Revista Española de Documentación Científica,
34, 3, julio-septiembre, 385-407, 2011
ISSN: 0210-0614. doi: 10.3989/redc.2011.3.834

La comunidad «Recursos y datos primarios» de la Universitat Pompeu Fabra: los repositorios institucionales como infraestructuras científicas: estudio de caso

Silvia Arano*, Gemma Martínez*, Marina Losada**, Marta Villegas*, Anna Casaldàliga** y Núria Bel*.

Resumen: El artículo presenta una primera aproximación a la publicación en acceso abierto de datos resultantes de la investigación en el área de humanidades. Describe el estudio de caso implementado en el repositorio institucional de la Universitat Pompeu Fabra, a partir de la creación de dos colecciones con datos, una para los anexos presentes en las tesis y otra para las herramientas y recursos lingüísticos. Finalmente se analiza la experiencia realizada y se extraen algunas conclusiones respecto a la utilización de los repositorios como infraestructuras.

Palabras clave: datos de investigación, datos primarios, repositorios, infraestructuras científicas, humanidades, lingüística aplicada, Universitat Pompeu Fabra.

The community «resources and primary data» of the Universitat Pompeu Fabra: institucional repositories as scientific infrastructures: a case study

Abstract: *This article presents an initial approach to open access publication of research data from the humanities. It is a case study of the creation of two data collections at the institutional repository of the Universitat Pompeu Fabra, one containing the annexes of doctoral theses and the other, language tools and resources. Finally, we analyze the experience and draw conclusions regarding the use of repositories as infrastructure.*

Keywords: *research data, repositories, scientific infrastructure, humanities, applied linguistics, Universitat Pompeu Fabra.*

«The coolest thing to do with your data will be thought of by someone else...»

(Crig, 2008)

* Institut Universitari de Lingüística Aplicada.

** Biblioteca UPF, Universitat Pompeu Fabra. Barcelona. Correo-e: silvia.arano@upf.edu; gemma.martinez@upf.edu; marina.losada@upf.edu; marta.villegas@upf.edu; anna.casaldaliga@upf.edu; nuria.bel@upf.edu.

Recibido: 09-02-2010; 2.ª versión: 04-03-2011; aceptado: 20-03-2011.

1. Introducción

En la actualidad, tanto el quehacer cotidiano como también el de las actividades culturales, económicas, sociales y de investigación se han visto revolucionados por el desarrollo de las tecnologías de comunicación e información. Los recientes avances en materia de digitalización creciente de datos, conjuntamente con la progresiva virtualización de los entornos de trabajo, han cambiado tanto la forma de procesar, acceder y distribuir los datos, como la forma de relacionarse entre las propias personas. Especialmente en el ámbito científico y académico, se está produciendo una migración de los espacios físicos de trabajo hacia los espacios virtuales, donde el concepto de *e-ciencia* se comienza a consolidar y pone de manifiesto la necesidad de generar infraestructuras científicas de soporte adecuadas para trabajar y preservar grandes volúmenes de datos en forma virtual y colaborativa, las llamadas *e-infraestructuras*. Este cambio de modalidad en cuanto a la forma de realizar investigación, también denominado el «cuarto paradigma» (Hey, 2009), no es privativo de las ciencias básicas sino que también involucra a las humanidades y ciencias sociales.

La implementación de una infraestructura científica requiere de un largo proceso de planificación y desarrollo, donde se necesitan gran cantidad de recursos humanos y económicos. Los beneficios más evidentes son para la comunidad académica y científica, ya que permite a diferentes investigadores encontrar y reunir gran cantidad de datos, trabajar sobre datos ajenos resultantes de otra investigación, generar nuevo conocimiento a partir de ellos y, además, facilitar el trabajo en un entorno colaborativo, preservando la integridad y autoría de los datos. De esta manera los datos se usan, se reutilizan y se combinan, incrementando la productividad y la capacidad de correlación a una escala nunca vista con anterioridad. Asimismo, existen otros colectivos beneficiados tales como los proveedores de datos, las entidades públicas de financiación, el sector de innovación empresarial e industrial, el sector político y el público en general. Todos ellos tienen la posibilidad, a través del uso de la infraestructura científica y con diferentes objetivos, de acceder, depositar y utilizar los datos publicados en una infraestructura científica.

En definitiva, una infraestructura científica es un conjunto de instalaciones, medios técnicos y servicios necesarios, tanto para permitir el acceso, uso y reutilización de los datos, como para garantizar la autoría e integridad de los mismos. Esta amplia conceptualización de infraestructura permite considerar como tal, elementos tan diversos como redes, *grids,* aplicaciones y recursos informáticos, bancos de trabajo experimentales, centros de supercomputación, repositorios y comunidades virtuales. Si bien todos los elementos mencionados se pueden entender como subtipos de infraestructuras científicas, no todos tienen las mismas funciones, por ejemplo los *grids* se orientarían más hacia la generación de plataformas colaborativas de investigación, en tanto que los repositorios se constituirían en instalaciones para preservar y difundir en acceso abierto datos e informaciones resultantes de la investigación.

El presente artículo constituye una primera aproximación a la publicación en acceso abierto de datos resultantes de la investigación del área de humanidades, a partir del análisis de la capacidad de un repositorio institucional (RI) como infraestructura científica para publicarlos. A partir de dicha reflexión, se plantea un estudio de caso que tiene como objetivos: a) describir la puesta en marcha de una comunidad para datos primarios y recursos provenientes de la investigación en el RI de la Universitat Pompeu Fabra (e-Repositori: Repositori Digital de la UPF); b) analizar las implicaciones de gestión que tiene dicha implementación, tanto para el personal encargado de la administración del repositorio como para los propios investigadores involucrados.

De acuerdo con los objetivos anteriormente mencionados, el estudio de caso se desarrolla en tres apartados. En el primer apartado, se realiza una breve introducción a la situación de los repositorios institucionales y sus contenidos. En el segundo, se presenta la investigación realizada por el grupo de Tecnologías de los Recursos Lingüísticos (TRL) del Instituto Universitario de Lingüística Aplicada (IULA) en el marco del proyecto europeo CLARIN. En el tercero, se desarrollan las características del repositorio institucional y se describe la implementación de dos nuevas colecciones dedicadas a los datos primarios y recursos resultantes de la investigación. Finalmente se presentan las conclusiones y perspectivas de investigación futuras.

