Hacia una universidad abierta

Recomendaciones para el SUE

CRUE TIC Comisión Sectorial de las Tecnologías de la Información y las Comunicaciones

Hacia una universidad abierta:

Recomendaciones para el SUE

Coordinación y edición:

Enrique Teruel Doñate

Delegado del Rector para los Servicios de Informática y Comunicaciones – Universidad de Zaragoza

*Coordinador del **subgrupo de Gobierno Abierto***

José Pascual Gumbau Mezquita

Director del Gabinete de Planificación y Prospectiva Tecnológica – Universitat Jaume I

*Coordinador del **grupo de Análisis, Planificación y Gobierno de las TI** de la Sectorial TIC de la CRUE*

Autores:

Este documento es fruto del trabajo del subgrupo de Gobierno Abierto del grupo de Análisis, Planificación y Gobierno de las TI de la Sectorial TIC de la CRUE. Naturalmente, el trabajo, y el grupo, está abierto a la participación y colaboración.

- Ricardo Borillo Doménech (Universitat Jaume I)
- Sonia Castro Muñoz (Red.es)
- María José García (Universidad Autónoma de Madrid)
- Irene Garrigós Fernández (Universidad de Alicante)
- José Pascual Gumbau Mezquita (Universitat Jaume I)
- Francisco Javier López Pellicer (Universidad de Zaragoza)
- José Norberto Mazón López (Universidad de Alicante)
- Jorge Pantoja (Universitat Pompeu Fabra)
- Miguel Angel Sicilia (Universidad de Alcalá de Henares)
- Enrique Teruel Doñate (Universidad de Zaragoza)[1]
- Alicia Troncoso Lora (Universidad Pablo de Olavide)
- José Jacobo Zubcoff Vallejo (Universidad de Alicante)

Naturaleza y licencia de este documento (open & linked):

Este documento, es la versión 20131210 de un documento vivo de trabajo. El documento vivo está radicalmente abierto a nuevas participaciones, especialmente universitarias

Ha sido deliberadamente preparado para prevenir su impresión en papel, las fuentes están enlazadas en lugar de citadas, y algunas de ellas son audiovisuales, no textuales.

[1] *Agradezco a mis compañeros Javier Luna, Marta de Miguel, Sergio Montesa y Estefanía Serrano su impulso inicial, incluida su respuesta a la petición de ayuda para superar el síndrome de la página en blanco. Su colaboración pasada, presente y seguro que futura siempre es bienvenida.*

1. Introducción

El objeto de este documento, fruto del trabajo conjunto del grupo constituido en el marco CRUE-TIC, es promover y facilitar la aplicación de los principios de gobierno abierto en las universidades, sirviendo como guía que describa los puntos que deben ser abordados en la puesta en marcha de una iniciativa así, con el propósito de contribuir a la construcción de una universidad abierta del siglo XXI.

Dado que el primer principio del gobierno abierto es la transparencia, y ésta se practica de forma idónea mediante la apertura de datos, los primeros pasos deberán sentar las bases para que las universidades empiecen a abrir sus datos institucionales de forma coherente, desarrollando políticas de transparencia que conviene adoptar, conjuntamente con políticas de acceso abierto al conocimiento, tanto docente como investigador.

1.1. ¿Qué es gobierno abierto?

El gobierno abierto surge como un nuevo modelo de relación entre los gobernantes, las administraciones y la sociedad, del que las universidades no deben permanecer al margen.

El concepto de gobierno abierto considera a los ciudadanos responsables de la dirección y del control de los servicios que las administraciones públicas les prestan, y les da el poder de solucionar los problemas.

La postura habitual de los gobiernos ha sido la de salvaguardar los datos, las informaciones y el conocimiento que les sirven de base para la toma de decisiones. En este celo protector se han volcado grandes esfuerzos, y en muchos casos hasta hace poco tiempo hasta ahora, no se había planteado qué ocurriría si en lugar de ocultar los abriesen, los hiciesen de dominio público, y dejasen interactuar con ellos a los ciudadanos.

Las primeras experiencias han sido tan alentadoras que se ha producido un movimiento global de apertura de datos y gobiernos, cuyo nacimiento todavía estamos presenciando. Por ejemplo, la reciente declaración del G8 en Lough Erne (18 de Junio de 2013), que es un conciso decálogo de políticas sobre la responsabilidad de los gobiernos, incluye un punto sobre apertura y reutilización de información.

Los tres principios fundamentales de este movimiento, recogidos en el celebrado "memorando de Obama" son:

- Transparencia: oferta de información clara y actualizada, accesible y reutilizable.
- Participación: intervención de la ciudadanía en todas las actividades del gobierno.
- Colaboración: entre instituciones y llamando a los ciudadanos a compartir lo que saben y a generar soluciones en las áreas donde tienen conocimientos.

De forma más visual, y en tres minutos, los principios de gobierno abierto se ilustran en este video promocional de la Open Government Partnership.

1.2. ¿Qué es transparencia y datos abiertos?

La transparencia permite a todos los ciudadanos conocer y vigilar el empleo de los recursos públicos y estimula a las instituciones a funcionar de modo más eficiente. Para éstas supone además una herramienta imprescindible para recuperar la confianza del ciudadano.

La administración es una gran fuente de información que cuenta con cantidad de datos de carácter público: mapas, meteorología, tráfico, datos financieros, subvenciones, planes urbanísticos, acuerdos políticos, informes de investigación financiados públicamente en base a los cuales se aprueban normas... Toda esta información pertenece a la ciudadanía.

Los datos serán la materia prima para que nuevos actores elaboren nuevos productos y servicios, creando valor y riqueza. Es una forma de capacitar a los ciudadanos y hacerlos más responsables regenerando la conexión

entre políticos y ciudadanos, facilitando la participación de estos últimos en la vida pública. La Reutilización de la Información del Sector Público (RISP), avalada oficialmente por directivas europeas y sus transposiciones locales, es un objetivo principal de cualquier iniciativa Open Data, así que conviene elaborar un plan de reutilización. El W3C también ha elaborado un documento de recomendaciones, aparte de trabajar activamente en tecnologías de interés (formatos, web semántica, etc).

A su vez el proceso de limpiar y preparar los datos para su publicación es en sí mismo beneficioso para los miembros de las administraciones que necesitan acceder a los datos. En otras palabras, abrir la información a los ciudadanos significa hacer la información más accesible dentro de la propia administración. De hecho, hoy por hoy, uno de los reutilizadores más señalados de la información es la propia administración, y en menor medida la universidad.

Dado que la accesibilidad de los datos, su estructuración, su disponibilidad en formatos abiertos y su interconexión favorecen la reutilización, los catálogos de datos abiertos se clasifican por estos criterios, siguiendo la propuesta de "cinco estrellas" de Tim Berners-Lee. Las ventajas de la apertura de datos tecnológicamente avanzada se ilustran en este video promocional de la fundación CTIC.

La transparencia es demandada de forma cada vez más clamarosa a todas las administraciones y otras entidades a resultas de escándalos reiterados. Es una exigencia que también afecta a las universidades, incluidas las privadas, pues nos debemos a la sociedad a la que servimos, y es inexcusable rendirle cuentas. La Fundación Compromiso y Transparencia publicó en septiembre de 2012 un informe sobre la transparencia de las universidades revisando sus portales web, y lo ha actualizado en 2013.

1.3. ¿Qué es acceso abierto?

La necesidad de promover la difusión de la investigación y la docencia en el contexto de la sociedad de la información ha favorecido que surgiera el movimiento del acceso abierto, conocido con las siglas OA (Open Access), que promueve que el conocimiento sufragado públicamente sea de dominio público y postula un acceso permanente, gratuito y libre de restricciones a los contenidos científicos y académicos.

Declaraciones internacionales sobre acceso abierto:

- Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (22 de octubre de 2003).
- Bethesda Statement on Open Access Publishing (20 de Junio de 2003).
- Budapest Open Access Initiative (14 de febrero de 2002).
- Declaración de CRUE-REBIUN en apoyo del modelo de acceso abierto (2004).

Recomendaciones e iniciativas europeas:

- Recomendación de la Comisión, de 17 de julio de 2012, relativa al acceso a la información científica y a su preservación (2012/417/UE)
- Iniciativa OpenAIRE (Open Access Infrastructure for Research in Europe, Infraestructura de libre acceso a la investigación en Europa. http://www.openaire.eu/

Legislación española:

- Ley 14/2011, de 1 de junio, de Ciencia, Tecnología e Innovación,
- Real Decreto 99/2011, de 28 de enero, por el que se regulan las enseñanzas oficiales de doctorado

Varias universidades españolas tienen políticas institucionales de acceso abierto, a pesar de que se excluyen de las obligaciones y recomendaciones RISP "*los documentos conservados por instituciones educativas y de investigación, tales como centros escolares, universidades, archivos, bibliotecas y centros de investigación, con inclusión de organizaciones creadas para la transferencia de los resultados de la investigación (art. 3.3. Ley 37/2007)*", es decir, se desaprovechó la ocasión para impulsar la reutilización de los contenidos científicos

y académicos, aunque afortunadamente ésta se impulsa a menudo por propia iniciativa, dado que son principios que están en sus genes.

