable Network Graphics (PNG) (s/d). En Wikipedia. Recuperado el 25 de junio de 2014, http://en.wikipedia.org/wiki/Portable_Network_Graphics.

Portal de Revistas Científicas de la Universidad Nacional de La Plata (2013). Recuperado el 19 de junio de 2014, de http://revistas.unlp.edu.ar/cientificas/.

QaamGo Media UG (s/d). Online-converter. Recuperado el 19 de junio de 2014, de http://www.online-convert.com/.

Sama Rojo, V.; Sevillano Asensio, E. (2012). *Guía de accesibilidad de documentos electrónicos*. Recuperado el 24 de junio de 2014, de http://portal.uned.es/archivos/Capitulo_II_Accesibilidad_Word.pdf.

Stack Exchange (s/d). Superuser. Recuperado el 24 de junio de 2014, de http://superuser.com/questions/25598/linux-pdf-version-converter.

Standardization of Office Open XML (s/d). En Wikipedia. Recuperado el 29 de junio de 2014, de http://en.wikipedia.org/wiki/Standardization_of_Office_Open_XML.

Capítulo 7 | Conclusiones y trabajos futuros

Un repositorio institucional es un archivo digital donde una institución, por caso una universidad, alberga su producción de manera total o parcial, con criterios de selección de naturaleza diversa: contenidos científicos como artículos, tesis de posgrado, presentaciones en congresos, materiales educativos, incluidos complejos objetos de aprendizaje, proyectos de investigación, extensión, currículas, ordenanzas y otros materiales administrativos. Es resorte de la institución el determinar los alcances, la amplitud y las tipologías de materiales a depositar. La institución ofrece así un espacio único de albergue de su quehacer.

En las fases preliminares de planeamiento de un repositorio es central definir los contenidos, conocer las prácticas de publicación de los autores, los circuitos de trabajo y las necesidades institucionales; también es muy importante que el grupo que organiza un anteproyecto de repositorio, consulte otras experiencias similares, que revise los registros de repositorios del mundo, conozca los más importantes y especialmente aquéllos con cometidos semejantes y analice su experiencia, tratando de tomar los mejores ejemplos para cada aspecto del repositorio.

El repositorio institucional es una estructura tecnológicamente compleja a través de la cual la institución se compromete a brindar servicios a una comunidad designada, la cual puede ser muy compleja y diversa, tanto como la propia institución lo defina en su proyecto. La institución debe asimismo configurar previamente numerosas políticas que determinarán qué herramientas tecnológicas y flujos de trabajo tendrá su repositorio; deberá definir de antemano, por ejemplo, los materiales que va a guardar y esto deberá especificarse en la política de contenidos; deberá detallar cómo expondrá estos materiales, pues siempre es deseable que estén en acceso abierto, pero en algunos casos el repositorio deberá dar la posibilidad a los autores de elegir períodos de embargo en atención a compromisos con editoriales, patentes comerciales o convenios de confidencialidad. Desde luego, los responsables institucionales deberán informar

sobre los derechos de uso y difusión de los materiales albergados en el repositorio... Todos estos puntos deben quedar delineados en la política de acceso y se debe informar sobre los metadatos a agregar para gestionar los contenidos y exponerlos al mundo y compartirlos adecuadamente, permitiendo de este modo su localización, es decir, la llamada política de metadatos. Un repositorio debe tener claro también cuál será su política de preservación, de modo de dar accesibilidad y legibilidad de los contenidos a lo largo del tiempo. La institución deberá asimismo definir una política de servicios para toda la comunidad a la que pretende llegar, incluso servicios diferenciales según los usuarios, de modo de generar interés.

A partir de todas las definiciones precedentes se debe formar un equipo de trabajo (que, como puede dilucidarse rápidamente, deberá ser interdisciplinario y muy específico a la vez), pensar en la tecnología, en el plan de marketing y en la puesta en marcha de toda esta compleja estructura. Como se ve, todo este proceso de planificación no es trivial, y la correcta definición de las instancias mencionadas lleva a que el equipo que se arme para atender todos los aspectos actúe con liderazgo en cada uno de sus ámbitos de alcance: bibliotecarios, técnicos, administradores, informáticos, personal de diseño y comunicación.

Existen numerosos retos y desafíos a enfrentar durante la vida de un repositorio institucional; siempre resulta relativamente sencillo llegar a la puesta en marcha, pero los problemas vienen luego y derivan de numerosas áreas. Puestos en orden de importancia, pero también de inminencia (esto es lo que acontece en la propia experiencia de inicio de un repositorio) se pueden citar:

1. Difusión del repositorio; generar confianza en la institución.
2. Compartir y hacer conocer las ventajas del autoarchivo, para favorecer la ingesta.
3. Enseñar y crear conciencia sobre los derechos de autor.
4. Facilitar y orientar el autoarchivo de materiales, distribuyendo de este modo la tarea y creando la posibilidad de aumentar la tasa de carga.

Estos cuatro puntos son determinantes para lograr el cometido de poblar el repositorio, de modo que no sea sólo una magnífica realización tecnológica sin contenidos. A continuación se realiza un comentario más detallado de los problemas

expuestos.

1. A lo largo de toda la vida del repositorio será necesario mantener un fuerte compromiso institucional para generar mandatos, políticas de depósito, recursos para mantener el personal y cambiar las tecnologías, para realizar planes de formación de los usuarios, materiales de apoyo, planes de marketing. Este punto, que se podría llamar "compromiso institucional", es trascendental para que el repositorio sobreviva.
2. El repositorio deberá integrarse a otras redes, seguir estándares nacionales e internacionales para proveer sus datos, generar conjuntos de colecciones de acuerdo a los intereses de los proveedores de servicios.
3. El repositorio es una herramienta de la institución para brindar servicios, pero también para generar visibilidad de su producción y aumentar de este modo el impacto de obras, autores y quehaceres. En este sentido, el repositorio debe interactuar con otros portales universitarios de exposición de contenidos, gestores, portales de revistas, congresos y gestores de noticias, entre otros. Estos puntos están vinculados a la interoperabilidad del repositorio con otras realizaciones similares o compatibles, y significa un fuerte trabajo humano y tecnológico para lograr compartir recursos de manera automática y transparente.
4. Otro gran punto a mantener durante todo el ciclo de vida de los contenidos albergados en el repositorio es su accesibilidad, su legibilidad sin que esto cambie por el paso del tiempo, es decir, que las personas (y máquinas) siempre puedan encontrar los contenidos, comprenderlos y verlos de manera adecuada. Esta tarea es muy importante ya que significa trabajar muchos de los aspectos más relevantes vinculados al repositorio.

Tal es así, que este trabajo ha versado fuertemente sobre este último punto, y para arribar a él se ha realizado un extenso recorrido de descripción del acceso abierto, que es el objetivo general de una institución al crear el repositorio: compartir lo que hace con equidad y de manera ubicua. Lo importante de detenerse en el acceso abierto ha sido especificar en mayor detalle qué vías existen y, fundamentalmente, qué debe cumplir un repositorio para lograr compartir sus contenidos con comunidades cada día más amplias.