2. Estudio de caso

2.1. Los repositorios institucionales y sus contenidos

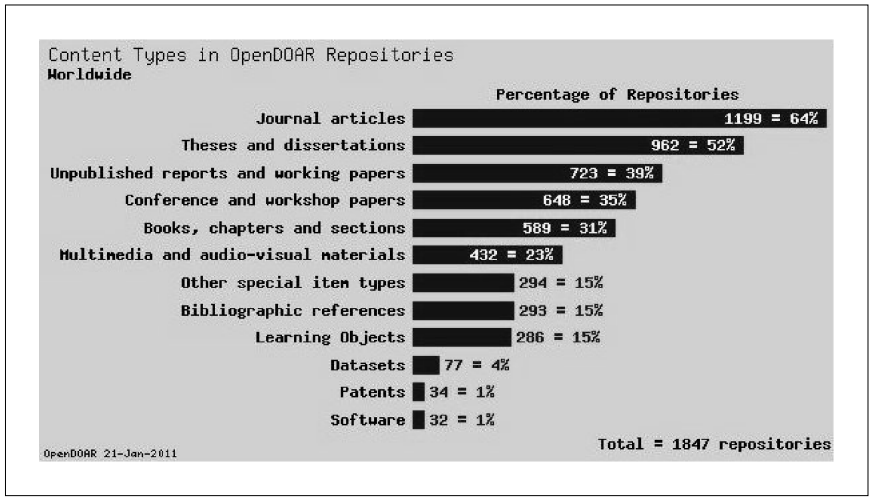
Según datos de directorio OpenDoar [http://www.opendoar.org/index.html], gestionado por la Universidad de Nottingham, a enero del 2011 existen 1847 repositorios registrados, de los cuales 1508 son institucionales, cantidad que representa un 82% de los repositorios existentes a nivel internacional. España no es ajena a esta tendencia internacional, teniendo 55 RIs censados sobre un total de 68 repositorios, lo cual representa un 81% de los repositorios españoles registrados en el directorio OpenDoar.

Si bien en sus orígenes los RIs se planteaban casi como bibliotecas digitales compuestas por una diversa tipología de objetos digitales, su implementación real se basaba fundamentalmente en materiales de producción científica e institucional. Esta tipología gradualmente se ha ampliado para dar cabida a materiales tales como presentaciones a congresos, recursos educativos y docentes, materiales multimedia, programas de ordenador y datos de investigación. Sin embargo, aún existe una clara tendencia hacia el predominio de los objetos digitales de tipo publicación (artículos de revista, tesis, *working papers,* informes de proyectos, etc.), en detrimento de otros formatos menos comunes como materiales multimedia, programas de ordenador, patentes o datos.

A nivel internacional, como se muestra en el siguiente gráfico, entre el 40 y 64% de los RIs tienen en depósito los materiales tradicionalmente asociados con la producción científica, en cambio los que incorporan materiales menos tradicionales rondan entre el 1 y 4%. Cabe mencionar también, que los RIs que incorporan recursos educativos y docentes y materiales multimedia presentan una incidencia media, alcanzando porcentajes entre el 15 y 23%. (figura 1)

FIGURA 1

Content Types in OpenDOAR Repositories - Worldwide



Fuente: OpenDOAR.

La situación en España no escapa a dichas pautas de contenido, teniendo una fuerte presencia los contenidos digitales tales como artículos y tesis doctorales a texto completo, además de *working papers*, presentaciones a congresos y monografías. (figura 2)

También se registran resultados similares en relación con el resto de tipos de documentos menos tradicionales, los cuales presentan volúmenes más bajos de depósito. Lo más significativo es que los objetos digitales tales como programas de ordenador o datos primarios no aparecen diferenciados en las tipologías de materiales depositados, aunque posiblemente dada su baja incidencia numérica puede ser que hayan sido incluidos en el ítem «Otros».

2.1.1. Los datos de investigación en los RIs

La investigación como actividad intelectual genera una gran cantidad de datos de diversa naturaleza y origen, además de las conocidas publicaciones e informes científicos.

FIGURA 2

Cantidad y tipo de objetos digitales en los repositorios españoles: años 2007-2008

Tipología objetos digitales	OD totales 2007	2008
Artículos (texto + metadatos)	17538	68357
Artículos (sólo metadatos)	100	235
Libros/capítulos de libros (texto + metadatos)	429	1145
Libros/capítulos de libros (sólo metadatos)	2005	3053
Tesis doctorales (texto + metadatos)	3155	8532
Tesis doctorales (sólo metadatos)	260	800
Actas de congresos (texto + metadatos)	444	2080
Actas de congresos (metadatos)	0	332
Documentos de trabajo "working papers" (texto + metadatos)	1306	3729
Documentos de trabajo "working papers" (sólo metadatos)	0	270
Material docente	1725	5889
Hojas de datos	-	-
Imágenes	1129457	4814493
Vídeos	283	2789
Música	-	-
Otros	2153	9439

Fuente: Melero, R. y otros (2009). Situación de los RIs en España: informe 2009.

Dicha heterogeneidad reúne bajo la común denominación de datos de investigación (RIN, 2008), o simplemente datos o datos primarios, materiales tan dispares como:

- Experimentos científicos.
- Modelos y simulaciones (donde se incluyen tanto el modelo con sus metadatos asociados, como los datos generados por el propio modelo).
- Observaciones (donde los datos usualmente constituyen un registro único e irremplazable, como pueden ser formularios, censos, registros de votos, etc.).
- Datos derivados (resultantes del proceso o combinación de datos sin procesar (*raw data*) con otros tipos de datos).
- Datos referenciales (p.e. secuencias genómicas, estructuras químicas, etc.).
- Material complementario (también de índole diversa, entendiéndose como parte del paquete informativo de los datos de investigación. P.e.: instrucciones de codificación, guías para entrevistadores, gráficos de los datos recolectados, instrumentos para la recolección de datos, diccionarios, etc. (Green, MacDonald y Rice, 2009).

De esta amplia gama de materiales, aquellos que no son textuales y escritos, como por ejemplo estructuras químicas, programas de ordenador o bases de datos tienen una incipiente incorporación al contenido de los RIs.

Específicamente, la incorporación de los datos de investigación en infraestructuras científicas de acceso abierto ha tenido una fuerte expansión en los últimos años, gracias al desarrollo de iniciativas oficiales y proyectos emprendidos en diversos países con la participación de organismos gubernamentales y universidades.

En el Reino Unido, proyectos tales como *eBank UK* (2003-2007) y *DISC-UK DataShare* (2007-2009), lograron poner de manifiesto la necesidad de disponer de datos primarios en plataformas de acceso abierto, de relacionar dichos datos con otros productos generados durante el proceso de investigación y de promover buenas prácticas comunes en diferentes instituciones (Rice, 2007). En definitiva, hicieron aflorar la necesidad de aunar esfuerzos y experiencias con el fin de optimizar los servicios actuales de los RIs para incorporar los datos de investigación.

A partir del año 2008 en Australia comienzan a consolidarse las iniciativas relacionadas con la gestión de los datos de investigación, con la fundación del *Australian National Data Service (ANDS*) en acuerdo con la *Monash University*. Los proyectos llevados a cabo, tanto por universidades como por entes públicos, tenían la misión común de promover la creación y el fortalecimiento de infraestructuras institucionales para datos de investigación, el desarrollo de servicios específicos por disciplina para nutrir las comunidades nacionales de investigación y la optimización en la recolección, administración y gestión de datos y metadatos.

En Estados Unidos se encuentra uno de los recursos de referencia en la publicación de datos primarios de investigación en acceso abierto, el *Genbank*, base de datos de secuencias genéticas en funcionamiento desde 1982. Actualmente el *Genbank* es gestionado por el *National Institutes of Health* (NIH) y trabaja colaborativamente con el *DNA DataBank of Japan* (DDBJ) y el Laboratorio Europeo de Biología Molecular (*European Molecular Biology Laboratory*, EMBL), constituyendo un recurso de gran prestigio entre los investigadores del área para difundir los resultados primarios de sus investigaciones. Según datos de diciembre de 2010, dicha base de datos cuenta con 129.902.276 secuencias genéticas disponibles para su consulta en acceso abierto (NCBI, 2010).