Los datos a que se refieren estas iniciativas son típicamente de tres clases:

- Material docente: OCW, MOOC
- Las publicaciones y otros documentos más o menos formales que recogen los resultados de la investigación o informan sobre ella.
- Datos experimentales o técnicos sobre los que se realiza una investigación, para que pueda ser replicada por otros, o para que otros puedan usarlos para sus propias investigaciones.

1.4. Iniciativas de apertura

A nivel nacional, desde 2001, el proceso de apertura es impulsado ministerialmente mediante el proyecto interministerial Aporta, y su portal datos.gob.es. Las políticas de gobierno abierto se encuentran en una fase de auténtica eclosión por lo que es difícil mantener un catálogo actualizado de iniciativas, así que mucho mejor enlazar el catálogo CTIC, aunque sea razonablemente incompleto. Varias de estas iniciativas recogen en sus portales interesante información introductoria y de referencia: definiciones, principios, catálogos de buenas prácticas, licencias, ejemplos de uso, etc.

Por ahora la única iniciativa de apertura de datos universitaria española recogida en el catálogo CTIC es la de la Universidad Pablo Olavide de Sevilla, aunque recientemente se han celebrado días de datos abiertos también en la Universidad de Granada y en la de Deusto, y nos consta que otras universidades están preparando sus pilotos de apertura de datos. En cualquier caso, en general, no abundan iniciativas en las universidades del mundo que puedan servirnos de ejemplo, hay un puñado en linkeduniversities, siendo destacable la de la Universidad de Southampton, la de la Universidad de Ege, o la reciente incorporación de la Universidad Pompeu Fabra. Así mismo se empiezan a recoger colecciones de datos abiertos, como en linkededucation, que vienen a sumarse a otros recursos en abierto tradicionalmente ofrecidos de forma más o menos dispersa, y que hoy en día las tecnologías de la web semántica nos permiten enlazar, como expone el informe Linked Data for Open and Distance Learning.

Parece evidente que la apertura está en los "genes universitarios", por lo que debieran ser la norma, no la excepción, políticas explícitas como la de la Universidad Jaume I de Castellón, que dispone en sus Estatutos que "*fomentará el uso de formatos informáticos abiertos en la comunicación interna y externa, promoverá el desarrollo y el uso de software libre y favorecerá la libre difusión del conocimiento creado por la comunidad universitaria*". Desde luego, existen unos cuantos núcleos de interés en nuestras universidades, así como grupos académicos con perfiles distintos (legal, político, de comunicación, técnico, semántico...). En el recientemente creado capítulo español de la Open Knowledge Foundation se ha creado un subgrupo para trabajar hacia la universidad abierta. La Open Knowledge Foundation tiene un grupo de trabajo específicamente dedicado a educación, el Open Education Working Group.

La apertura en la universidad se manifestará además en otros ámbitos, complementarios, como el acceso abierto a los resultados científicos, la educación en abierto, o mediante procesos participativos/abiertos en sus acciones de gobierno, de lo que el reciente proceso participativo integrado en la elaboración del Plan Estratégico de la Universidad Pablo de Olavide es un ejemplo a imitar.

Como muestra de la complementariedad de las aproximaciones a la apertura, un botón: el curso sobre datos abiertos en la plataforma de cursos online masivos en abierto de la UNED.

Una revisión detallada de los portales de datos abiertos de universidades permite identificar qué tipo de conjuntos de datos son los habitualmente más publicados:

Información organizativa

- Organigrama. Organizaciones y unidades organizativas que forman parte de la estructura de la universidad.

- Perfiles del personal de la universidad (profesores, investigadores, administrativos, etc.)
- Directorio del personal de la universidad.
- Información de ofertas de trabajo en la universidad.

Docencia

- Catálogo de estudios.
- Calendarios y horarios de los cursos.
- Programas de movilidad.
- Catálogos de contenidos educativos.

Alumnado

- Perfil de los alumnos matriculados.
- Programas de becas
- Rendimiento académico (calificaciones de los estudiantes matriculados, porcentaje de alumnos aprobados, tasa de duración de estudios, tasa de abandonos,…).
- Índices de satisfacción de los alumnos.

Economía

- Estado de cuentas púbico. Informe financiero anual.
- Presupuesto de ingresos.
- Presupuesto de gastos.
- Distribuciones del presupuesto.
- Ayudas y subvenciones.
- Licitaciones

Infraestructuras y servicios

- Edificios y otras entidades geoespaciales presentes en la universidad.
- Información geoespacial de los edificios de la universidad (rectorado, facultades, departamentos, aulas,…).
- Imágenes de los edificios e instalaciones de la universidad.
- Información de accesibilidad de los edificios y servicios de la universidad.
- Equipamiento de las aulas, tanto su localización como sus características (capacidad, plazas, accesibilidad, etc.).
- Catering. Menús de la cafetería.
- Instalaciones y equipos de la universidad.
- Accesos.
- Catálogo de servicios informáticos.
- Catálogo de servicios bibliotecarios y catálogo bibliográfico de la biblioteca de la universidad.
- Servicios a estudiantes: alojamientos (número de plazas/tipo, ubicación, precio...), cafeterías, máquinas expendedoras, instalaciones deportivas, actividades culturales.

Comunicación

- Eventos que se realizan en la universidad.
- Noticias relacionadas con la universidad.

En cuanto a los formatos empleados a la hora de publicar la información, estos difieren de una universidad a otra. Así, la Universidad de Southampton, una de las pioneras en la apertura de datos universitarios, la de Oxford o la de Münster sitúan sus conjuntos de datos en el marco de la web semántica (vinculación de datos) mediante la utilización de formatos RDF (en alguna de sus formas de serialización: XML, N3, Turtle), mientras que otras universidades se inclinan más por el uso de formatos de uso más común como hojas de cálculo (XLS) o archivos separados por comas (CSV). Cabe destacar que los conjuntos de datos ofrecidos a través de estos portales se disponen bajo licencias que permiten la reutilización de la información tanto para fines

comerciales como no comerciales, tales como licencias Open Government Licence o del tipo Creative Commons.

2. Políticas de apertura

Recogiendo las experiencias previas, y entendiendo que el primer paso para avanzar hacia la apertura es la adopción institucional de políticas claras que fijen la orientación y alcance de las iniciativas, en esta sección se proponen formulaciones básicas de estas políticas, consensuadas dentro del grupo de trabajo de CRUE-TIC, que se propone que sean revisadas y debatidas por un conjunto amplio de universidades interesadas. Así, este trabajo servirá a todas las universidades, que adoptarían, o adaptarían, estas formulaciones básicas.

2.1. Política de transparencia

La *Universidad* <...> se compromete con el principio de transparencia, y declara su disposición a publicar datos abiertos de todos sus ámbitos de actividad (académico, investigación y transferencia, recursos humanos y materiales, y economía) para que esta información pueda ser reutilizada, con las únicas limitaciones que imponga la legislación, en particular la protección de datos personales o los acuerdos de confidencialidad.

2.2. Política de gobierno abierto

La *Universidad* <...> se compromete con los principios de gobierno abierto, pues los estima imprescindibles para un buen gobierno y los reconoce plenamente alineados con el espíritu universitario:

- Transparencia: oferta de información clara y actualizada, accesible y reutilizable, mediante la que se rinden cuentas de la actividad y se justifican las acciones de gobierno.
- Participación: establecimiento de canales apropiados para que la ciudadanía, informada gracias a la transparencia, pueda intervenir en los procesos de toma de decisiones.
- Colaboración: llamamiento a los ciudadanos e instituciones a compartir lo que saben y a generar soluciones conjuntamente.

2.3. Política de acceso abierto

La *Universidad* <...> se compromete con los principios en favor del acceso abierto, pues los reconoce plenamente alineados con el espíritu universitario, y en consecuencia promoverá las siguientes acciones:

- Dictar mandatos y recomendaciones dirigidos a la comunidad universitaria para que publiquen en abierto los trabajos financiados total o parcialmente por la Universidad.
- Dotar al repositorio institucional de los medios y recursos necesarios para desarrollar su función de archivo, difusión y preservación de los documentos y otros objetos digitales para su acceso en abierto por la comunidad global, garantizando el cumplimiento de la legislación relativa a los derechos de autor en los materiales depositados.
- Difundir el acceso abierto y las licencias abiertas.