Luego, el presente trabajo ha buscado definir el repositorio institucional,

deslindando sus diferencias con las bibliotecas digitales, las electrónicas y las híbridas. Se ha hablado también de las variantes posibles de repositorios para luego comenzar a ver de qué modo se podría representar, a través de una abstracción, un modelo de repositorio institucional. El capítulo dedicado a los modelos, así, ha corroborado el hecho de que, más allá de las variantes léxicas y los avances de las TIC, desde el año 2000, en que maduró el concepto de repositorio institucional, las principales organizaciones dedicadas a dar las funciones, actores y niveles dentro del repositorio, ya estaban en lo cierto. Por ejemplo, la separación conceptual de las tres esferas (datos, tecnología y usuarios) es fundamental y es, en definitiva y en otras palabras, hablar en términos de la ISO 14721, esto es, el modelo OAIS (2012), donde los datos son el paquete de información, la tecnología es el OAIS con sus seis entidades funcionales y los usuarios son el entorno. El pensar la práctica, por otra parte, en comunidades de investigación y bibliotecarios es estar anticipando la esfera multidisciplinar del repositorio, donde conviven el sistema con distintas comunidades de usuarios con permisos diferentes y roles distintos.

En cuanto al modelo OAIS huelga decir que su red de procesos da comienzo con un productor de contenido que ingresa una obra, la cual es transformada en el interior del repositorio con numerosas operaciones técnicas, como el trabajo administrativo de agregado de metadatos de diversos esquemas, trazabilidad, operaciones de chequeo, agregado de identificadores unívocos, replicación, medidas de seguridad y muchas más, que reflejan el modelo esperado para un repositorio capaz de preservar los contenidos y de dar acceso a los mismos sin restricciones; ese mismo paquete de información queda dotado, en las operaciones propuestas por el modelo, de las condiciones necesarias para su entrega a un usuario persona o sistema.

Tras elegido un modelo y una forma de evaluación, un aspecto, o mejor, un plano de evaluación vinculado a los contenidos y la posibilidad de legibilidad de los mismos a lo largo del tiempo, es posible ver que este plano inicial de evaluación es mucho más que eso. La evaluación del paquete de información es una evaluación compleja que involucra al contenido, su formato y muchos metadatos, y en este proceso no sólo hay usuarios y contenidos a preservar, sino que el paquete de información atraviesa el repositorio y es el sistema que sustenta al OAIS, el que realiza el empaquetado, el que agrega un identificador persistente, el que se asegura de que el contenido no ha sido

alterado, el que guarda la trazabilidad, los eventos y los agentes. Todo esto apunta a que aquella evaluación que parecía estar dedicada, en lo abstracto, tan sólo a verificar la accesibilidad de los contenidos, en realidad también chequea la correcta realización del flujo de trabajo y muchas operaciones no evidentes del sistema que sustenta el repositorio. Así, la evaluación, cuyo modelo propuesto es novedoso, ha recorrido, en el orden del experimento, las tres dimensiones de sistema, datos y usuarios.

La tarea de estudio e investigación demandó buena parte del recorrido de esta tesis, así como la lectura de extensísima bibliografía a lo largo de más de veinte años de trabajos en el área de las bibliotecas y los repositorios digitales. La experimentación en este trabajo ha sido extensa y ha involucrado a muchas personas dentro del repositorio, como se ha intentado plasmar en el capítulo precedente, dedicado al experimento en sí.

Se ha evaluado, entonces, el repositorio SEDICI de la Universidad Nacional de La Plata completo, en todos sus contenidos, revisando formatos y versiones, las cuales han sido contrastadas con registros internacionales, se han generado tareas (más exactamente, 99 tickets en el gestor de incidencias de SEDICI), afectando colecciones de revistas, artículos, tesis, de los cuales se han resuelto ya 89 y sólo quedan pendientes 10, por ser éstos de los más complejos de resolver por diversas causas (entre ellas, la cantidad de ítems con problemas a corregir).

La visión integral de los contenidos ha permitido detectar ítems duplicados, ítems de los cuales sólo se tenía el registro sin tan siquiera una localización física o electrónica, ítems sin ningún resumen que diera idea de su contenido, ítems con problemas debido a la exportación desde el software que sustentaba SEDICI (Celsius-DL) antes de su migración a DSpace, ítems con formatos antiguos que deberán ser sometidos a migración, lo que ha llevado a realizar *scripts* que automaticen la prioridad en la atención de los archivos a migrar, y se han pensado diversos criterios que guíen el ordenamiento. Se han descubierto, paralelamente, ítems sin licencia o con problemas en la licencia. Toda esta tarea, ya extensamente descripta en el capítulo 5, ha resultado muy movilizadora para todo el equipo humano del repositorio y ha mejorado sustancialmente la calidad de los datos del repositorio.

Tras detallados análisis de las tipologías documentales existentes en el repositorio y los formatos, desde el punto de vista de la preservación y la legibilidad, se han indicado

procedimientos a realizar tanto en los procesos de digitalización de materiales cuanto en los relativos a los materiales nacidos digitales, indicando recomendaciones tanto para los formatos de preservación cuanto para los de visualización, de modo de asegurar la mejor legibilidad de los contenidos. Se ha marcado la conveniencia de los formatos abiertos y de la necesidad de formar a los usuarios hacia prácticas que aseguren el acceso a las obras. Se han estudiado estándares de formatos para preservación; en algunos casos, como PDF, se ha hecho un extenso análisis de las versiones y las normas que aseguran la preservación y la accesibilidad, como es el caso del estándar PDF/A propuesto.

El perfilamiento también demostró que no existían contenidos en formatos PDF/A en el repositorio. Si bien en las tareas de digitalización se obtenía un PDF/A, al hacer la edición del archivo por parte de la administración con Acrobat (por ejemplo, para agregar una carátula o para corregir el archivo obtenido tras el OCR), se perdía el formato PDF/A. Como resultado de esto, el repositorio no tiene contenidos en formatos aptos para la preservación a largo plazo. De allí que se plantee la necesidad de encontrar metodologías que, por ejemplo, se adosen al flujo de trabajo para automatizar lo más posible estos procedimientos o bien de instruir a los administradores del repositorio para que utilicen otros programas de edición a la hora de realizar correcciones en los archivos en PDF/A obtenidos tras la digitalización, para que no se pierda esta característica fundamental para la preservación.

Se han verificado también los metadatos presentes en los ítems del repositorio de modo de asegurar la existencia de la totalidad de los metadatos vinculados a la información descriptiva de la preservación y a la información descriptiva. El estado de los ítems en cuanto a sus metadatos se ha averiguado a través de consultas al indexador Solr, a la base de datos e, implementando un validador de un metadato o un conjunto de metadatos, los contrastes propuestos para la información descriptiva siguieron directrices internacionales, en las que están basadas las que se siguen a nivel nacional en el Sistema Nacional de Repositorios del Ministerio de Ciencia, Tecnología e Innovación e Innovación Productiva de Argentina.