Entre los años 2005 y 2008, también comienza a crecer y consolidarse el volumen de datos de investigación como contenido en los RIs académicos. En el informe del proyecto *Making Institutional Repositories in A Collaborative Learning Environment* (MIRACLE) desarrollado entre los años 2005 y 2008, se atribuye dicho crecimiento al incremento del control institucional en el cumplimiento del protocolo de depósito de los datos de investigación por parte del alumnado (Markey y otros, 2007). Más recientemente, los proyectos emprendidos a nivel de universidades, se orientan hacia la gestión de datos *(data management)* en consonancia con los lineamientos sugeridos por la *National Science Foundation*

(NSF) en materia de publicación de datos de investigaciones financiadas con dinero público. Por ejemplo, universidades tales como *University of California, University of Idaho* o *University of Virginia*, incluyen apartados específicos en sus sitios web con información práctica sobre la publicación de datos de investigación para los investigadores de sus comunidades académicas.

En julio del 2010, se ha dado a conocer en España una iniciativa pionera en la publicación en acceso abierto de datos de investigación. A instancias de la Estación Experimental Aula DEI (EEAD), instituto del Consejo Superior de Investigaciones Científicas (CSIC) dedicado a la investigación en el sector agrícola, se ha iniciado una colección denominada «Conjuntos de datos» en el RI Digital CSIC (CSIC, 2010). Esta nueva colección contiene la base de datos SPEIbase con un nuevo índice de sequía, el *Standarized Precipitation-Evapotranspiration Index,* donde se incluyen los datos mensuales (desde enero de 1900 a la actualidad) y cubre la totalidad de la superficie terrestre (exceptuando los polos) con una distribución en rejilla de 0,5º de latitud y longitud. La mencionada colección ofrece para la consulta en abierto la SPEIbase en tres formatos distintos: texto, binario y NetCDF *(Network Common Data Form,* formato específico para la publicación de datos atmosféricos).

Si bien las experiencias presentadas no son las únicas emprendidas en estos países, como tampoco a nivel internacional, tienen como denominador común evidenciar el creciente interés de las universidades e instituciones públicas por hacer visible los resultados del proceso de investigación, contribuyendo de esta forma a la consolidación de infraestructuras donde se accedan a los datos de investigación en forma abierta.

Los RIs brindan una excelente oportunidad para dar un paso adelante en dar acceso en abierto a los datos de investigación, posibilitando una tipificación y descripción especializada de dichos datos y promoviendo la elaboración de manuales que fomenten las buenas prácticas en el desarrollo e implementación de colecciones digitales con estos contenidos.

2.2. El proyecto CLARIN

Desde el año 1994 el Instituto Universitario de Lingüística Aplicada (IULA) es un centro de investigación y formación de investigadores en el contexto de la lingüística aplicada y, en particular, en las temáticas de especialización de los investigadores del centro. Teniendo como marco la lingüística aplicada, los grupos de investigación del Instituto trabajan en temas que, bajo distintas perspectivas, se construyen en torno a cuatro ejes principales: comunicación, información, lenguaje y tecnologías aplicadas. Específicamente, el grupo de Tecnologías de los Recursos Lingüísticos (TRL) trabaja en la creación, desarrollo y aplicaciones de las tecnologías relacionadas con la adquisición, producción, formalización, gestión, validación y evaluación de los recursos lingüísticos necesarios en sistemas de Procesamiento del Lenguaje Natural y Lingüística Computacional.

El grupo TRL participa, desde el año 2004, en diversos proyectos competitivos europeos tales como: *Common Language Resources and Technologies Infrastructure* (CLARIN); *CLARA Initial Training Network for Common Language Resources and their Applications,* ambos dedicados a la creación de infraestructuras para la investigación en humanidades y ciencias sociales; *Platform for the Automatic, Normalized Annotation and Cost-Effective Acquisition of Language Resources for Human Language Technologies* (PANACEA); y *Fostering Language Resources Network* (FlaReNet).

De estos cuatro proyectos, el proyecto CLARIN es el marco bajo el cual se plantea la iniciativa de la publicación en acceso abierto de los resultados de investigación como parte indispensable de una infraestructura para las humanidades y ciencias sociales en general. CLARIN, financiado por el 7PM de la Unión Europea, se encuentra en el final de su primera fase de carácter preparatorio (2008-2010), donde se busca planificar detalladamente la construcción de una infraestructura científica, con una estimación de costos reales, definición del uso de la red, identificación de los centros que la componen y de los recursos y las tecnologías que aseguren su funcionamiento estable. En las fases siguientes, se llevará a cabo la construcción de la plataforma CLARIN y el desarrollo de prototipos de aplicaciones (segunda fase), como también su posterior consolidación y explotación (tercera fase) (Bel y otros, 2008).

El objetivo de CLARIN, siguiendo el modelo de las infraestructuras promovidas por la e-ciencia en otras disciplinas, es el establecimiento de un escenario adecuado que permita acceder a recursos lingüísticos existentes (textos e información lingüística entendidos como datos de investigación), instrumentos para su análisis y explotación. La infraestructura CLARIN se basa en la interacción de los conceptos de tecnología grid, metadatos y servicios web. La tecnología grid proporcionará un contexto de trabajo colaborativo donde se garantice la interoperabilidad de las aplicaciones y el acceso a los recursos lingüísticos, los metadatos permitirán la identificación y ubicación de dichos recursos, en tanto que los servicios web posibilitarán su explotación. El desarrollo del entorno de trabajo propuesto por CLARIN contribuiría a evitar el problema de la dificultad de conocer la existencia de recursos, herramientas y tecnologías lingüísticas y, por tanto, evitar también la consecuente problemática que presenta para su visibilidad, acceso y disponibilidad.

2.2.1. *The Harvesting Day* y el e-Repositori de la UPF

Esta problemática vinculada a la visibilidad, acceso y disponibilidad de los recursos, herramientas y tecnologías lingüísticas no es un tema nuevo. Como exponen Parra y otros (2009) en su artículo, a lo largo del tiempo se han ido desarrollando diferentes iniciativas que buscan subsanar estos inconvenientes.

Es así que experiencias como el proyecto europeo *European National Activities for Basic Language Resources* (ENABLER) en el año 2003, el propio proyecto CLARIN, el catálogo universal de la *European Language Resources Association*

*(*ELRA), el registro del *German Research Center for Artificial Intelligence* (DFKI GmbH), o la encuesta realizada por la *Fostering Language Resources Network (*FlaReNet), tienen como denominador común la preocupación por la identificación, descripción y acceso de los tecnologías, herramientas y recursos lingüísticos. Temas tales como: altos costos de mantenimiento y gestión, dificultad de acceso a los datos de identificación y descripción, utilización no uniforme de esquemas de metadatos, documentación no sistemática, e inestabilidad en el acceso y disponibilidad, permiten componer el mosaico de dificultades existentes.

Las soluciones probadas hasta el momento no han proporcionado mejoras sustanciales pero una posible vía de solución que va ganando adeptos es el fortalecimiento del uso de los metadatos por parte de los proveedores de herramientas, tecnologías y recursos lingüísticos. Una activa utilización de los metadatos permitiría obtener una descripción homogénea y unificada que pueda ser automáticamente recolectada por los diversos catálogos y observatorios del área y así ser descubierta por otros investigadores interesados.