3. Buenas prácticas de apertura

Aprovechando que este es un camino que ya ha sido emprendido por otras administraciones, lo que se propone, en un espíritu de reutilización, es adoptar o adaptar el decálogo general elaborado a partir del día open data en Euskadi:

1. Armonización entre Administraciones.
2. Publicar datos en formatos abiertos y estándares.
3. Usar esquemas y vocabularios consensuados y utilizar metadatos abiertos.
4. Inventario en un catálogo de datos estructurado.

5. Datos accesibles desde direcciones web persistentes y amigables.
6. Exponer un mínimo conjunto de datos relativos al nivel de competencias del organismo y su estrategia de exposición de datos
7. Compromiso de servicio, actualización y calidad del dato, manteniendo un canal eficiente de comunicación reutilizador <-> AAPP.
8. Monitorizar y evaluar el uso y servicio mediante métricas.
9. Datos bajo condiciones de uso no restrictivas y comunes.
10. Evangelizar y educar en el uso de datos.
11. Recopilar aplicaciones, herramientas y manuales para motivar y facilitar la reutilización.

3.0. Armonización entre Administraciones

Todos los puntos del decálogo se basan en la premisa de que debe existir una armonización entre todas las Administraciones. Todas las iniciativas Open Data deben compartir los mismos principios y definiciones que se listan en el decálogo. Este punto 0 es básico para la interoperabilidad y aprovechamiento eficiente de las sinergias llevadas a cabo por todos los actores Open Data - RISP. En el caso de las universidades, entendidas como distintas administraciones, naturalmente el primer nivel de armonización es entre ellas.

3.1. Publicar datos en formatos abiertos y estándares

Cualquier iniciativa Open Data debería publicar sus conjuntos de datos, así como la documentación, en formatos abiertos (no-propietarios) y que sean adecuados para permitir la reutilización de los mismos por parte del colectivo reutilizador destinatario, para lo que está disponible un catálogo de estándares.

3.2. Usar esquemas y vocabularios consensuados y utilizar metadatos abiertos

Además de los formatos abiertos y estándar, la estructura de los datos debería seguir un convenio o unos esquemas definidos, en particular la norma técnica de interoperabilidad de RISP. Si se crean vocabularios o esquemas de representación de la información específicos, éstos se deberían exponer públicamente para que el colectivo reutilizador pueda interpretar correctamente la información. Utilizar metadatos documentados y disponibles como "Metadatos Abiertos" preparados para su reutilización (Nivel 3 en el documento de ISA "Towards Open Government Metadata").

3.3. Inventario en un catálogo de datos estructurado

Cualquier iniciativa Open Data debe tener un punto de consulta donde se incluya un inventario con información descriptiva y técnica sobre los conjuntos de datos que se exponen. Los metadatos que informan sobre cada conjunto de datos debería seguir una estructura común y estándar, y en el caso de datos específicamente universitarios sería necesario consensuarla. Asimismo, se deberían compartir las taxonomías de temáticas u otras necesarias -p.e., toponimia- para clasificar los conjuntos de datos dentro de los catálogos.

3.4. Datos accesibles desde direcciones web persistentes y amigables

Tanto las fichas de los conjuntos de datos, como la distribución de la propia información (volcado en un archivo, API de consulta, RSS, etc.) deberían de estar accesibles desde URLs (direcciones web) que persistan en el tiempo y así evitar que se pierdan las referencias en el futuro. Además deben seguir una estructura homogénea y bien definida, con información legible para que los reutilizadores conozcan o "intuyan" el contenido referido por dichas direcciones web.

3.5. Exponer un mínimo conjunto de datos relativos al nivel de competencias del organismo y su estrategia de exposición de datos

Cada Administración que impulse una iniciativa Open Data debería crear una hoja de ruta donde especifique la estrategia de exposición de los conjuntos de datos y sus prioridades. Inicialmente, debería publicar los

conjuntos de mayor interés según las competencias del propio organismo. En el caso de las universidades convendría mucho que las hojas de ruta estuviesen alineadas para promover la combinación/comparación de datos y la eficiencia en el desarrollo de aplicaciones y servicios específicos.

3.6. Compromiso de servicio, actualización y calidad del dato, manteniendo un canal eficiente de comunicación entre reutilizador y AAPP

La Administración debe mantener un mínimo de calidad y servicio en su iniciativa Open Data, manteniendo lo expuesto en la estrategia de publicación y comprometiéndose con su colectivo reutilizador. Debe establecer un canal eficiente de comunicación que permita la interacción bidireccional organismo público - reutilizadores.

3.7. Monitorizar y evaluar el uso y servicio mediante métricas

La Administración debe crear métricas y evaluar sus indicadores de uso y servicio de la iniciativa Open Data. De esta forma puede monitorizar el funcionamiento y uso, y así analizar si se está cumpliendo el compromiso con la comunidad de reutilizadores y cuales son las potenciales carencias del sistema o de la estrategia.

3.8. Datos bajo condiciones de uso no restrictivas y comunes

Las condiciones de uso deberían ser lo menos restrictivas posible y permitir la reutilización libre, incluso para fines comerciales. Se recomienda la creación y uso de licencias tipo, autodocumentadas y que sean comunes entre distintas administraciones.

3.9. Evangelizar y educar en el uso de datos

Es necesario educar en el uso de los datos, tanto a los colectivos de reutilización específicos (sector TIC, periodismo, investigación, etc.) como a la sociedad en general y así fomentar el conocimiento y la inquietud por procesar información de una forma autónoma. Contar con una estrategia de comunicación para dar a conocer los datos que se vayan abriendo y potenciar su reutilización. Por ejemplo, a través de redes sociales y medios de comunicación "tradicionales". Evitar el "disgusto" por los datos.

3.10. Recopilar aplicaciones, herramientas y manuales para motivar y facilitar la reutilización

Cualquier iniciativa Open Data debería recopilar ejemplos de uso y herramientas que faciliten y motiven la reutilización de los datos que se publican.

En el caso universitario parece obvio que los primeros reutilizadores serán los propios universitarios (investigadores, tecnólogos y estudiantes, montar iniciativas hacktivistas… partiendo de lo local pero sabiendo que es posible - y muy estimulante - competir globalmente), y es fundamental que se difundan las aplicaciones que se desarrollen (en datos.gob.es, por ejemplo, hay un apartado de Aplicaciones, además de los datos). Por supuesto esto tiene una importante derivada de costes: las aplicaciones interesantes que desarrolle la universidad (o reutilizador) A basadas en datos abiertos consensuados le sirven automáticamente a todas, esto incluso cuando se trate de datos “no comparables”, es decir, además del beneficio de poder elaborar servicios combinando los datos de todos (los estudios ofrecidos por todos, la investigación realizada por todos, etc).

4. Datos universitarios susceptibles de abrirse

Es importante que las universidades muestren, desde un primer momento, el compromiso que tienen como institución con la iniciativa de datos abiertos y quizá la mejor forma de demostrarlo es poniendo a prueba su resistencia a la liberación de información sensible.

Si nuestra universidad nos pregunta ¿qué datos debo abrir? deberíamos responderle con otra pregunta ¿qué

datos podemos (poder como sinónimo de capacidad técnica y legal de hacerlo) abrir?

Deberíamos empezar por aquellos que combinen el ser el “más útil-más barato-más sensible”. Liberando información de todas las áreas de negocio de nuestra institución, de aquellas áreas de las que ya dispongamos de datos estructurados y que sean fiables. Debemos liberar datos lo más desagregados posible. Debemos mantenerlos actualizados con una frecuencia adecuada a cada caso. Y también datos que permitan generar servicios, posiblemente los datos georreferenciados de recursos y actividades en un campus sean de los más aprovechables, véase el ejemplo de Southampton.

Un resultado importante del grupo, lo que toca trabajar ahora más intensamente, será un censo inicial consensuado de catálogos de datos de todos los ámbitos (acceso, matrícula, calificaciones y graduación de estudiantes, docencia impartida, investigación realizada, recursos humanos, economía, y recursos materiales). Alcanzar este consenso aumentará el valor de los datos liberados, al facilitar su combinación y comparación, y permitirá enlazarlos técnicamente (dando calidad “cinco estrellas” a su publicación), y por otra parte permitirá que las aplicaciones y servicios que se construyan sobre los datos sean compartidas (eficiencia).

La intención es mantener este censo (inicial y evoluciones siguiendo un procedimiento ordenado y ágil) en el marco del ENI, donde seguramente habrá también información sobre las codificaciones generales (empezando por la de las propias universidades, e incluyendo titulaciones, municipios, países, etc).