Para realizar esta experimentación se han recorrido proyectos y trabajos actuales vinculados específicamente a la evaluación de la preservación y la confiabilidad de los repositorios, dado que desde un modelo de evaluación general y un modelo con un alto

nivel de abstracción como el OAIS (aún con su claro cometido funcional de preservación y accesibilidad) resultaba necesario, antes de entrar en la experimentación, ver otros proyectos y propuestas con resultados concretos.

La experimentación también demandó una investigación en herramientas de acceso abierto capaces de realizar el perfilamiento de los contenidos del repositorio, el chequeo y la validación de formatos, aprender a utilizarlas y ver de qué modo podrían ser incorporadas al repositorio.

Para lograr plasmar el experimento en una implementación en DSpace, como es el caso del repositorio SEDICI, ha sido necesario comprender las entidades y relaciones del modelo de datos de DSpace y las propias de SEDICI-DSpace, la organización del *assetstore*, la estructura de los ítems con sus bundles y bitstreams y los metadatos elegidos en el repositorio.

El estudio y la experimentación realizados han arrojado luz sobre las prácticas, pero también sobre los errores, los procesos de trabajo a mejorar, la necesidad de utilizar herramientas adicionales a DSpace, la necesidad de chequeos y validaciones constantes de los metadatos, la necesidad última de elaborar un plan de preservación completo para el repositorio que no ataque los problemas ya existentes sino que se anticipe a ellos.

Algo de esto se ha dicho ya en el capítulo 5, pero a continuación se expondrá todo lo que resta realizar no ya de manera individual sino en un trabajo en equipo, que dará lugar a nuevos aprendizajes y, sobre todo, a una mejora en la calidad de los contenidos y los servicios ofrecidos por el repositorio central de la UNLP.

A lo largo del devenir de esta tesis se han publicado numerosos trabajos, dictado conferencias y talleres para compartir y contrastar el punto de vista aquí expuesto. Todos los trabajos se adjuntarán con esta tesis.

Trabajos a futuro

La realización del experimento generó distintos cursos de acción a futuro. El primero es la propuesta de utilización de estándares como PDF/A, en su forma A1-a. El estándar -a- significa que el PDF tiene una estructura de modo tal que permite la accesibilidad aún en el caso de personas con capacidades diferentes, ya que al tener

dicha estructura, un lector para una persona con capacidad visual disminuida, por ejemplo, puede reconocer el orden de la lectura, aún en el caso de que el archivo tenga, por ejemplo, columnas que dificulten su comprensión.

Si bien en el momento actual no es posible migrar todos los contenidos a este formato, se propone una metodología de acceso a un estándar -b-, sin perder de vista el objetivo más ambicioso, que excede la propuesta inicial de legibilidad y accesibilidad en el tiempo por la de accesibilidad en un sentido más universal, para todo tipo de persona, independientemente de sus capacidades.

Una cuestión subyacente, más allá del trabajo puntual de migración de los ítems de manera individual, es determinar con qué herramienta se realizará, de ser posible, la migración masiva de modo de no perder la calidad en el archivo resultante ni aumentar de manera exagerada el tamaño de los archivos. Se ha realizado una propuesta preliminar, utilizando una herramienta libre para migración masiva de los PDF, pero es necesario considerar qué significaría para la administración del repositorio el ingreso masivo de ítems que debieran someterse a esta verificación. Este tipo de acciones pueden parecer excelentes desde el punto de vista tecnológico, pero se debe tener en cuenta que el repositorio debe seguir incorporando nuevos materiales y que hay que ser cuidadosos en cuanto a que ocurran errores por este tipo de acciones, que podrían afectar a un porcentaje enorme de sus ítems.

Otro objetivo a futuro está vinculado a la extensión de la herramienta de validación, pues se pretende avanzar en el diseño, la descripción y la implementación de un mecanismo que permita a la administración de un repositorio realizar distintos tipos de controles sobre sus recursos, atendiendo al cumplimiento de una o varias condiciones preestablecidas en la política de preservación. Por ejemplo, se podría pensar, en un nivel de abstracción mayor, en testear los cinco elementos constitutivos de la PDI del modelo OAIS, de acuerdo a lo presentado en el capítulo 5. Esto sería muy importante para el repositorio, ya que brindaría una herramienta a la administración que automatizaría la realización de controles sobre los contenidos y la misma herramienta sería capaz de generar un reporte comprensible por parte de la administración.

Un objetivo futuro aún más ambicioso es pensar en la introducción en el flujo de trabajo de otras herramientas agregadas a DSpace que permitan el chequeo, la

verificación y la validación de los formatos. Por ejemplo, en las figuras que siguen se muestra la propuesta de introducir una herramienta de extracción de características como JHOVE2 o Xena en los procesos de trabajos. La figura 6.1 muestra cómo trabajaría JHOVE2 en los procesos de ingreso de contenidos (SIP), particularmente en la operación de autoarchivo. La propuesta es que al momento de ingresar el SIP se realice la operación de desempaquetado para separar contenido de metadatos y que la herramienta de extracción de características realice la caracterización del objeto digital, proceso en el cual este desarrollo determina las propiedades significativas de un objeto digital que está en un formato dado e incluso ese objeto digital puede ser complejo, es decir no un simple archivo digital, sino uno compuesto por varios archivos.

El proceso de caracterización es un excelente punto de partida para el proceso de preservación digital (Abrams *et al.*, 2009), muy especialmente en los casos en los cuales el objetivo digital es pasible de sufrir transformaciones que pudieran alterarlo, como puede ser el caso de una migración como la que se exhibe en la figura 6.2. En un proceso de migración ya no se habla de SIP sino de un AIP que se encuentra en el propio repositorio.

Es interesante observar que en la figura 7.1 también se muestra un proceso de validación en el cual las características extraídas se contrastan con las reglas de preservación que se han establecido el repositorio digital. Del mismo modo, en la figura 7.2 se muestra que, tras la operación de desempaquetado del AIP, se evalúa si es necesario el proceso de migración en casos como este, mientras que el nuevo objeto de contenido (contenido') pasa al proceso de ingesta, y los nuevos metadatos deben ser validados al igual que en el caso del SIP.

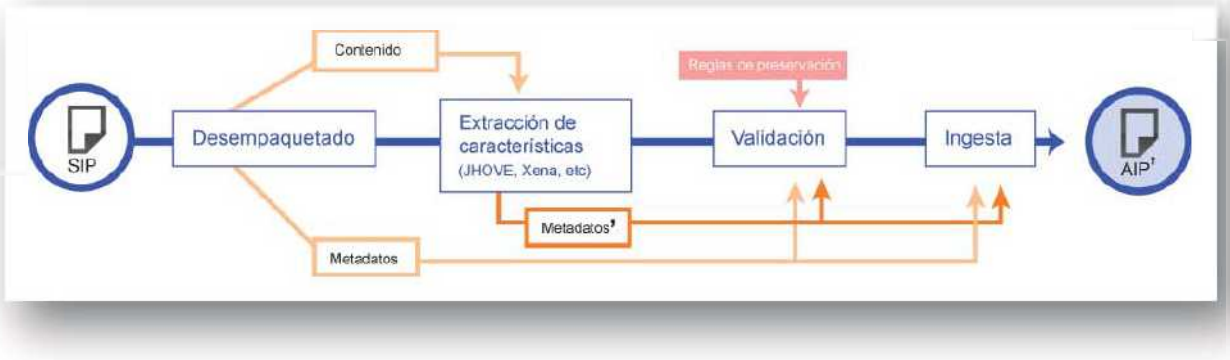


Figura 7.1: Intervención de una herramienta de extracción de metadatos en la ingesta

Fuente propia.