Con esta finalidad el grupo TRL, como participante del proyecto CLARIN, ha promovido dos estrategias. Por una lado una iniciativa denominada *The Harvesting Day* realizada para facilitar la autogestión tanto en la descripción de recursos y herramientas lingüísticas con metadatos como para su publicación en acceso abierto para una posterior recolección automática. Y por otro la utilización del RI de la UPF para la difusión en acceso abierto de datos primarios, recursos y herramientas lingüísticas generadas en la propia universidad.

La iniciativa *The Harvesting Day* se sustenta en la realización de cuatro acciones básicas: el establecimiento de un conjunto de metadatos mínimos; la descripción de recursos y herramientas lingüísticas de acuerdo con dicho conjunto de metadatos; el archivo de estas descripciones en un formato XML (susceptible de ser recopilado e interpretado automáticamente); y la implementación de una infraestructura de fácil instalación que permita la recolección automática de las descripciones de acuerdo con el protocolo OAI-PMH.

Cabe aclarar que el conjunto de metadatos mínimo no es un esquema propio de metadatos, sino que es una versión simplificada de un esquema de metadatos ya establecido en el área de los recursos lingüísticos, el de ENABLER, que a su vez es compatible con otros esquemas de metadatos tales como ISOcat (del proyecto CLARIN), OLAC y LREC-Map.

Adherirse a *The Harvesting Day* es tan sencillo como conectarse a su sitio web [http://www.theharvestingday.eu/] y completar un formulario en línea por cada recurso y/o herramienta lingüística que se tenga. Una vez se completa el formulario se salva la información y se descarga un archivo XML que se guarda en el ordenador del proveedor. Si estos datos se quieren ubicar en acceso abierto para su recolección automática, se debe descargar una aplicación autoejecutable que permite al proveedor instalar un servidor propio. En el año 2010, los días 21 de julio y 21 de setiembre, fueron realizadas dos campañas de recolección automática de descripciones de recursos y herramientas lingüísticas.

La utilización del e-Repositori de la UPF [http://repositori.upf.edu/] para la difusión de datos primarios, recursos y herramientas lingüísticas generadas en la propia universidad tiene dos finalidades complementarias. En primer lugar, busca estudiar la capacidad del repositorio institucional como infraestructura científica para disponer de dichos materiales resultantes de la investigación. En segundo lugar, intenta posicionar como producto de la investigación a dichos objetos digitales, hasta ahora medianamente visibles a través de la página web del instituto de investigación correspondiente, en el caso de recursos y herramientas, o encapsulados en otros objetos digitales, como es el caso de los datos primarios de investigación de los anexos de las tesis.

2.3. La comunidad «Recursos y datos primarios de investigación»

El proyecto del RI de la UPF (e-Repositori) comienza su funcionamiento público el 19 de mayo de 2009, tras un proceso de desarrollo llevado a cabo por la Biblioteca y el Servicio de Informática de la Universitat Pompeu Fabra. En enero del presente año, el e-Repositori da acceso a 4.613 ítems depositados.

La misión del e-Repositori es recoger, difundir y preservar la producción intelectual en formato digital que resulta de la actividad académica e investigadora de la universidad, las revistas científicas y las publicaciones institucionales. Dicha iniciativa complementa las funciones del Portal de Producción Científica (PPC), que tiene entre sus objetivos el de contribuir a la visibilidad de la producción científica de la UPF.

El RI incluye objetos digitales en acceso abierto tanto recolectados de los repositorios consorciados que están funcionando, como documentación introducida directamente por la institución. El contenido en el RI se organiza a partir de espacios virtuales, denominados comunidades, en los cuales se agrupan los materiales digitales de acuerdo a su ámbito (docencia, investigación, vida universitaria, contenidos institucionales) y también a su tipología (artículos de revista, tesis, *working papers*).

Las comunidades que forman en la actualidad el e-Repositori son las siguientes:

Docencia: reúne materiales docentes e informes sobre docencia producidos por el profesorado y el personal de administración y servicios. Estos materiales también se encuentran disponibles en Materials Docents en Xarxa (MDX), repositorio cooperativo del Consorcio de Bibliotecas Universitarias de Cataluña (CBUC).

Investigación (artículos, congresos, libros): incluye documentos resultantes de la investigación como son artículos de revista, libros, comunicaciones, ponencias o pósters presentados en jornadas y congresos, etc.

Investigación (tesis): incluye las tesis doctorales leídas en la universidad. También se encuentran disponibles en Tesis doctorales en red (TDR), repositorio cooperativo gestionado por el Consorcio de Bibliotecas Universitarias de Cataluña (CBUC) y el Centro de Supercomputación de Cataluña (CESCA).

Investigación (working papers, PFC, etc.): documentos de investigación como *working papers,* informes de investigación, artículos aún no publicados (*preprints*), comunicaciones de congresos, proyectos de fin de carrera, memorias técnicas, etc. También se encuentran disponibles en Recercat: Dipòsit de la Recerca a Catalunya.

Revistas científicas: artículos de las revistas científicas publicadas por la universidad o con su colaboración. También se encuentran disponibles en RACO: Revistes Catalanes amb Accés Obert.

Contenidos institucionales: documentación institucional y/o administrativa de la universidad (memorias, informes, discursos, conferencias, etc.)

Vida universitaria: documentos y materiales audiovisuales resultado de las actividades socioacadémicas y de participación de los estudiantes de la universidad.

La implementación de una nueva comunidad en el e-Repositori, dedicada a la publicación de datos de investigación, es el resultado de la conjunción de intereses de dos colectivos institucionales. Por un lado, el interés por parte de los profesionales encargados de la gestión del RI que buscan tanto dotarlo de la máxima variedad de contenidos institucionales (con el objetivo de ofrecer a la comunidad académica en general una visión completa de la producción científica e institucional de la UPF) así como también el acceso y la preservación a largo plazo dicha producción. Por otro lado, el interés de un grupo de investigación de la UPF preocupado también por asegurar el acceso y la preservación de los datos resultantes de la investigación, pero, además, con idea de facilitar la validación de dichos datos, incrementar la visibilidad de los resultados de las investigaciones así como posibilitar la búsqueda y recuperación por parte de otros investigadores (de la propia universidad o de la comunidad académica en general).

Para lograr que el conjunto de ambos intereses, distintos pero complementarios, confluyeran en una misma propuesta, se llevaron a cabo una serie de entrevistas donde se puntualizaron marcos comunes de acción y pautas para el trabajo colaborativo. Fruto de esta tarea de gestión y trabajo colaborativo entre ambos colectivos se implementa la creación de la comunidad «Recursos y datos primarios de investigación».

La experiencia piloto de publicar los datos de investigación en acceso abierto se inicia con la creación de dos colecciones de materiales generados a partir de las actividades de investigación del IULA, pero la idea de futuro es que otros departamentos, institutos u órganos de la universidad sumen sus resultados de investigación, y así mostrar otro aspecto más sobre la investigación desarrollada en la UPF a la comunidad académica en general.

Las dos nuevas colecciones tienen contenidos relativos a la investigación, ya sean datos primarios, recursos o herramientas lingüísticas o de áreas afines, pero con un origen diferente: las tesis doctorales y los proyectos de investigación.