Con respecto a las descripciones semánticas, deben utilizarse categorías o metadatos “genéricos”, de vocabularios ampliamente utilizados como el Data Catalog Vocabulary (DCAT) (ejemplo de su uso en en la UPF), y si el organismo tiene algunos más específicos una buena práctica es incluirlos como etiquetas. Estos datos de catalogación también es muy importante que se usen de forma coordinada (punto 2 del decálogo).

Dentro de las numerosas opciones disponibles para la selección de unos conjuntos de datos iniciales en el ámbito universitario, resulta especialmente interesante la que ofrece el SIIU (Sistema Integral de Información Universitaria). Por el momento el SIIU, además de no estar abierto, cubre únicamente datos sobre enseñanzas, la oferta y rendimiento.

Otro conjunto de datos relativamente consolidado que cubre otros ámbitos de la actividad (investigación, docencia, RRHH, economía) son las tablas solicitadas por el Tribunal de Cuentas, aceptando la exigencia de transparencia que este proceso de fiscalización supone, y llevándola más allá, entendiendo que es la sociedad entera a quien hay que rendir cuentas, mediante dicho tribunal, o/y directamente. Entre estos datos consideramos que los más interesantes en primera instancia son los datos sobre actividad investigadora, dado que las dos principales actividades de la universidad al servicio de la sociedad son la educación y la investigación y transferencia.

Por su potencial como servicio, y su potencial para originar una fértil compartición de recursos desarrollados sobre el mismo, cabe destacar los recursos materiales (edificios, aulas, laboratorios,…) que pueden georeferenciarse tomando como base la información que proporciona Openstreetmap.org (es espectacular el detalle de la información de los Campus) y herramientas similares a las que ha construido GeoSpatiumLab (spin-off UZ) para el ayuntamiento de Zaragoza. Es más, es factible organizar hacktivismos de datos para que los alumnos (y PDI y PAS) participen, con lo que se daría difusión a la iniciativa de apertura. Una vez creados los datos se puede dar difusión continuada mediante aplicaciones que "consuman" esta información.

4.0. Necesidad de normalización

Orientado a todos los datasets que se quieran publicar.
Dos criterios a tener en cuenta al redactar:
- La necesidad de normalización tiene que extenderse a los datos que consideremos de referencia, no solo para los datos espaciales; por ejemplo, habría que normalizar la información que se proporciona de cada investigador
- Una aproximación pragmática identificando potenciales modelos core sencillos (el menor común denominador entre los sistemas de información de todas las Universidades); el nomenclátor es un ejemplo, los datos del SIIU (no todos) son otro.

4.1. Datasets sobre enseñanzas: oferta y rendimiento, basados en el SIIU

El principal objetivo del SIIU es disponer de unos indicadores del Sistema Universitario Español que sean de calidad, fiables, que reflejen fielmente la realidad, y que permitan la comparabilidad. Además, deben aportar la información necesaria a todos los agentes del sistema (Ministerio, CCAA, Universidades, Estudiantes, Profesores e Investigadores,…).

Para el desarrollo del SIIU, las Universidades ya realizan un importante esfuerzo de normalización y extracción de datos en un formato predefinido y con semántica común.

Basta con aplicarle a los datos una pequeña transformación de consenso para conseguir una publicación de datos muy digna y útil con un esfuerzo razonable.

Debe tenerse en cuenta además, que los ficheros del SIIU están estructurados en formato xml, por lo que sería relativamente fácil llegar a alcanzar un nivel de apertura técnica calificada con 4*

La promoción de la reutilización de la información pública por parte de terceros es otro de los motivos que nos mueve a seleccionar un catálogo de conjuntos de datos derivados de los ficheros que se entregan al SIIU; si cada universidad seleccionara conjuntos de datos distintos, sin acuerdo previo de formatos y semántica, y aún tratándose de información similar en esencia, la comparación no sería posible. Sin embargo, seleccionando conjuntos de datos como los que se propondrán, con formatos y semántica común ya conocidos, incluso con un número pequeño de universidades que se unan a la iniciativa, pueden realizarse comparaciones entre ellas, y promover la creación de aplicaciones o servicios que utilicen esos datos.

Por último, seleccionar el SIIU como origen del catálogo parece razonable dentro del entorno de restricción de recursos en el que nos encontramos. Desarrollar “desde cero” nuevos procedimientos de extracción para nuevos catálogos sería un esfuerzo que no toda universidad podrá acometer, lo que puede constituir un obstáculo para el desarrollo uniforme de la iniciativa en el tejido universitario.

Como punto de partida para elaborar ese censo consensuado, y que resulte económico y aceptado, se revisan los ficheros del SIIU, y las eventuales necesidades de disociación.

De entre todos los ficheros del SIIU, se analizan los ficheros que forman parte del área académica por ser actualmente los datos más depurados en todas las universidades españolas, y los que menos resistencia a su apertura pueden ofrecer por parte de las universidades. De entre todos los ficheros analizados, que han sido un total de 31 ficheros, se propone que 18 ficheros formen parte del citado catálogo común y se recomienda también su apertura inicial en todas las universidades.

Estos 18 ficheros contienen los siguientes datos:

(NOTA GENERAL: la “identidad” de los estudiantes ha de ser anonimizada, es decir, cada estudiante es representado por un código disociado de su identidad personal..

Fichero 01.01. AVANCE DE MATRÍCULA EN ESTUDIOS DE GRADO (CENTROS PROPIOS)

En este fichero se encuentran las características generales de los estudiantes matriculados en estudios de grado que se imparten en centros propios de la universidad. Contiene la titulación en la que se encuentran matriculados los estudiantes.

Fichero 01.05. AVANCE DE MATRÍCULA EN ESTUDIOS DE MÁSTER (CENTROS PROPIOS)

En este fichero se encuentran las características generales de los estudiantes matriculados en estudios másteres oficiales que se imparten en centros propios de la universidad. Contiene la identificación y ubicación de másteres que cursan los estudiantes.

Fichero 01.08. AVANCE DE FORMACIÓN EN DOCTORADO RD 56/2005, 1393/2007 Y 99/2011 (CENTROS

PROPIOS)

En este fichero se encuentran las características generales de los estudiantes matriculados en el periodo de formación de estudios de doctorado en los centros propios de la universidad. Contiene la identificación y ubicación de los programas de doctorado que cursan los estudiantes.

Fichero 02.01. MATRICULA DE ACCESO AL ESTUDIO DE GRADO (CENTROS PROPIOS)

En este fichero se encuentran datos de estudiantes que se matriculan por primera vez en el grado en el curso académico actual en centros propios de la universidad. Están incluidos los estudiantes procedentes de otro grado así como los estudiantes que trasladen expediente desde planes antiguos. Contiene las titulaciones que cursan los estudiantes, datos estadísticos socio-económicos familiares, estudios con los que acceden y nota de admisión.

Fichero 03.01. RENDIMIENTO DE ESTUDIANTES MATRICULADOS EN ESTUDIOS DE GRADO (CENTROS PROPIOS)

En este fichero se encuentran datos del rendimiento de todos los estudiantes matriculados en el curso de referencia en estudios de grado impartidos en centros propios de la universidad. Contiene créditos matriculados, superados, presentados, reconocidos y transferidos para cada estudiante. Asimismo contiene también la nota media del expediente académico si el estudiante es titulado en el curso de referencia.

Fichero 03.05. RENDIMIENTO DE ESTUDIANTES MATRICULADOS EN ESTUDIOS DE MÁSTER (CENTROS PROPIOS)

En este fichero se encuentran datos del rendimiento de todos los estudiantes matriculados en el curso de referencia en estudios de Máster impartidos en los centros propios de la universidad. Contiene datos de residencia familiar y del estudiante, identificación de los máster e información sobre créditos matriculados, superados, presentados, reconocidos y transferidos. Asimismo contiene también la nota media del expediente académico si el o la estudiante es titulado/a en el curso de referencia.

Fichero 03.08. FORMACIÓN EN DOCTORADO 99/2011, 56/2005 y 1393/2007 (CENTROS PROPIOS)

En este fichero se encuentran datos del rendimiento de los estudiantes matriculados en el periodo de formación de estudios de doctorado en los centros propios de la universidad. Contiene datos de residencia familiar y del estudiante, acceso al doctorado, identificación de los doctorados e información sobre créditos matriculados y superados.

Fichero 03.10. LECTURA DE TESIS (CENTROS PROPIOS)

En este fichero se encuentran datos sobre los doctorandos que han leído la tesis doctoral en el curso de referencia en los centros propios de la universidad. Se incluirán los doctorandos procedentes de los doctorados 99/2011, 56/2005, Y 1393/2007. Contiene información sobre residencia familiar y personal del estudiante, identificación de la universidad que otorgó la suficiencia y calificación de la tesis.