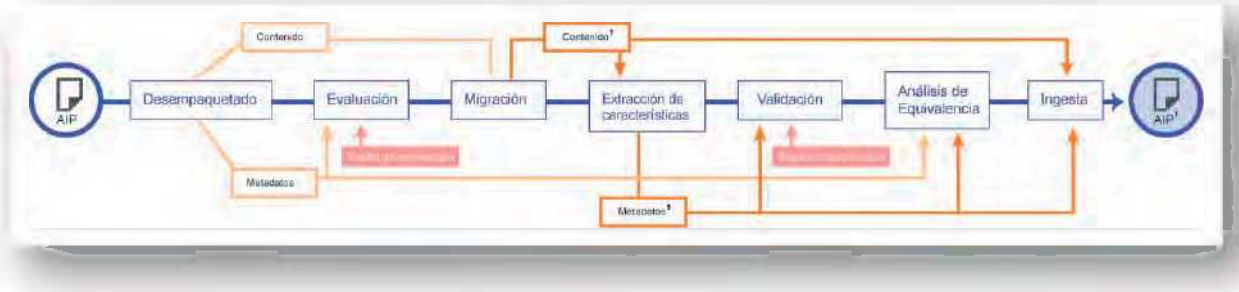


Figura 7.2: Intervención de una herramienta de extracción de metadatos en la migración

Fuente propia.

Durante la experimentación se realizó una primera aproximación a la herramienta Plato que permite generar y compartir planes de preservación; dada la complejidad (y la potencialidad de la herramienta) se invirtió bastante tiempo para leer bibliografía vinculada al planeamiento sistemático de la preservación, sus estrategias y, particularmente, la generación de planes de preservación (Strodl *et al.*, 2007; Becker *et al.*, 2008; Becker *et al.*, 2009; Becker, Rauber, 2011). Tras estas primeras inmersiones en el tema, se continuó el análisis de la herramienta, que es un sistema de soporte, un servicio al cual se accede de manera libre y gratuita y en el que es posible preparar un plan de preservación. En el propio sitio de Plato pueden verse las posibilidades y el flujo de trabajo, tal cual se reproduce en la figura 7.3.

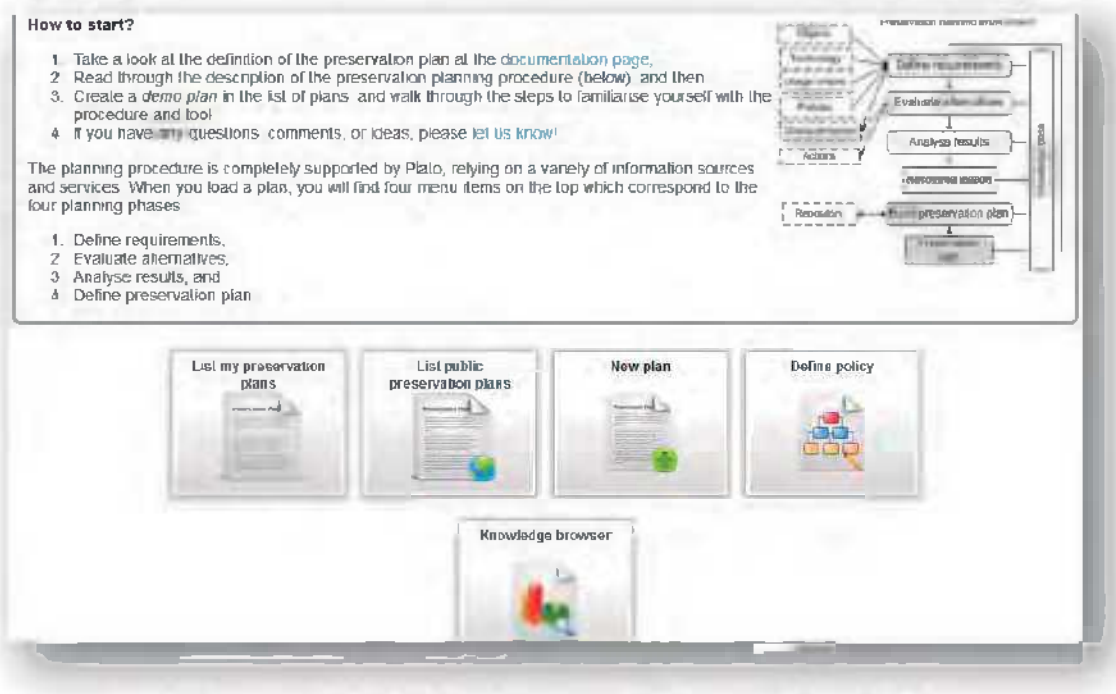


Figura 7.3: Flujo de trabajo y facilidades que ofrece Plato

Fuente: Plato.

Como se ve, se puede obtener una página de documentación muy completa, que muestra un conjunto de importantes trabajos dedicados a definir los pasos y los criterios para el diseño de un plan de preservación. En el mismo sitio, se ofrece la posibilidad de ver planes de preservación públicos, realizados por instituciones, como el plan de preservación de la colección de periódicos de la British Library; incluso es posible abrir esos planes y estudiar cómo se armaron, y hasta hay casos documentados de migraciones masivas a PDF/A. Cada usuario dispone de un espacio propio para generar su plan de preservación y lo hace público al momento en que lo considera oportuno.

Como puede observarse en la figura 7.4, ya se han realizado algunas pruebas para cargar planes y ver los pasos y qué resultados ofrece la herramienta. Un plan de preservación para un repositorio de la complejidad de SEDICI demandará seguramente un tiempo sustancial de aprendizaje e iteraciones. Es necesario remarcar que este es un desafío a largo plazo, de hecho para toda la vida del repositorio institucional y que, por supuesto, abarca mucho más que la vigilancia de formatos y engloba aspectos políticos a definir por la propia Universidad Nacional de La Plata en cuanto a los tiempos de

preservación y los recursos humanos para realizar tales tareas.