No obstante, cabe aclarar que existen dos limitaciones relacionadas con el proyecto CLARIN que afectan desde el punto de partida la implementación de ambas colecciones.

La primera limitación la constituye la temporalización prevista para la realización del estudio de caso, que se encuentra ligada al calendario de plazos del proyecto CLARIN el cual acaba su fase preparatoria en diciembre de 2010. Debido a esta restricción temporal no se ha podido proponer una adaptación, para obtener una estructura más en consonancia con los nuevos contenidos, en la configuración del programa del e-Repositori.

La segunda limitación la constituye la propia especialización de los recursos y herramientas lingüísticas. Si bien el programa de gestión (DSpace) del e-Repositori permite la adición de otro esquema de metadatos complementario, los tiempos de implementación de esta solución exceden la temporalización prevista para la realización del presente estudio de caso. El uso de un esquema de metadatos de amplia difusión, como el Dublin Core, es una garantía de interoperabilidad en la recolección automática de datos, pero también presenta una faceta negativa en la descripción y recuperación de información con requerimientos temáticos especializados. Para solventar dicho desajuste se planteó el mapeo de ambos esquemas de metadatos, con la consecuente pérdida de especificidad en la descripción y recuperación especializada de los datos de investigación.

El resultado del mapeo se muestra en la siguiente tabla:

TABLA I

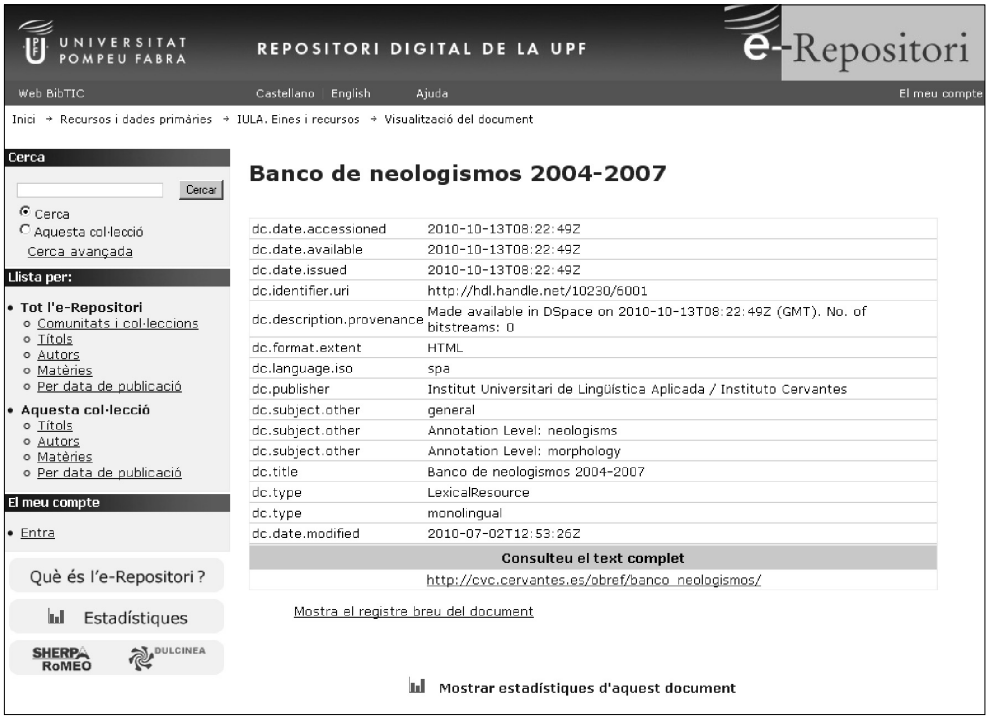
Mapeo entre esquema de metadatos CLARIN y Dublin Core

Clarin	Dublin Core
Resource Title	Title
Url	Identifier
Organization	Publisher
Language	Language
Format	Format
Resource Type	Type
Lexicon Type	Type
Corpus Type	Type
Domain	Subject
Information Contained	Subject
Annotation Level Type	Subject
Operations	Subject

Las celdas de la tabla marcadas con fondo gris contienen los metadatos CLARIN que no tienen equivalencia con los metadatos Dublin Core. Sin embargo, se ha adaptado el mapeo para que se redireccionaran hacia otros metadatos semánticamente compatibles. El metadato que se ha encontrado más apropiado para incluir dicha información, en el esquema Dublin Core, es el de palabra clave (*dc.subject*). Dado el carácter especializado de los metadatos CLARIN, se ha agregado automáticamente, mediante *script,* una etiqueta que identifica el tipo de información especificada en el valor del metadato. Las etiquetas de especificación agregadas a los metadatos son: *Information Contained*, *AnnotationLevel* y *Operations*. (figura 3).

FIGURA 3

Impresión de pantalla del registro en el repositorio institucional: recurso lingüístico (formato ampliado). Metadatos con etiquetas de especificación en los valores



Se puede acceder al esquema completo de metadatos en la siguiente dirección: [http://theharvestingday.eu/schemas/clarin_bamdes-1.1.xsd].

El estudio de la implementación y desarrollo de ambas colecciones se realiza teniendo en cuenta los siguientes ítems: contenido, proceso de incorporación, formato de visualización, resultados y propuestas de mejora.

2.3.1. La colección «IULA. Herramientas y recursos»

Contenido

Es una colección donde se recopilan los recursos y herramientas generadas por los diferentes grupos de investigación del IULA en los proyectos de investigación llevados a cabo. Se encuentran recursos y herramientas tales como corpus lingüísticos (orales y escritos), gestores de diccionarios, vocabularios, plataformas de trabajo, gestores de índices y mapas conceptuales, herramientas para procesamiento lingüístico, etc. Actualmente la colección cuenta con 34 ítems.

Dichas herramientas y recursos, además de incorporarse como colección en el RI, ya estaban accesibles desde la página principal del sitio web del Instituto en el apartado «Recursos. IULA» [http://www.iula.upf.edu/recurs06ca.htm]. Asimismo, los ítems de esta colección también constituyeron la materia prima para la participación en la iniciativa *Harvesting Day*, para lo cual se describieron siguiendo el esquema de metadatos propuesto por el proyecto CLARIN. Por consiguiente, se realizó la descripción de cada uno de los recursos y herramientas a través del formulario web y se generaron los archivos XML que fueron guardados en el servidor del grupo de investigación TRL para su futura recolección automática.

Proceso de incorporación

Dado que las herramientas y recursos incluidos en la colección ya estaban descritos según el esquema de metadatos CLARIN (para participar en la iniciativa *Harvesting Day,* véase apartado 2.2.1), fue necesaria la creación de un *script* para gestionar las equivalencias de mapeo con Dublin Core.

Tras pasar el *script* de mapeo se generaron los archivos XML que permitieron la recolección automática de metadatos compatible con el esquema Dublin Core.

Formato de visualización

La visualización de las entradas en el RI tiene dos formatos: el primero, por defecto, con un conjunto de metadatos básico y el segundo, disponible en la opción «Mostrar el registro completo del documento», con un conjunto de metadatos más amplio.

Los metadatos visibles en el formato básico, como se muestra en la figura 4, son: título, fecha, citación (donde se exportan los datos al programa de citación bibliográfica *Ref Works*) y el vínculo que enlaza con el recurso o herramienta en cuestión.