Fichero 04.01. ENTRADA EN EL SISTEMA UNIVERSITARIO ESPAÑOL CON PROGRAMA DE MOVILIDAD

En este fichero se encuentran datos de todos los estudiantes procedentes de instituciones de educación superior extranjeras que están matriculados en el curso actual en la universidad con algún programa de movilidad. Contiene, entre otros, datos de identificación de los estudiantes, universidad de procedencia, tipo de programa de movilidad y fechas de inicio y finalización. Este fichero debe ser anonimizado para su uso por entidades externas a la universidad.

Fichero 04.02. SALIDA A UNA UNIVERSIDAD EXTRANJERA CON PROGRAMA DE MOVILIDAD

En este fichero se encuentran datos de todos los estudiantes de cualquier ciclo formativo (1er y 2º ciclo, grado, máster y doctorado) que están matriculados/as en la universidad y en el curso actual cursan parte de sus estudios en una institución de educación superior extranjera con algún programa de movilidad. Contiene

universidad de destino, tipo de programa de movilidad y fechas de inicio y finalización.

Fichero 04.03. SALIDA A OTRA UNIVERSIDAD ESPAÑOLA CON PROGRAMA DE MOVILIDAD

En este fichero se encuentran datos de todos los estudiantes de cualquier ciclo formativo (1er y 2º ciclo, grado, máster y doctorado) que están matriculadas/os en la universidad y en el curso actual cursan parte de sus estudios en otra universidad española con algún programa de movilidad. Contiene la universidad de destino, tipo de programa de movilidad y fechas de inicio y finalización.

También se propone la apertura inicial de ficheros auxiliares del SIIU como aquellos ficheros que contienen datos relacionados con la estructura como los centros, departamentos u otras unidades como escuelas de doctorado, centros de investigación, etc. que forman parte de la universidad y estudios de Grado, Máster, Doctorado y programaciones conjuntas que se imparten en la universidad:

Fichero 01.01. LISTADO DE CENTROS

En este fichero se encuentran datos sobre las características de los centros asociados a la universidad, en este fichero sólo se recogen las escuelas y facultades. Contiene, entre otros, datos de los nombres y ubicación de los centros.

Fichero 01.02. LISTADO DE DEPARTAMENTOS

En este fichero se encuentran datos sobre las características de los departamentos que forman parte de la universidad. Contiene, entre otros, datos de los nombres y ubicación de los departamentos.

Fichero 01.03. LISTADO DE OTRAS UNIDADES

En este fichero se encuentran datos sobre nombre, ubicación y tipo de aquellas otras unidades que forman parte de la universidad. Contiene, entre otros, datos de los nombres, ubicación de las unidades y CIF.

Fichero 02.01. ESTUDIOS DE GRADO

En este fichero se encuentran datos de los estudios de grado regulados por el R.D. 1393/2007 que oferta la universidad. Contiene, entre otros, datos de los nombres de las titulaciones, centros de impartición, ramas de enseñanzas, número de créditos necesarios y número de créditos ofertados.

Fichero 02.02. ESTUDIOS DE MÁSTER

En este fichero se encuentran datos de los estudios de máster regulados por el R.D. 1393/2007 y por el R.D. 56/2005 que oferta la universidad. Contiene, entre otros, datos de los nombres de los másteres, unidades responsables, ramas de enseñanzas, número de créditos necesarios y número de créditos ofertados.

Fichero 02.03. ESTUDIOS DE DOCTORADO

En este fichero se encuentran datos de los estudios de doctorado regulados por el por el R.D. 99/2011, R.D. 56/2005 y por el R.D. 1393/2007 que oferta la universidad. Contiene, entre otros, datos de los nombres de los programas de doctorado, unidades responsables y ramas de enseñanzas.

Fichero 02.04. PROGRAMACIÓN CONJUNTA DE ESTUDIOS OFICIALES DE GRADO

En este fichero se encuentran datos de las programaciones conjuntas de estudios oficiales de grado que oferta la universidad. Contiene, entre otros, datos de los nombres de los programas conjuntos, centros responsables, ramas de enseñanzas asociadas, número de créditos necesarios y número de créditos ofertados.

En una primera fase, para la puesta en marcha inicial, sería conveniente seleccionar un número reducido de estos ficheros por los cuales comenzar el proceso de apertura y que permitan hacer algún tipo de análisis útil. Por tanto, se proponen los 6 conjuntos de datos siguientes:

Fichero 02.01. ESTUDIOS DE GRADO

Fichero 02.02. ESTUDIOS DE MÁSTER

Fichero 01.01. AVANCE DE MATRÍCULA EN ESTUDIOS DE GRADO (CENTROS PROPIOS)

Fichero 01.05. AVANCE DE MATRÍCULA EN ESTUDIOS DE MÁSTER (CENTROS PROPIOS)

Fichero 03.01. RENDIMIENTO DE ESTUDIANTES MATRICULADOS EN ESTUDIOS DE GRADO (CENTROS PROPIOS)

Fichero 03.05. RENDIMIENTO DE ESTUDIANTES MATRICULADOS EN ESTUDIOS DE MÁSTER (CENTROS PROPIOS)

4.2. Datasets sobre actividad investigadora

Alguna consideración inicial:

- Habrá datos importantes que no abriremos (nosotros) porque no nos pertencen, pero que es altamente recomendable combinar con los datos que abrimos: índices de impacto, posición relativa de las revistas en las categorías, citas a los artículos, entidades financiadoras...
- Posiblemente no se incluyan en un dataset parte de los datos disponibles, cuando no sean públicos. Por ejemplo no incluir la cuantía de contratos que no sean públicos, o incluirla de forma agregada en un "contrato" sin investigadores.
- Se identificarán nominalmente los investigadores (son autores de PUBLICaciOneS, y son trabajadores públicos), preferentemente mediante su ORCID, pero no exclusivamente porque (todavía) no es universal (usar un identificador propio prefijado por el código de la universidad, por ejemplo).

Conceptualmente, básicamente se incluyen cuatro entidades:

Investigador

Un *Investigador* es toda persona que realice una labor investigadora reconocida por alguna entidad. Este concepto incluye desde personal investigador permanente, hasta investigadores en formación. Todo investigador tiene un identificador único (atributo *uid*) y opcionalmente un nombre (atributo *nombre*).

Entidad

Una *Entidad* representa a una colección de personas organizadas formando una comunidad u otra estructura social. Toda entidad tiene un identificador único (atributo *uid*) y opcionalmente un nombre (atributo *nombre*). El término es amplio ya que los investigadores pueden formar parte de comunidades cuyo propósito no es la investigación, sino, por ejemplo, la docencia. Normalmente se utilizarán especializaciones de este concepto como por ejemplo Universidad, Departamento, Instituto Universitario de Investigación o Grupo de Investigación. La relación *miembro* se utiliza para relacionar investigadores con entidades. Un investigador puede ser miembro de varias entidades de forma simultánea. Se puede representar de forma explícita la duración de dicha relación (atributos *desde* y *hasta*) y se podría incluir el papel en dicha organización (atributo *rol*).

Proyecto

Un *Proyecto* es la representación de las ayudas públicas, contratos y convenios para el desarrollo de líneas de investigación, servicio, asesorías y otros que los investigadores pueden participar. Todo proyecto tiene un identificador único (atributo *uid*) y opcionalmente un título descriptivo (atributo *título*). Como atributos (según sea el tipo del proyecto) pueden incluirse la financiación, duración, entidad financiadora (y su ámbito o categoría, y eventualmente relacionarla con el exterior).

La relación *participa* se utiliza para relacionar investigadores con proyectos. Un investigador puede participar en varios proyectos de forma simultánea. Se puede representar de forma explícita la duración de dicha relación (atributos *desde* y *hasta*) y el papel en dicho proyecto (atributo *rol*, en particular, si es investigador principal del proyecto).

Resultado

Un *Resultado* es la representación de cualquier resultado de investigación disponible a partir de un momento determinado (atributo *fecha*). Todo resultado tiene un identificador único (atributo *uid*) y opcionalmente un título descriptivo (atributo *título*). Normalmente se utilizarán especializaciones de este concepto como por ejemplo Artículo y Patente.

Como atributos (según sea el tipo de resultado) interesan por ejemplo la fuente (e.g., revista, y de esta su factor de impacto y su mejor posición relativa en la lista de categorías - esto podría ser un dato externo, o no incluirse ya que que es un dato licenciado, en cuyo caso de querer usarse habría de combinarse con la fuente autorizada) y el número de citas (nuevamente es un dato externo, que idealmente se ha de obtener independientemente, por ejemplo de google scholar, y combinarse si se quiere utilizar).