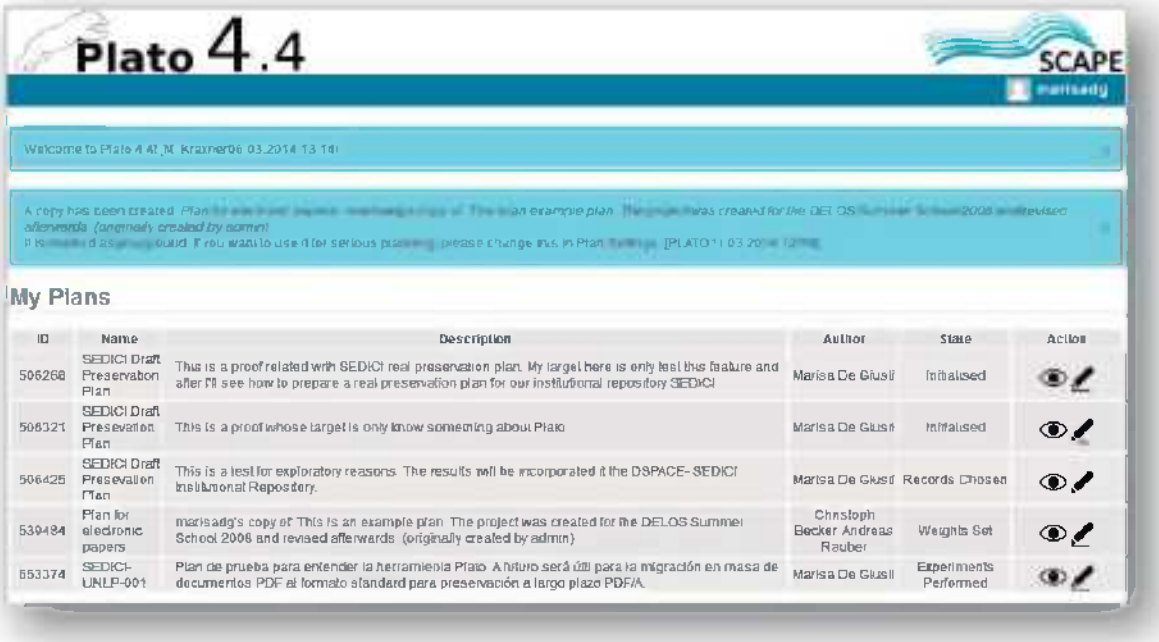


Figura 7.4: Planes de preservación en etapa de inicio del repositorio SEDICI

Fuente propia.

Para finalizar, sólo resta decir que si bien se ha logrado bastante y el trabajo ha sido muy gratificante, muchos más son los desafíos por venir. Por lo pronto, el repositorio SEDICI está en un proceso de migración a DSpace 4, y se espera que con esta herramienta puedan agregarse de manera sencilla cuestiones que son de gran importancia y que se han mencionado ya al hablar de la información descriptiva de la preservación, como las vinculadas a la trazabilidad de cambios y agentes; también se espera que las actividades propuestas a futuro, referidas a las acciones de validación, puedan llevarse adelante y su realización no consuma mucho tiempo de cómputo. El objetivo más ambicioso, desde luego, es el plan de preservación, el cual, más allá de la herramienta con que se genere, deberá integrarse al propio repositorio, modificando los flujos de trabajo y con el advenimiento de herramientas adicionales de verificación, validación y extracción de características. El camino es largo y significará la necesidad de un compromiso grupal e institucional para llevar estas propuestas a hechos concretos.

Bibliografía del capítulo

Abrams, S.; Morrissey, S.; Cramer, T. (2009). "What? So What": The Next-Generation JHOVE2 architecture for Format-Aware Characterization. *The International Journal of Digital Curation*, 3 (4).

Becker, C.; Kulovits, H.; Guttenbrunner, M.; Strodl, S.; Rauber, A.; Hofman, H. (2009). Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries* 10. DOI 10.1007/s00799-009-0057-1.

Becker C.; Rauber, A. (2011). "Preservation Decisions: Terms and Conditions Apply Challenges, Misperceptions and Lessons Learned in Preservation Planning". *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* JCDL'11, June 13-17, Ottawa, Ontario.

Becker, C.; Kulovitis, H.; Rauber, A.; Hofman, H. (2008). "Plato: A Service Oriented Decision Support System for Preservation Planning". *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries* JCDL'08, June 16-20, Pittsburgh, Pennsylvania.

Strodl S.; Becker Ch.; Neumayer R.; Rauber A. (2007). How to Choose a Digital Preservation Strategy: evaluating a Preservation Planning Procedure. *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, p. 29-38.

ANEXOS

PRONOM: Detailed Report

File Format Information: Acrobat PDF 1.6 - Portable Document Format 1.6

Summary

Name	Acrobat PDF 1.6 - Portable Document Format
Version	1.6
Other names	PDF (1.6)
Identifier(s)	MIME: application/pdf Apple Uniform Type Identifier: com.adobe.pdf PUID: fmt/20
Family	
Classification	Page Description
Disclosure	Full
Description	Portable Document Format is a platform-independent format for representing formatted documents, developed by Adobe Systems Incorporated. It is the native format of Adobe's Acrobat family of software products, version 1.6 corresponding to the release of Acrobat 7.0. PDF is based on, and shares the same imaging model as, the PostScript page description language. A PDF file comprises a Header section, a Body section containing the objects which make up the document, a Cross Reference Table, and a Trailer section. PDF files can contain a wide variety of content, including text, images, video and audio.
Orientation	Binary
Byte orders	Big-endian (Motorola)
Related file formats	Has lower priority than Acrobat PDF/A - Portable Document Format (1b) Has lower priority than Acrobat PDF/A - Portable Document Format (2a) Has lower priority than Acrobat PDF/A - Portable Document Format (2b) Has lower priority than Acrobat PDF/A - Portable Document Format (2u) Has lower priority than Acrobat PDF/A - Portable Document Format (3a) Has lower priority than Acrobat PDF/A - Portable Document Format (3b) Has lower priority than Acrobat PDF/A - Portable Document Format (3u) Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4 Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4p Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5g Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5pg Has lower priority than Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5n Has lower priority than Acrobat PDF/E - Portable Document Format for Engineering PDF/E-1 Has priority over Acrobat PDF 1.0 - Portable Document Format (1.0) Has priority over Acrobat PDF 1.1 - Portable Document Format (1.1) Has priority over Acrobat PDF 1.2 - Portable Document Format (1.2) Has priority over Acrobat PDF 1.3 - Portable Document Format (1.3) Has priority over Acrobat PDF 1.4 - Portable Document Format (1.4) Has priority over Acrobat PDF 1.5 - Portable Document Format (1.5) Is subsequent version of Acrobat PDF 1.5 - Portable Document Format (1.5) Is supertype of Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4 Is supertype of Acrobat PDF/X - Portable Document Format - Exchange PDF/X-4p Is supertype of Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5g Is supertype of Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5pg Is supertype of Acrobat PDF/X - Portable Document Format - Exchange PDF/X-5n Is supertype of Acrobat PDF/E - Portable Document Format for Engineering PDF/E-1
Technical Environment	
Released	01 Jan 2004
Supported until	
Developed by	Adobe Systems Incorporated

Supported by	None.
Source	Digital Preservation Department / The National Archives
Source date	11 Mar 2005
Source description	
Last updated	22 Oct 2009
Note	

Documentation

Adobe Systems Incorporated, 2004, PDF Reference, fifth edition: Adobe Portable Document Format, Version 1.6

Type	Authoritative
Availability	Public
Availability note	
Author	Adobe Systems Incorporated
Publication date	14 Nov 2004
Title	PDF Reference, fifth edition: Adobe Portable Document Format, Version 1.6
Publisher	Adobe Systems Incorporated
Identifier(s)	
Rights	
Note	

Signatures

External signatures File extension: pdf

Internal signatures

Name	PDF 1.6
Description	Header and footer
Byte sequences	

Position type	Absolute from BOF
Offset	0
Byte ordering	
Value	255044462D312E36
Position type	Absolute from EOF
Offset	0
Byte ordering	
Value	2525454F46

Compression Types

None.