En cambio los metadatos visibles en el formato ampliado, como se muestra en la figura 5, son: fecha (de publicación, de modificación, de incorporación y de disponibilidad), identificador persistente Handle, descripción (procedencia), formato, lengua (código ISO), editor, materia, título, tipo y el vínculo que enlaza con el recurso o herramienta en cuestión.

FIGURA 4

Impresión de pantalla de un registro del repositorio institucional: recurso lingüístico (formato básico)

Cabe destacar que las etiquetas de especificación agregadas al valor del metadato *dc.subject* son recuperables tanto a través de la opción de búsqueda como por el índice predeterminado de materias. Por tanto, las especificaciones temáticas que se habían mapeado del esquema de metadatos CLARIN a Dublin Core constituyen una mejora en la visibilidad, búsqueda y recuperación de dichos recursos.

Resultados y propuestas de mejora

Los objetos digitales de la colección tienen tres características fundamentales: presentan un conjunto de datos diferente de identificación, la ubicación física no se encuentra en el servidor del repositorio institucional y los derechos de autor no están identificados.

Tener un conjunto de datos de identificación atípico produce vacíos de información en los recursos de búsqueda predeterminados en el RI. Por ejemplo, en los recursos y herramientas lingüísticas se produce una situación poco común a nivel documental, dada su naturaleza, pues no cuentan con un autor identifi-

FIGURA 5

Impresión de pantalla de un registro del repositorio institucional: recurso lingüístico (formato ampliado)

cado. Esta situación provoca que dichos objetos digitales no incluyan el metadato necesario para que tengan presencia tanto en las opciones de búsqueda (simple y avanzada) como en los índices predeterminados del RI, entre los cuales figura el de autores. Una posible mejora para incrementar las opciones de recuperación de este tipo de objetos digitales seria ampliar la lista de índices predeterminados, por ejemplo, con la adición de un índice por tipología (a partir del metadato *dc.type*), debido a la importancia que adquiere dicho metadato para recuperar contenido en colecciones digitales cada vez más diversificadas.

Con relación a que la ubicación física no se encuentre en el servidor del RI, sino en servidores nativos en donde fueron generados y funcionan, implica una cierta sombra de duda sobre su persistencia y accesibilidad. No obstante, conjuntamente con el personal de gestión del repositorio, se tuvo en cuenta que son recursos y herramientas con un probado funcionamiento y uso por parte de la comunidad investigadora, por lo cual no se darían de baja de los servidores nativos. Como proyección de futuro, a nivel institucional, quizá se debería plantear la necesidad de contar con un servidor dedicado a los datos y aplicaciones generadas en la universidad.

La identificación imprecisa de los derechos de autor tiene su explicación en que dichos recursos y herramientas fueron generados para realizar un uso libre de ellos, pero esta condición no fue reflejada en una declaración de derechos de autor o licencia de uso explícita y de libre acceso. Esta «ausencia» aparente de derechos puede ser el punto de partida para debatir, a nivel institucional, sobre cómo deben publicarse este tipo de objetos digitales, los recursos y herramientas lingüísticas y si es posible aplicar licencias *Creative Commons, Scientific Commons* o si es necesario crear un grupo específico de licencias.

2.3.2. La colección «IULA. Datos primarios de las tesis doctorales»

Contenido

Las tesis son materiales que tienen gran importancia a nivel de producción científica debido a que son trabajos realizados a partir de una investigación original y especializada, cuya finalidad última es aportar nuevos conocimientos a un área temática específica. Si bien su acceso, difusión y preservación ya tienen un camino consolidado en el ámbito de los RIs (y repositorios cooperativos), los contenidos accesibles son principalmente el texto completo acompañado de una descripción donde no figuran los materiales anexos, los cuales generalmente reproducen los datos generados durante el transcurso de la investigación.

Quizá no todas las tesis contienen anexos con datos primarios susceptibles de publicar, ya sea por la naturaleza del contenido de la tesis o por los derechos de autor de terceras partes, pero en cualquier caso, de existir dichos materiales se quedan fuera de los circuitos actuales de acceso, difusión y preservación de la producción científica en acceso abierto.

El IULA, a diciembre de 2010, cuenta con 39 tesis doctorales defendidas. Para la inclusión de los anexos en la colección del repositorio se fijaron como condiciones:

a) Que los anexos fueran herramientas o recursos lingüísticos o de áreas afines con capacidad para ser reutilizados por otros investigadores.
b) Que no estuvieran sujetos a derechos de autor de terceros.

Teniendo en cuenta ambos criterios, finalmente se solicitó la autorización a ocho autores de tesis doctorales. Se les envió una carta, de estructura similar a la que se envía para el depósito de las tesis en el repositorio cooperativo de tesis electrónicas Tesis doctorales en Red (TDR), donde se indicaba expresamente sobre que anexos se solicitaba la autorización para publicarlos en el RI. Luego de realizar dicha gestión se obtuvieron seis respuestas positivas.

Los anexos que se incluyen en el estudio de caso son un código fuente de programa, dos bases de datos Access y cuatro interfaces en HTML de consulta. Una de las tesis tiene dos anexos incluidos en el estudio, una base de datos y una interfaz HTML de acceso a datos primarios. Actualmente la colección incluye 6 ítems.

Proceso de incorporación

Debido a que dichos objetos digitales se encontraban en el CD-ROM adjunto de las tesis, como tarea previa se recuperaron los archivos que constituían los anexos que podían formar parte del estudio. Dada su variada tipología, a sugerencia del personal encargado de la gestión del repositorio, se incluyó el conjunto de archivos que conformaba un anexo determinado en un solo archivo zip, el cual fue publicado en el RI. Además de dicho archivo comprimido, el personal encargado de la gestión del repositorio, incluiría un enlace para ejecutar la aplicación directamente en los casos en que fuera necesario, por ejemplo, en las interfaces HTML de consulta a los datos primarios.

Cada uno de los anexos fue introducido en el e-Repositori de forma manual, a través del formulario de entrada de datos.

Formato de visualización

La visualización de las entradas en el RI tiene dos formatos habilitados, el primero por defecto, con un conjunto de metadatos básico y el segundo disponible en la opción «Mostrar el registro completo del documento», con un conjunto de metadatos más amplio.

Los metadatos visibles en el formato básico, como se muestra en la figura 6, son: título, autor, descripción, fecha, citación (donde se exportan los datos al programa de citación bibliográfica *Ref Works*), identificador persistente Handle, mención de derechos y enlace al fichero publicado.

En tanto, los metadatos visibles en el formato ampliado, como se muestra en la figura 7, son: fecha (de publicación, de modificación, de incorporación y de disponibilidad), identificador persistente Handle, descripción, descripción (procedencia), formato, lengua (código ISO), mención de derechos, título, tipo, palabras clave y enlace al fichero publicado.

Cabe recordar, que los metadatos presentes en ambos formatos de visualización (básico y ampliado), son recuperables tanto por la opción de búsqueda (simple y avanzada) como por los índices predeterminados (en el caso de metadatos relativos a los autores, títulos, materias y fecha de publicación). Destaca la posibilidad de realizar búsquedas por los valores del metadato *dc.type*, con lo cual se incrementa la posibilidad de recuperar concretamente datos primarios entre la diversidad de contenidos posibles del RI.