Los resultados de investigación son generados por uno o varios investigadores.

4.3 Datasets sobre equipamientos, infraestructuras y mapas

A la hora de hablar sobre datos abiertos de naturaleza espacial o geográfica se debe comentar que la comunidad GIS tiene una amplia tradición en la liberalización de datos. Ejemplos de ello son los esfuerzos en estandarización de formatos realizados por la OGC (Open Geospatial Consortium) o el portal de acceso a la información geográfica (geoportal) del Gobierno de España. De esa manera resulta adecuado definir los objetivos de un portal de datos abiertos para una universidad y de un geoportal universitario, además de situar a cada uno en una infraestructura común:

- Un portal de datos abiertos para una universidad debe ser una fuente de datos de diferente procedencia (docencia, investigación, etc.). El portal de datos abiertos puede actuar como interfaz para el acceso a una base de datos espacial.
- Un geoportal universitario sería uno de los principales reutilizadores de los datos suministrados por un portal de datos abiertos con una alta componente geográfica, es decir, el objetivo es situar diversos datos universitarios como capas en un mapa. Para ello también accede a una base de datos espacial, además de obtener datos temáticos del portal de datos abiertos.

Uno de las claves de la información geoespacial reside en el formatos elegido para su representación. En concreto, los formatos espaciales elegidos deben permitir maximizar la reutilización de los datos. Además, esta reutilización vendrá condicionada por el nivel de conocimiento de bases de datos espaciales que tenga el potencial reutilizador, no teniendo las mismas necesidades un reutilizador que simplemente quiera dibujar algo sencillo en un mapa de un campus universitario que aquel que quiera elaborar una aplicación para el cálculo de rutas entre edificios de un campus. Por ello, se considera conveniente establecer los formatos de datos espaciales en dependencia de diferentes perfiles de usuarios, es decir, la elección de los formatos viene dada por la audiencia que tendrá el portal de datos abiertos. Además, cobra especial importancia la diferencia entre un servicio que ofrece datos espaciales y la descarga de un archivo de datos con la información geográfica. Siguiendo estas directrices se pueden proponer dos tipos de formatos:

- GML (Geography Markup Language): lenguaje basado en XML para modelar sistemas geográficos y para el intercambio de datos geográficos.
- KML (Keyhole Markup Language): lenguaje de basado en GML para representar datos geográficos. Es muy popular por su uso en Google Earth.
- WFS (Web Feature Service): servicio que permite la realización de peticiones de información geográfica en diversos formatos. Útil para reutilizadores "avanzados".

En cuanto a la tipología de datos espaciales susceptibles a ser abiertos en una universidad, existe una amplia casuística desde datos acerca de las propias estancias de la universidad (coordenadas, áreas, uso dado, etc.) hasta consumos energéticos y de recursos (electricidad, gas, agua, etc.) pasando por datos sobre el material inventariable.

5. Ejemplos de reutilizaciones de estos datos

En este apartado se describen a modo de ejemplos o sugerencias algunas aplicaciones/reutilizaciones vistas en universidades:

- **Simplificadores**. Aplicaciones y soluciones que facilitan el acceso a información compleja resumiendola en una forma más simple. *Ejemplos: College Affordability and Transparency Center (EE.UU.), studentaid.gov (EE.UU.), Análisis de alumnos de nuevo ingreso (U. Pablo de Olavide), Vacancy TreeMap (U. Oxford)*
- **Directorios y agendas**. Aplicaciones y soluciones que permiten localizar ofertas y demandas de recursos y servicios (personas, eventos, edificios, salas, equipamiento, etc.). Pueden presentar la información sobre un mapa para facilitar la localización. *Ejemplos: iSoton (U. Southampton), Campusplan (U. Münster), University Science Area Map (U. Oxford), Portal de datos (Open University)*
- **Herramientas de gestión**. Aplicaciones y soluciones equivalentes a los (pequeños) aplicativos de gestión corporativos, pero construidas sobre los datos abiertos, en lugar de directamente sobre los sistemas corporativos. *Ejemplos: Room Finder (U. Southampton).*

6. Recomendaciones técnicas

En el presente apartado trataremos de fijar las bases técnicas iniciales que nos permitan ofrecer un soporte adecuado a la hora de publicar nuestros conjuntos de datos (datasets).

Pero, antes de entrar en las tecnologías y herramientas, organizadas de forma temática, es importante detenerse a reflexionar sobre la proporcionalidad entre los medios y los fines, o entre las soluciones y los problemas. En el caso que nos ocupa, la "dimensión del problema" viene determinada por dos medidas:

- Número de colecciones de datos (datasets)
- Frecuencia de actualización o refresco

De forma orientativa, tentativa, y obligadamente difusa (especialmente en sus fronteras) podríamos hablar de tres niveles de complejidad creciente para el desarrollo de un portal de datos abiertos:

1. Hasta 100 datasets, hasta 10 datasets requieren actualización semanal.
2. Hasta 1000 datasets, hasta 100 datasets requieren actualización semanal.
3. Por encima de 1000 datasets o de 100 que requieran actualización semanal.

El Nivel 1 puede describir las necesidades de un prototipo de portal de Datos Abiertos, o las necesidades "en explotación" de una Universidad de tamaño pequeño, mientras que el Nivel 3 puede ser el que corresponde a las grandes Universidades, grupos de Universidades (e.g., todas las catalanas bajo UNEIX, el Campus Iberus), o a un nodo central (e.g., CRUE).

A la hora de analizar las diferentes soluciones tecnológicas en cada escenario que se presentan hay que tener en cuenta que son de carácter orientativo, es decir, no son las únicas posibles, y los sistemas indicados son sólo ejemplos para comprender mejor cómo es la solución propuesta. Además, lo que se describe a continuación es una infraestructura orientativa básica para cada escenario que puede ser extendida. Por ejemplo el nivel más básico (Nivel 1) puede ser complementado con la creación de Web APIs que usan los datos publicados y cuyas URL se indican en los metadatos de dichos datos .

Nivel 1: menos de 100 colecciones con menos de 10 colecciones actualizadas cada semana. En este caso la solución más adecuada es la construcción de un portal Web convencional utilizando un servidor Web de uso habitual (Apache, Nginx, IIS). Este portal se caracterizaría por:

- **Sitio web**. Un sitio web convencional sobre un servidor Web de uso general.
- **Datos publicados**. Estarían almacenados en formatos estructurados abiertos o de uso generalizado dentro del espacio del servidor Web, en una nube pública (S3), o en portales temáticos (geoportal,

bibliotecas). En este último caso sólo se permite si hay compromiso de actualización por parte de un tercero.

- **Metadatos publicados**. Estarían como microdatos, microformatos o RDFa en las páginas HTML estáticas que enlazan a los datos publicados. Debe verificarse que los motores de búsqueda comerciales como Google y Bing son capaces de extraerlos para asegurar la visibilidad del dato publicado.
- **Catálogo de datos**. Formado por las páginas HTML anteriores. Debe asegurarse que el catálogo puede ser recorrido completamente vía enlaces para facilitar su indexación por motores de búsqueda. La búsqueda en el catálogo se implementa mediante herramientas como Apache Sorl y Google Site Search.
- **Datos accesibles**. Los datos tienen una URL accesible y persistente con independencia de dónde estén almacenados. Las páginas HTML con sus metadatos tienen una URL accesible y persistente. Los ficheros recuperados están en formatos estructurados abiertos o de uso generalizado.
- **Compromiso de actualización**. Actualización manual de los datos, metadatos y catálogo de datos por un equipo del Servicio de Informática de la Universidad.
- **Realimentación**. Las páginas HTML incluyen la posibilidad de enviar comentarios a los administradores.
- **Monitorización**. Ya sea a nivel de servidor Web y a nivel de las páginas HTML se ha habilitado algún mecanismo básico de monitorización de acceso y descarga, como los logs de acceso del servidor Web o Google Analytics.