Character Encoding

None.

Rights

None.

Reference Files

None.

Inherent Properties

Instance Properties

Creation Date	Description
	Risk
Encrypted	Description
	Risk
Number of Pages	Description
	Risk
Number of Images	Description
	Risk
Creating Application	Description
	Risk
Title	Description
	Risk
Creator	Description
	Risk
Subject	Description
	Risk

Comprehensive breakdown

Profile Summary

Name	Signature version	Container version	Started	Finished	Filters
Untitled-1	79	20140922	16 dic 2014	16 dic 2014	

File & folder count

Report field	Grouping fields
FILE_NAME	

Profile	Count	Sum	Min	Max	Average
Untitled-1	19231				
Profile totals	**19231**				

Unreadable files

Report field	Grouping fields
FILE_NAME	

Filter fields: (all filter criteria below must be true)		
Field	**Operator**	**Values**
RESOURCE_TYPE	NONE_OF	"Folder"
JOB_STATUS	ANY_OF	"Not found" "Access denied" "Error"

Profile	Count	Sum	Min	Max	Average
Untitled-1	0				
Profile totals	**0**				

Unreadable folders

Report field	Grouping fields
FILE_NAME	

Filter fields: (all filter criteria below must be true)		
Field	**Operator**	**Values**
RESOURCE_TYPE	ANY_OF	"Folder"
JOB_STATUS	ANY_OF	"Not found" "Access denied" "Error"

Profile	Count	Sum	Min	Max	Average

Untitled-1	0
Profile totals	**0**

File count and sizes

Report field	Grouping fields
FILE_SIZE	

Filter fields:		
Field	**Operator**	**Values**
RESOURCE_TYPE	NONE_OF	"Folder"

Profile	Count	Sum	Min	Max	Average
Untitled-1	18523	26864373276	1120	701906423	1450325
Profile totals	**18523**	**26864373276**	**1120**	**701906423**	**1450325**

File sizes per extension

Report field	Grouping fields
FILE_SIZE	FILE_EXTENSION

Filter fields:		
Field	**Operator**	**Values**
RESOURCE_TYPE	NONE_OF	"Folder"

Profile	Count	Sum	Min	Max	Average
Untitled-1	18522	26863674031	1120	701906423	1450365
Profile totals	**18522**	**26863674031**	**1120**	**701906423**	**1450365**

pdf					
Profile	**Count**	**Sum**	**Min**	**Max**	**Average**
Untitled-1	1	699245	699245	699245	699245
Profile totals	**1**	**699245**	**699245**	**699245**	**699245**

Group totals

Count	Sum	Min	Max	Average
18523	26864373276	1120	701906423	1450324

File sizes per PUID

Report field		Grouping fields		
FILE_SIZE	PUID	FILE_FORMAT	FORMAT_VERSION	MIME_TYPE

Filter fields:

Field	Operator	Values
RESOURCE_TYPE	NONE_OF	"Folder"

Profile	Count	Sum	Min	Max	Average
Untitled-1	16	67175565	64151	12605693	4198472
Profile totals	**16**	**67175565**	**64151**	**12605693**	**4198472**

fmt/108	Macromedia Flash	5	application/x-shockwave-flash		
Profile	Count	Sum	Min	Max	Average
Untitled-1	47	7462353	1120	1629614	158773
Profile totals	**47**	**7462353**	**1120**	**1629614**	**158773**

fmt/116	Windows Bitmap	3.0	image/bmp		
Profile	Count	Sum	Min	Max	Average
Untitled-1	4	203240	50454	51054	50810
Profile totals	**4**	**203240**	**50454**	**51054**	**50810**

fmt/126	Microsoft Powerpoint Presentation	97-2003	application/vnd.ms-powerpoint		
Profile	Count	Sum	Min	Max	Average
Untitled-1	5	21845504	856576	13942784	4369100
Profile totals	**5**	**21845504**	**856576**	**13942784**	**4369100**

fmt/134	MPEG 1/2 Audio Layer 3		audio/mpeg		
Profile	Count	Sum	Min	Max	Average
Untitled-1	583	5526826294	205952	180529426	9479976
Profile totals	**583**	**5526826294**	**205952**	**180529426**	**9479976**

fmt/14	Acrobat PDF 1.0 - Portable Document Format	1.0	application/pdf		
Profile	Count	Sum	Min	Max	Average
Untitled-1	1	1648476	1648476	1648476	1648476
Profile totals	**1**	**1648476**	**1648476**	**1648476**	**1648476**

fmt/15	Acrobat PDF 1.1 - Portable Document Format	1.1	application/pdf

Profile	Count	Sum	Min	Max	Average
Untitled-1	137	85968018	5607	46825322	627503
Profile totals	**137**	**85968018**	**5607**	**46825322**	**627503**

fmt/16	Acrobat PDF 1.2 - Portable Document Format			1.2	application/pdf
Profile	Count	Sum	Min	Max	Average
Untitled-1	1021	326802778	2834	13872547	320081
Profile totals	**1021**	**326802778**	**2834**	**13872547**	**320081**

fmt/17	Acrobat PDF 1.3 - Portable Document Format			1.3	application/pdf
Profile	Count	Sum	Min	Max	Average
Untitled-1	1469	1268927497	3473	149412845	863803
Profile totals	**1469**	**1268927497**	**3473**	**149412845**	**863803**

fmt/18	Acrobat PDF 1.4 - Portable Document Format			1.4	application/pdf
Profile	Count	Sum	Min	Max	Average
Untitled-1	5599	4183655704	7254	548132894	747214
Profile totals	**5599**	**4183655704**	**7254**	**548132894**	**747214**

fmt/19	Acrobat PDF 1.5 - Portable Document Format			1.5	application/pdf
Profile	Count	Sum	Min	Max	Average
Untitled-1	2426	2545224966	6013	86687661	1049144
Profile totals	**2426**	**2545224966**	**6013**	**86687661**	**1049144**

fmt/20	Acrobat PDF 1.6 - Portable Document Format			1.6	application/pdf
Profile	Count	Sum	Min	Max	Average
Untitled-1	6710	10810780213	6509	676724282	1611144
Profile totals	**6710**	**10810780213**	**6509**	**676724282**	**1611144**

fmt/203	Ogg Vorbis Codec Compressed Multimedia File				audio/ogg
Profile	Count	Sum	Min	Max	Average
Untitled-1	7	5767875	751996	874698	823982
Profile totals	**7**	**5767875**	**751996**	**874698**	**823982**

fmt/276	Acrobat PDF 1.7 - Portable Document Format			1.7	application/pdf
Profile	Count	Sum	Min	Max	Average