Resultados y propuestas de mejora

Los anexos de las tesis incluidos en el estudio son objetos digitales de diferente naturaleza y con distintos formatos: código fuente en Perl, intefaces HTML y bases de datos Access, que se encuentran guardados en el CD-ROM de anexos que incluyen las tesis doctorales.

FIGURA 6

Impresión de pantalla de un registro del repositorio institucional: anexo tesis doctoral (formato básico)

Debido a que son materiales digitales con formatos y tipos no tradicionales en la producción científica en general, se debería contar, además de las posibilidades de búsqueda (simple y avanzada), con índices predeterminados de otros metadatos más útiles para la recuperación, por ejemplo *dc.type*.

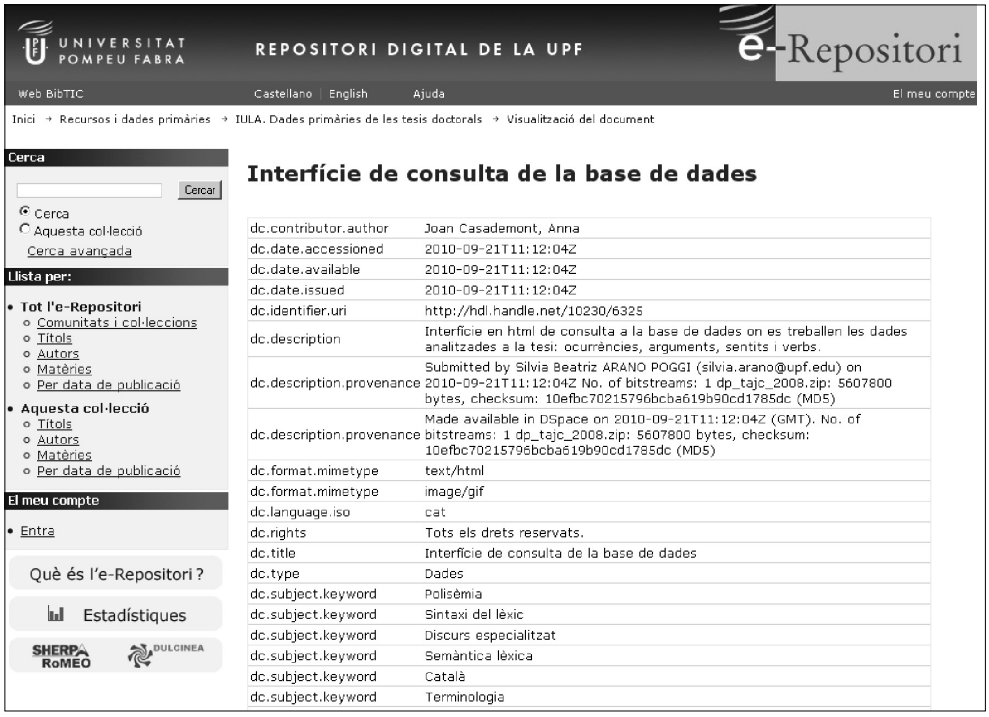
De la misma forma, motivado por su diversa naturaleza y formato, sería adecuado agregar metadatos al formato básico de visualización, por ejemplo la descripción. Dicho elemento descriptivo ayudaría a la comprensión de objetos digitales con una cierta complejidad dada por su carácter especializado.

3. Conclusiones y perspectivas de futuro

Los RIs comienzan a consolidar su presencia como infraestructuras científicas de la mano de la progresiva publicación de datos de investigación en abierto. Muestra de ello son los múltiples proyectos e iniciativas emprendidos tanto a nivel institucional como gubernamental en diversos países, de los cuales hemos

FIGURA 7

Impresión de pantalla de un registro del repositorio institucional: anexo tesis doctoral (formato ampliado)



mencionado una breve muestra (*eBank UK*, *DISC-UK DataShare*, *Australian National Data Service*, *Genbank*, CSIC). La realización del presente estudio de caso pretende contribuir de forma práctica, al desarrollo y diversificación de los RIs en su papel de infraestructura científica de apoyo para la visibilidad, acceso y gestión de la producción científica institucional, cualquiera que sea su formato.

Es evidente que con la implementación de las colecciones se han obtenido beneficios, pero también se han abierto nuevas incógnitas y problemáticas por resolver.

En el ámbito de los beneficios obtenidos de esta iniciativa, en primer lugar cabe mencionar el logro de hacer visibles y accesibles los datos, recursos y herramientas de investigación como parte de la producción científica de la UPF. De otro modo, en el mejor de los casos para recursos y herramientas, solamente estarían presentes en las páginas web de los grupos de investigación, y, en el peor de los casos, para los datos primarios de investigación, seguirían ocultos en los CD-ROMs anexos a las tesis.

En segundo lugar, también reporta un beneficio la asignación de un identificador Handle al ser incluidos en el RI datos primarios y recursos lingüísticos, con el cual se ayuda a su persistencia y accesibilidad a largo plazo. Sin embargo, aún

queda reflexionar sobre el tema de la ubicación física de los recursos y herramientas lingüísticas, las cuales se encuentran en los servidores nativos donde fueron creadas y son utilizadas. En este sentido, en parte también dependiendo de la suma de otras áreas disciplinarias de la institución a dicha iniciativa, se debería estudiar la posibilidad de contar con un servidor propio para los datos y aplicaciones generadas en la universidad.

Por último, también es un beneficio la posibilidad de recolectar automáticamente, por parte de otros repositorios, las descripciones de tales materiales. Esto facilita la difusión y conocimiento de este tipo de producción científica, incrementando las posibilidades de reutilización en otras investigaciones.

Como problemática de base a resolver, se presenta la flexibilización de la configuración de los programas gestores de los repositorios. En general, dichos programas cuentan con una configuración ligada al depósito de objetos digitales de tipo publicación o multimedia, característica que los limita en la actualidad debido a la diversificación de contenidos que puede albergar un RI. Particularmente en el presente estudio de caso, la estructura del programa gestor del repositorio (el e-Repositori) vincula tanto los formularios de entrada de datos como los metadatos de búsqueda e índices predeterminados al esquema de metadatos Dublin Core representando una dificultad. Dicha restricción constituye una limitación importante para la publicación en abierto de los datos de investigación, recursos o herramientas, que debido a su carácter fuertemente especializado necesitan de una descripción con mayor detalle, para así propiciar su búsqueda y recuperación por parte de otros investigadores de la misma área temática.

Otro tema sobre el cual es necesario realizar un análisis profundo, es la identificación compleja de los derechos de autor en relación con los objetos digitales que no son publicaciones. Actualmente, con las posibilidades de publicación en acceso abierto, los investigadores deben plantearse la necesidad de establecer los derechos y tipos de licencia que sus datos primarios, recursos o herramientas tienen para que puedan ser reutilizados en otras investigaciones. Si bien el equipo de trabajo de personas encargadas de la gestión del e-Repositori también cuenta con especialistas en derechos de autor, es necesario que esta cuestión sea reflexionada tanto a nivel institucional como por los propios creadores, los investigadores, estableciendo pautas y protocolos sobre la posibilidad de adjudicar licencias *Creative Commons*, *Scientific Commons* o, si es necesario, crear un grupo específico de licencias.