Nivel 2: 100-1000 colecciones o con 10-100 colecciones actualizadas cada semana. En este caso la solución más adecuada pasa por la construcción de un portal Web basado en algún tipo de CMS (e.g., Drupal, Wordpress, Joomla) que dé soporte al catálogo de datos. Entendemos que una colección de datos es un tipo algo especial de contenido, cuya publicación puede ser realizada mediante un CMS. Las diferencias con el Nivel 2 se indican en color verde. Este portal se caracterizaría por:

- **Sitio web**. Un sitio web basado en el CMS.
- **Datos publicados.** Estarían almacenados en formatos estructurados abiertos o de uso generalizado dentro del espacio del servidor Web, en una nube pública (S3), o en portales temáticos (geoportal, bibliotecas). En este último caso sólo se permite si hay compromiso de actualización por parte de un tercero.
- **Metadatos publicados.** Estarían almacenados en una base de datos, siendo servidos dinámicamente como microdatos, microformatos o RDFa en las páginas HTML estáticas que enlazan a los datos publicados. Debe verificarse que los motores de búsqueda comerciales como Google y Bing son capaces de extraerlos para asegurar la visibilidad del dato publicado. Considerar el almacenamiento de dichas páginas HTML en una caché para aumentar el rendimiento del sistema.
- **Catálogo de datos.** Formado por las páginas HTML anteriores. Debe asegurarse que el catálogo puede ser recorrido completamente vía enlaces para facilitar su indexación por motores de búsqueda. La búsqueda en el catálogo se implementa mediante una búsqueda sobre la base de datos donde están los metadatos y/o herramientas como Apache Sorl o Google Site Search.
- **Datos accesibles.** Los datos tienen una URL accesible y persistente con independencia de dónde estén almacenados. Las páginas HTML con sus metadatos tienen una URL accesible y persistente. Los ficheros recuperados están en formatos estructurados abiertos o de uso generalizado. Se verifica de forma automática que no hay enlaces rotos a los datos.
- **Compromiso de actualización.** Actualización manual de los datos y la base de datos de metadatos por un equipo del Servicio de Informática de la Universidad.
- **Realimentación.** Hay disponibles herramientas básicas de gestión de la comunidad de usuarios.
- **Monitorización.** Hay disponibles herramientas de monitorización de acceso y descarga de datos.

Nivel 3: Más de 1000 colecciones o con más de 100 colecciones actualizadas cada semana. En este caso identificamos como solución más adecuada la construcción de un portal Web basado en algún tipo de CMS avanzado y opcionalmente un sistema especializado en servir datos (e.g., CKAN, Virtuoso Universal Server). Las diferencias con el Nivel 2 se indican en color verde. Este portal se caracterizaría por:

- **Sitio web**. Un sitio web basado en el CMS que ofrece diferentes Web APIs sobre los datos y metadatos publicados (acceso, visualización, integración).
- **Datos publicados.** Estarían almacenados en formatos estructurados abiertos o de uso generalizado dentro del espacio del servidor Web, en una nube pública (S3), en un sistema especializado en servir datos o en portales temáticos (geoportal, bibliotecas). En este último caso sólo se permite si hay compromiso de actualización por parte de un tercero.
- **Metadatos publicados.** Estarían almacenados en una base de datos especializada (e.g., la del propio CMS, Virtuoso, Neo4j), siendo servidos dinámicamente como microdatos, microformatos o RDFa en las páginas HTML estáticas que enlazan a los datos publicados. Debe verificarse que los motores de búsqueda comerciales como Google y Bing son capaces de extraerlos para asegurar la visibilidad del dato publicado. Considerar el almacenamiento de dichas páginas HTML en una caché para aumentar el rendimiento del sistema.
- **Catálogo de datos.** Formado por las páginas HTML anteriores. Debe asegurarse que el catálogo puede ser recorrido completamente vía enlaces para facilitar su indexación por motores de búsqueda. La búsqueda en el catálogo se implementa mediante una búsqueda sobre la base de datos donde están los metadatos y/o herramientas como Apache Sorl o Google Site Search.
- **Datos accesibles.** Los datos tienen una URL accesible y persistente con independencia de dónde estén almacenados. Las páginas HTML con sus metadatos tienen una URL accesible y persistente. Los ficheros recuperados están en formatos estructurados abiertos o de uso generalizado. Se verifica de forma automática que no hay enlaces rotos a los datos.
- **Compromiso de actualización.** Delegación de la incorporación y actualización de los datos al sistema mediante formularios, flujos de trabajo, validación automática y acceso controlado basado en roles. Los metadatos pueden ser adquiridos y actualizados automáticamente a partir de la información publicada en otros catálogos de datos remotos.
- **Realimentación.** Herramientas avanzadas de gestión de la comunidad de usuarios que incluyen la posibilidad de comunicarse con el propietario real del dato.
- **Monitorización.** Hay disponibles herramientas de monitorización de acceso y descarga de datos.

6.1. Estructura de los datasets

Cada una de las unidades de datos definidas debería de tener la siguiente estructura con carácter general:

- El tema o categoría al que pertenece y etiquetas, que ayuden a su clasificación. Esta clasificación, debido a la variabilidad y cambios que pueden surgir, debería verse más como una asignación de marcas o tags. Un dataset podrá recibir uno o más tags que permitirán su búsqueda y clasificación de forma más flexible.
- Un resumen del recurso: una frase corta y comprensible por cualquiera.
- La fecha o desde cuándo este dataset está disponible (cuando se publicó por primera vez).
- La fecha de la última actualización (podría también considerarse mantenerse público el histórico, cuando sea adecuado).
- La frecuencia con la que el dataset es actualizado.
- La extensión en el tiempo o duración del dataset (por ejemplo: calificaciones del curso académico 2012/13 o presupuestos del año 2012) Se podrían definir unos valores enumerados iniciales: Anual, mensual, semanal o diario.
- Toda la documentación que se considere relevante y sirva para interpretar ese dataset de la forma menos ambigua posible, en particular diccionario de datos.
- Los estándares o formatos en los que ese conjunto de datos ha sido entregado, con referencias a un lugar genérico que detalle su tratamiento.
- La procedencia o quiénes han contribuido (unidades administrativas, por ejemplo) a la generación de ese dataset. Quizá podría ser interesante introducir la figura del “responsable de la información”, el cual puede ser un organismo, una unidad o un servicio.
- La versión (si en algún caso ha sufrido modificaciones con respecto a la estructura, relacionándolo con

los datasets de versiones anteriores).

- Un identificador único de recurso. Sólo si esto se refiere a su URI.
- Si es referenciado o tiene referencias a otros datasets. Sería aconsejable ofrecer una lista de vocabularios junto con la referencia de donde acceder para obtener una descripción del mismo.
- Lenguaje (para el caso de recursos que se publiquen en multi-idioma).
- La licencia (por ejemplo http://opendatacommons.org/licenses o la denominada *modalidad general básica* del art 8.1 del RD 1495/2011).

6.2. Formatos

El propósito de este apartado es tener en consideración tanto los formatos que permitan exportar datos enlazados y sus vocabularios relacionados, como aquellos enfocados a la simple publicación de datos planos acorde al esquema de 5* en cuanto a la calidad de los datos se refiere.

Los siguientes formatos serían formatos admisibles para codificar información conforme al esquema 4* y 5* al estar estandarizados por W3C (o prácticamente estandarizados), ser abiertos y basados en RDF.

- **RDF/XML**: El primer formato de intercambio de RDF en XML. Formato de referencia para Tim Berners-Lee.
- **Formatos SPARQL**: Los formatos estándar en los que devuelve la respuesta un servicio SPARQL 1.0 (XML) o 1.1 (XML, JSON, CSV, TSV).
- **N-Triples** (recomendación candidata): Formato de serialización en líneas de texto plano de RDF. Muy popular para intercambio de grandes colecciones de datos (e.g. dbpedia).
- **N-Quads** (recomendación candidata): Formato de serialización en líneas de texto plano de RDF que permite identificar la pertenencia de una sentencia RDF a una determinada colección. Popular para intercambio de grandes colecciones de datos (e.g. dbpedia).
- **Turtle** (recomendación candidata): Formato de serialización en texto plano. compatible con N-Triples pensado para ser leído por humanos. Muy popular.
- **TriG** (recomendación candidata): Extensión de Turtle que permite representar el concepto de un dataset en RDF.
- **Formatos OWL 2**: OWL/XML formato de serialización en XML; Functional Syntax formato de serialización en texto plano; Manchester Syntax formato de serialización en texto plano de pensado para ser leído por humanos. De muy reciente creación.
- **JSON-LD** (recomendación propuesta): Perfil de JSON capaz de intercambiar RDF. De muy reciente creación.

Los siguientes formatos se pueden utilizar para exportar datos enlazados y sus vocabularios relacionados, pero todavía no han sido estandarizados o no están basados en RDF. Serían pues formatos conformes al esquema 3*:

- **Familia RDF**: N3 (sometido), Binary RDF (sometido), Trix.
- **JSON**: Formato de intercambio de datos muy popular en APIs abiertos estandarizado por IETF.
- **XML**: Formato de intercambio de datos estandarizado por W3C.