Untitled-1	301	1877145112	8287	701906423	6236362
Profile totals	**301**	**1877145112**	**8287**	**701906423**	**6236362**

fmt/396	PocketMobi (Palm Resource) File				
Profile	Count	Sum	Min	Max	Average
Untitled-1	1	2381714	2381714	2381714	2381714
Profile totals	**1**	**2381714**	**2381714**	**2381714**	**2381714**

fmt/42	JPEG File Interchange Format			1.00	image/jpeg
Profile	Count	Sum	Min	Max	Average
Untitled-1	2	1451804	402703	1049101	725902
Profile totals	**2**	**1451804**	**402703**	**1049101**	**725902**

fmt/43	JPEG File Interchange Format			1.01	image/jpeg
Profile	Count	Sum	Min	Max	Average
Untitled-1	32	14127805	18030	3144896	441493
Profile totals	**32**	**14127805**	**18030**	**3144896**	**441493**

fmt/44	JPEG File Interchange Format			1.02	image/jpeg
Profile	Count	Sum	Min	Max	Average
Untitled-1	154	64492202	45710	8642693	418780
Profile totals	**154**	**64492202**	**45710**	**8642693**	**418780**

fmt/483	ePub format		application/epub+zip		
Profile	Count	Sum	Min	Max	Average
Untitled-1	1	6519092	6519092	6519092	6519092
Profile totals	**1**	**6519092**	**6519092**	**6519092**	**6519092**

fmt/5	Audio/Video Interleaved Format			video/x-msvideo	
Profile	Count	Sum	Min	Max	Average
Untitled-1	4	39275188	6647670	16685412	9818797
Profile totals	**4**	**39275188**	**6647670**	**16685412**	**9818797**

fmt/561	Adobe Illustrator				12.0
Profile	Count	Sum	Min	Max	Average
Untitled-1	1	6532132	6532132	6532132	6532132

Profile totals	1	6532132	6532132	6532132	6532132

fmt/61	Microsoft Excel 97 Workbook (xls)	8	application/vnd.ms-excel		
Profile	Count	Sum	Min	Max	Average
Untitled-1	2	159744	79872	79872	79872
Profile totals	**2**	**159744**	**79872**	**79872**	**79872**

Group totals

Count	Sum	Min	Max	Average
18523	26864373276	1120	701906423	1450325

File sizes per MIME type

Report field	Grouping fields
FILE_SIZE	MIME_TYPE

Filter fields:

Field	Operator	Values
RESOURCE_TYPE	NONE_OF	"Folder"

Profile	Count	Sum	Min	Max	Average
Untitled-1	18	76089411	64151	12605693	4227189
Profile totals	**18**	**76089411**	**64151**	**12605693**	**4227189**

application/epub+zip

Profile	Count	Sum	Min	Max	Average
Untitled-1	1	6519092	6519092	6519092	6519092
Profile totals	**1**	**6519092**	**6519092**	**6519092**	**6519092**

application/pdf

Profile	Count	Sum	Min	Max	Average
Untitled-1	17664	21100152764	2834	701906423	1194528
Profile totals	**17664**	**21100152764**	**2834**	**701906423**	**1194528**

application/vnd.ms-excel

Profile	Count	Sum	Min	Max	Average
Untitled-1	2	159744	79872	79872	79872

Profile totals	2	159744	79872	79872	79872

application/vnd.ms-powerpoint					
Profile	Count	Sum	Min	Max	Average
Untitled-1	5	21845504	856576	13942784	4369100
Profile totals	**5**	**21845504**	**856576**	**13942784**	**4369100**

application/x-shockwave-flash					
Profile	Count	Sum	Min	Max	Average
Untitled-1	47	7462353	1120	1629614	158773
Profile totals	**47**	**7462353**	**1120**	**1629614**	**158773**

audio/mpeg					
Profile	Count	Sum	Min	Max	Average
Untitled-1	583	5526826294	205952	180529426	9479976
Profile totals	**583**	**5526826294**	**205952**	**180529426**	**9479976**

audio/ogg					
Profile	Count	Sum	Min	Max	Average
Untitled-1	7	5767875	751996	874698	823982
Profile totals	**7**	**5767875**	**751996**	**874698**	**823982**

image/[illegible]					
Profile	Count	Sum	Min	Max	Average
Untitled-1	4	203240	50454	51054	50810
Profile totals	**4**	**203240**	**50454**	**51054**	**50810**

image/jpeg					
Profile	Count	Sum	Min	Max	Average
Untitled-1	188	80071811	18030	8642693	425913
Profile totals	**188**	**80071811**	**18030**	**8642693**	**425913**

video/x-msvideo					
Profile	Count	Sum	Min	Max	Average
Untitled-1	4	39275188	6647670	16685412	9818797
Profile totals	**4**	**39275188**	**6647670**	**16685412**	**9818797**

Group totals

Count	Sum	Min	Max	Average
18523	26864373276	1120	701906423	1450325

File count and sizes per year last modified

Report field	Grouping fields
FILE_SIZE	Year(LAST_MODIFIED_DATE)

Filter fields:

Field	Operator	Values
RESOURCE_TYPE	NONE_OF	"Folder"

2013

Profile	Count	Sum	Min	Max	Average
Untitled-1	18522	26863674031	1120	701906423	1450365
Profile totals	**18522**	**26863674031**	**1120**	**701906423**	**1450365**

2014

Profile	Count	Sum	Min	Max	Average
Untitled-1	1	699245	699245	699245	699245
Profile totals	**1**	**699245**	**699245**	**699245**	**699245**

Group totals

Count	Sum	Min	Max	Average
18523	26864373276	1120	701906423	1450324

File count and sizes per year and month last modified

Report field	Grouping fields	
FILE_SIZE	Year(LAST_MODIFIED_DATE)	Month(LAST_MODIFIED_DATE)

Filter fields:

Field	Operator	Values
RESOURCE_TYPE	NONE_OF	"Folder"

2013 12

Profile	Count	Sum	Min	Max	Average

Untitled-1	18522	26863674031	1120	701906423	1450365
Profile totals	**18522**	**26863674031**	**1120**	**701906423**	**1450365**

2014					4
Profile	Count	Sum	Min	Max	Average
Untitled-1	1	699245	699245	699245	699245
Profile totals	**1**	**699245**	**699245**	**699245**	**699245**

Group totals

Count	Sum	Min	Max	Average
18523	26864373276	1120	701906423	1450324

File count and sizes by month last modified

Report field	Grouping fields
FILE_SIZE	Month(LAST_MODIFIED_DATE)

Filter fields:		
Field	Operator	Values
RESOURCE_TYPE	NONE_OF	"Folder"

4					
Profile	Count	Sum	Min	Max	Average
Untitled-1	1	699245	699245	699245	699245
Profile totals	**1**	**699245**	**699245**	**699245**	**699245**