Una última cuestión para reflexionar y debatir es el tema de los formatos de los datos primarios. Actualmente estos datos quedan relegados en los CD-ROMs anexos a las tesis, razón por la cual no existe una real preocupación por su accesibilidad más allá de la consulta del propio CD-ROM. Con las posibilidades que se abren si dichos datos se incorporan a los RIs, es necesario planificar las formas en las cuales se puede garantizar la accesibilidad a largo plazo de dichos contenidos. Esta es la razón por la cual los futuros protocolos, que pauten la presentación de las tesis y sus anexos, deben tener en cuenta la gestión de los datos primarios tanto para su inclusión en la tesis como para su posterior incorporación a un RI.

Sin embargo, más allá de la existencia de cuestiones que no son fáciles de solventar, sí existen ciertas medidas que pueden adoptarse a mediano plazo y que posibilitarían la resolución de algunas de ellas.

Dada la posibilidad de publicar datos de investigación, recursos y herramientas en el RI, sería beneficioso revisar la política de depósito de materiales digitales en el e-Repositori, como también las normativas de depósito de trabajos de final de carrera y tesis doctorales. En ambos protocolos sería necesario detallar cómo deben ser las estructuras, formatos y licencias de explotación, en un lenguaje claro y sencillo, previendo posibles lagunas de conocimiento técnico de los investigadores.

En función del éxito y acogida de la implementación de una comunidad en el RI de la UPF para los datos de investigación, se podría comenzar a pensar a nivel institucional en la implementación de diversas soluciones. A modo de ejemplo, el estudio de caso propone la implementación de una estructurada federada de repositorio y catálogos o bases de datos por áreas temáticas a nivel institucional, para garantizar, por un lado, los temas de accesibilidad, persistencia y seguridad, y, por otro, la necesidad de descripciones detalladas para favorecer la búsqueda y recuperación en entornos especializados.

Como primera medida a adoptar se propone el fortalecimiento del RI de cara a que mejore el tratamiento de datos y aplicaciones, además de los contenidos actuales de texto y multimedia. Dicha mejora garantizaría la persistencia, accesibilidad y seguridad en torno a la gestión de datos y aplicaciones a nivel institucional, a la vez que permitiría acceder a una descripción básica de cada ítem según el esquema de metadatos Dublin Core. Paralelamente, cada área disciplinaria de la institución podría contar con catálogos propios donde se pudieran describir en forma detallada y con esquemas de metadatos especializados los datos y aplicaciones específicas del área, para incrementar las opciones de búsqueda y recuperación por parte de otros investigadores del área en cuestión. Cada área disciplinaria actuaría como un agente intermediario (*broker*), donde proporcionaría el entorno adecuado para realizar búsquedas especializadas, pero también se utilizaría el RI para otorgar el acceso a los datos o aplicaciones. De esta forma, se podrían satisfacer tanto las preocupaciones institucionales de gestión de dichos objetos digitales, como las de los investigadores de la búsqueda y recuperación especializadas, en definitiva del descubrimiento por parte de otros investigadores de su producción científica de datos primarios, recursos y aplicaciones.

La realización del presente estudio de caso solamente ha abordado algunas de las cuestiones que involucra utilizar un RI para la publicación de datos primarios, recursos y herramientas en acceso abierto. Aún quedan cuestiones por resolver y se perciben nuevos temas a tratar, tales como la integración de los repositorios en otras infraestructuras científicas, la concienciación de la importancia de la asignación de metadatos para posibilitar la búsqueda y recuperación por parte de otros investigadores así como la necesidad de preservar y conservar institucionalmente a largo plazo los datos primarios y la determinación de los roles y responsabilidades de los encargados e interlocutores de la gestión de dichos datos.

4. Agradecimientos

Este trabajo ha sido desarrollado en el marco del proyecto CLARIN, proyecto co-financiado por el Ministerio de Ciencia y Tecnología (ACI2009-0995) de España, por la Generalitat de Catalunya, y por la Unión Europea (FP7-INFRASTRUCTURES-2007-1-212230).

5. Bibliografía

Bel, N.; Bel, S.; Espeja, S.; Marimon, M., y Villegas, M. (2008): El proyecto CLARIN: una infraestructura de investigación científica para las humanidades y las ciencias sociales. *Digithum (*10). http://www.uoc.edu/digithum/10/dt/esp/bel_bel_espeja_marimon_villegas.pdf [consulta: 20 de noviembre de 2010].

CSIC (2010): La investigación en abierto: dos nuevas colecciones en Digital.CSIC. *CSIC Abierto* (1), 1-5. http://hdl.handle.net/10261/26261 [consulta: 28 de octubre de 2010].

CRIG (2008): *JISC Common Repository Interfaces Group (CRIG)*. http://www.ukoln.ac.uk/repositories/digirep/index/CRIG [consulta: 26 de noviembre de 2010].

Green, Ann; Macdonald, S., y Rice, R. (2009): *Policy-making for Research Data in Repositories: A Guide:* version 1.2. Londres: JISC, 40. http://www.disc-uk.org/docs/guide.pdf [consulta: 28 de octubre de 2010].

Hey, T.; Tansley, S., y Tolle, K. (eds.) (2009): *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond Wash: Microsoft Research, 252. http://research.microsoft.com/en-us/UM/redmond/about/collaboration/fourthparadigm/4th_PARADIGM_BOOK_complete_HR.pdf [consulta: 28 de octubre de 2010].

Markey, K.; Rieh, S. R.; St. Jean, B.; Kim, J., y Yakel, E. (2007): *Census of Institutional Repositories in the United States: MIRACLE Project Research Findings*. Washington, D.C.; Council on Library and Information Resources, 167. http://www.clir.org/pubs/reports/pub140/contents.html [consulta: 28 de octubre de 2010].

Melero, R.; Abadal, E.; Abad, F., y Rodríguez-Gairín, J. M. (2009): *Situación de los repositorios institucionales en España: informe 2009* [s.I.]; Grupo de investigación Acceso Abierto a la Ciencia, 54. http://hdl.handle.net/10261/11354 [consulta: 28 de octubre de 2010].

NCBI (2010): *NCBI-GenBank Flat File Release 181.0. Distribution Release Notes*. Bethesda: National Library of Medicine. ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt [consulta: 19 de enero de 2011].

Parra, C.; Villegas, M., y Bel, N. (2009): The Basic Metadata Description (BAMDES) and TheHarvestingDay.eu: Towards Sustainability and Visibility of LRT. Proceedings *Seventh Conference on International Language Resources and Evaluation (LREC'10)*, 49-53. Valletta, Malta: ELRA. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W20.pdf [consulta: 20 de noviembre de 2010].

Rice, R. (2007): DISC-UK Datashare Project: Building Exemplars for Institutional Data Repositories in the UK. *IASSIST Quarterly*, 31 (3/4), Fall & Winter 2007, 21-27. http://www.iassistdata.org/downloads/iqvol313rice.pdf [consulta: 28 de octubre de 2010].

RIN (2008): *Stewardship of digital research data – principles and guidelines*. Londres: RIN. 15. http://www.rin.ac.uk/system/files/attachments/Stewardship-data-guidelines.pdf [consulta: 28 de octubre de 2010].