12					
Profile	Count	Sum	Min	Max	Average
Untitled-1	18522	26863674031	1120	701906423	1450365
Profile totals	**18522**	**26863674031**	**1120**	**701906423**	**1450365**

Group totals

Count	Sum	Min	Max	Average
18523	26864373276	1120	701906423	1450324

Nombre amigable	[illegible] en [illegible]	standard + proximo	Dominio/Auth	Tipo	Oblig.	Auto Archivo	Workflow
Localizacion Electrónica	sedici.identifier.uri	dc.identifier	URI	*	MA	-	O
HANDLE (DSPACE)	dc.identifier.uri	dc.identifier	URI	*	S	-	-
DOI	sedici.identifier.doi	dc.identifier	URI	*	MA	-	O
Handle	**sedici.identifier.handle**	dc.identifier	URI	*	MA	-	O
ISBN	sedici.identifier.isbn	dc.identifier	string	libros, tesis, CO	MA	O2	M(libro)
ISSN	sedici.identifier.issn	dc.identifier	Journal?	CO, ART	MA	O	O
Nro Expediente	sedici.identifier.expediente	dc.identifier		Inst			
Otro identificador	sedici.identifier.other	dc.identifier	URI	*	R	-	O
Título	dc.title	dc.title	string	*	M	M1	R
Título Alternativo	dc.title.alternative	dcterms.alternative	string	*	R	-	O
Subtitulo	sedici.title.subtitle		string	*	MA	-	O
Autor	sedici.creator.person	dc.creator	Autores-s2003?	*	MA	M1	R(!ADMIN)
Autor Institucional	sedici.creator.corporate	dc.creator	Instituciones-s2003?	*	MA	-	O
Fecha de publicación	dc.date.issued	dcterms.issued	W3C-DTF	*	MA	M1	O
Fecha de disponibilidad en Dspace	dc.date.available	dcterms.available	W3C-DTF	*	-	-	-
Fecha de creacion en Dspace	dc.date.accessioned		W3C-DTF	*	-	-	-
Fecha de creación	dc.date.created	dcterms.created	W3C-DTF	Objeto Fisico	-	-	MA

Fecha de Presentación	sedici.date.exposure		W3C-DTF	tesis, objeto conferencia, audio	R	O	O
Periodo de Embargo	sedici.embargo.period	-	VP(1,3,6,12,18,24)	*	MA	-	O
Fecha de fin de embargo	**sedici.embargo.liftDate**		W3C-DTF	*	O	-	-
Colaborador	sedici.contributor.colaborator	dc.contributor	string	*	R	O	O
Traductor	sedici.contributor.translator	dc.creator	string	libro	MA	-	O
Compilador	sedici.contributor.compiler	dc.creator	Autores-s2003?	libro	MA	-	O
Director	sedici.contributor.director	dc.contributor	Autores-s2003?	tesis	MA	M2	M
Codirector de tesis	sedici.contributor.codirector	dc.contributor	Autores-s2003?	tesis	R	-	O
Jurado	sedici.contributor.juror	dc.contributor	string	tesis	O	-	O
Firmante	sedici.contributor.inscriber	dc.creator		Inst			
Editor	sedici.contributor.editor	dc.creator	string	libro	R	-	O
Editorial	dc.publisher	dc.publisher	string	libro, tesis	MA	-	O
Alcance Geografico	dc.coverage.spatial	dcterms.spatial	ISO3166? + TGN?	*	R	-	O
AlcanceTemporal	dc.coverage.temporal	dcterms.temporal	string	*	R	-	O
Idioma	dc.language	dc.language	VP-ISO639-1!	*	M	-	M
Extensión	dc.format.extent	dcterms.extent	string or hh:mm:ss	*	O	-	O
Materiales	dc.format.medium	dcterms.medium	string	Objeto Fisico	-	-	MA
Materia (Sedici old)	sedici.subject.materias	dc.subject	Materia-s2003!	*	M	-	M
Materia (LCSH)	sedici.subject.lcsh	dc.subject	LCHS!	*	M*	-	M

	Descriptores Decs	sedici.subject.decs	dc.subject	DECS-s2003!	*	R	-	O
	Descriptores Eurovoc	sedici.subject.eurovoc	dc.subject	Eurovoc-s2003!	*	R	-	O
	Descriptores Libres	sedici.subject.other	dc.subject	string	*	R	-	O
	Palabras Clave	sedici.subject.keyword	dc.subject	mega dictionary?	*	R	O2	O
	Descriptores ACM	sedici.subject.acmcss98	dc.subject	ACM-CSS98	*	R	-	O
	Notas	sedici.description.note	dc.description	string	*	O	O2	O
	Resumen	dc.description.abstract	dcterms.abstract	string	*	MA	O2	O
	Fulltext	sedici.description.fulltext	-	boolean	*	M	-	M
##	Evaluación	sedici.description.peerReview	eprints.status	VP(PeerReviewed, NonPeerReviewed)	CO, ART	R	-	O
	Registro de Origen	dc.description.provenance	dcterms.provenance	string	*	-	-	-
	Lugar de desarrollo	sedici.institucionDesarrollo	-	Insittuciones-s2003?	tesis	O	-	O
	Grado Alcanzado	thesis.degree.name	thesis.degree.name	Grados-s2003?	Tesis	MA	M	M
	Institucion Garante	thesis.degree.grantor	thesis.degree.grantor	Insittuciones-s2003?	Tesis	MA	M2	M
	Entidad de Origen	mods.originInfo.place	mods.originInfo.place	PIDU-s2003?	*	M	-	M
	Tipo de Documento	dc.type	dc.type	VP(Type)	*	M	M	M
	SubTipo de Documento	sedici.subtype	dc.type	VP(SubType)	*	M	M	M
	Licencia	sedici.rights.license	dcterms.license	VP(Licencias)	*	M	M	-
	URI de la Licencia	sedici.rights.uri	dc.rights	URL	*	R	-	-

Localizacion Fisica	mods.location	mods.location	string OR URL	*	R	-	O
Nombre del Evento	sedici.relation.event	dcterms.isPartOf	string	CO	MA	M	M
Titulo de la serie	sedici.relation.journalTitle	dcterms.isPartOf	-	ART, BookCh	MA	M	M
Número de la serie	sedici.relation.journalVolumeAndIssue	dcterms.isPartOf	-	ART, BookCh	MA	M	M
Nombre del Dossier	sedici.relation.dossier	dcterms.isPartOf	string	ART	R	-	O
Fuente	mods.recordInfo.recordContentSource	mods.recordInfo.recordContentSource	VP	*	MA	-	O
Revision de	sedici.relation.isReviewOf	dc.relation	URL	ART	O	-	O
Revisado por	sedici.relation.isReviewedBy	dc.relation	URL	*	O	-	O
Relacionado con	sedici.relation.isRelatedWith	dc.relation	URL	*	O	-	O
Título del Libro	sedici.relation.bookTitle	dcterms.isPartOf	String	BookCh	O	-	O
Titulo de la serie	sedici.relation.isPartOfSeries	dcterms.isPartOf	String	Book	O	O	O