Un punto de vista complementario al presentado son los formatos recogidos en la Norma Técnica de Interoperabilidad de Catálogo de estándares. Esta norma define un catálogo que entre otros recoge formatos estándar abiertos o de uso generalizado que se consideran necesarios para asegurar los aspectos más prácticos y operativos de la interoperabilidad entre las administraciones públicas y con el ciudadano. Extender este uso al ámbito de la apertura de la Universidad hacia la sociedad puede ser razonable. En particular, la versión actual del catálogo recoge los siguientes formatos para integración de datos: XML, RDF/XML

(implícito), Formatos SPARQL (implícito), Turtle, Formatos OWL2 (implícito), N3. Notar que en otra sección identifica formatos de fichero entre los cuales hay formatos estructurados susceptibles de usarse para intercambiar datos: CSV, GML (ISO 19136: XML para Cartografía vectorial y Sistemas de Información Geográfica) y OpenDocument (ISO 26300: XML para documentos).

Otro mecanismo a considerar en lo que se refiere a publicación abierta de datos, es la posibilidad de añadir cierto marcado al HTML ya existe en los portales de información actuales con el fin de darle semántica y enriquecer su contenido. Esto permitiría su procesamiento por terceras partes o por alguno de los buscadores existentes:

- RDFa: Enriquecimiento de documentos con marcado RDF previamente definido (incluído en la NTI de Catálogo de estándares).
- Schema.org: Aplicación de ciertos marcadores semánticos ya definiedos en su especificación pública y que permiten el mejor tratamiento de datos personales, eventos, etc.

Finalmente, W3C ha definido un mecanismo estándar denominado GRDDL que permite introducir marcado semántico en un documento XML (incluyendo páginas XHTML) y transformarlo en RDF mediante XSLT.

6.3. Almacenes de información enlazada

Almacenamiento de tripletas: Triple Store

Inicialmente se han evaluado las siguientes soluciones de código abierto para el almacenamiento de datos enlazados:

	Version	SPARQL	OWL	Licencia
OWLIM-Lite	5.3	1.1	Sí	Gratuita, no libre
AllegroGraph	4.1	1.1	Parcial	Gratuita para < 5M tuplas, no libre
4store	1.1.5	1.1	No	GPLv3
Virtuoso	6.1.6	1.1	Sí	GPLv2

Otras opciones a valorar: Sesame Store, Apache Jena TDB, BigData o Mulgara.

Referencias y comparativa entre los sistemas de almacenamiento propuestos:

- http://en.wikipedia.org/wiki/Triplestore
- http://docs.lib.purdue.edu/cgi/viewcontent.cgi?article=1046&context=techmasters
- http://www.w3.org/wiki/RdfStoreBenchmarking

Según la evaluación realizada y haciendo prevalecer como criterios la flexibilidad, rendimiento, soporte de estándares y licencia abierta, recomendamos el uso de Virtuoso como soporte para la persistencia de tripletas RDF.

Bases de datos de grafo

Como nota adicional con respecto a la persistencia de información, cabe decir que las bases de datos de grafo ofrecen un mecanismo mucho más general a la hora de relacionar información que los almacenes de tripletas,

los cuales están diseñados para trabajar exclusivamente con el metamodelo de RDF.

Existe la posibilidad de dar el mismo soporte con una base de datos de grafo, que con un almacén de tripletas, con la ventaja de poder relacionar información y publicarla en distintos formatos de una forma más general y abierta.

En este sentido, recomendamos el uso de bases de datos como Neo4j y su integración con RDF mediante Sail.

6.4. Lenguajes de consulta

Posiblemente, SPARQL es a día de hoy el estándar de facto a la hora de realizar consultas sobre conjuntos de datos enlazados. Es por esto que adicionalmente a la posibilidad de descarga directa de los datasets, es muy recomendable ofrecer un endpoint de consulta sobre los propios datos.

Por otra parte, también es interesante ofrecer un servicio de búsqueda full-text que permita buscar sobre el contenido en si. Las búsquedas full-text pueden de igual forma resultar interesantes a la hora de navegar por la información siempre que se defina un facetado adecuado de la información. Motores de búsqueda como SOLr o ElasticSearch, los cuales se basan en Apache Lucene, pueden ayudarnos a la hora de definir las facilidades de búsquedas full-text, siendo además los más utilizados en la mayoría de los entornos documentales.

También es de gran utilidad la posibilidad de recuperar toda la información relacionada a un identificador cualesquiera. Este servicio puede ofrecerse de manera conjunta al soporte a la dereferenciación en datos enlazados de todos los identificadores HTTP publicados por la organización.

Finalmente, con el objeto de facilitar el desarrollo de soluciones de integración, sería deseable ofrecer un servicio de conversión de etiquetas en identificadores. Un ejemplo de estas consultas lo ofrece el Reconciliation Service API desarrollado para OpenRefine (ex-GoogleRefine).

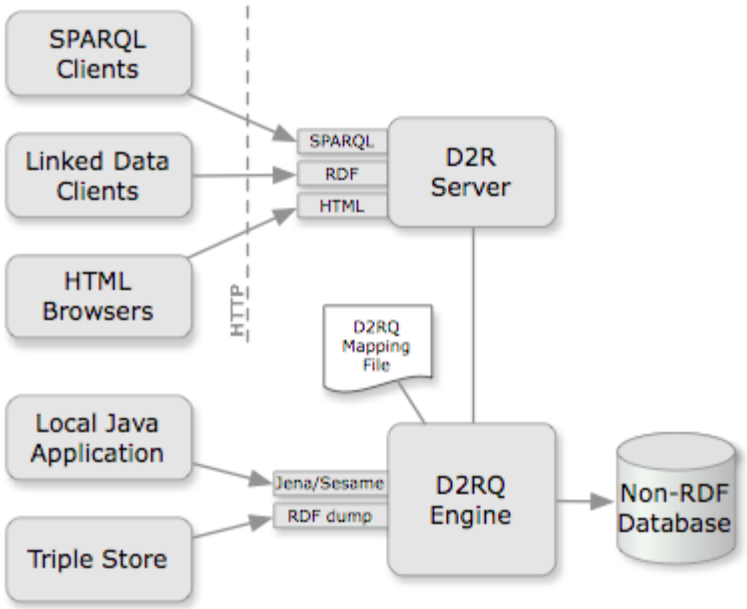
Un buen ejemplo de todas las anteriores aproximaciones es el portal de datos de Ordnance Survey (UK).

6.5. Frameworks

Existen en la actualidad varios frameworks que podemos considerar a la hora de desplegar nuestra plataforma de publicación de datos abiertos. De entre ellos, cabe destacar:

- D2R Server: A tool for publishing relational databases as Linked Data

  http://d2rq.org/



- Callimachus: Framework de trabajo integrado con Sesame Storage, edición visual de recursos publicados y soporte para búsquedas con SPARQL.

  http://callimachusproject.org/index.xhtml?view

- CKAN: (http://ckan.org/) es un framework completo para la publicación, edición y gestión de datos abiertos desarrollado por la Open Knowledge Fundation con soporte para Linked Data, SPARQL y RDFa. Destaca su interoperabilidad con catálogos de información estándar (Z39.50, CSW) y catálogos Web legados, las herramientas de búsqueda que proporciona, el soporte a la información geoespacial (tanto en búsqueda como en visualización), la inclusión de herramientas sociales, y la posibilidad de actuar como almacén de los datos publicados. publicdata.eu y data.gov.uk son casos de éxito de CKAN.

Como complemento a estos entornos o frameworks, existe la posibilidad de integrar ciertas librerías de utilidad en nuestros desarrollos, con el fin de trabajar de una forma más cómoda con los distintos formatos, conectar con almacenes de datos o realizar búsquedas. Algunos ejemplos destacables son:

- Tinkerpop blueprints: Interfaz genérica para el acceso a distintas implementaciones de base de datos de grafo. Incorpora un storage propio por defecto llamado blueprints, pero puede utilizar el de cualquier tipo de base de datos de grafo existente.

  http://www.tinkerpop.com/

- Spring Data REST HATEOAS: Sencilla exportación de entidades de datos vía REST.

  http://www.springsource.org/spring-data/rest

6.6. Referencias

Algunos sitios de referencia que pueden ser tomados como ejemplo a la hora de construir el punto de entrada al portal de publicación de datos:

http://lodum.de/

http://www.upo.es/datos-abiertos/#.USpiJx3gmAg

http://data.southampton.ac.uk/

http://datos.gob.es/datos/

http://opendata.aragon.es/

http://datos.fundacionctic.org/sandbox/catalog/faceted/

http://www.datosabiertos.jcyl.es/

http://opendata.euskadi.net/w79-home/es/

Algunos ejemplos de portales universitarios, en niveles muy dispares de desarrollo, no necesariamente para servir como ejemplo (a seguir) todos ellos:

http://data.upf.edu/similar_data_portals

6.7. Integración con los sistemas de información existentes

En el caso concreto de la UJI, es la empresa 4TIC la que está realizando el desarrollo de una capa abstracta intermedia que permita realizar ETL de datos desde los sistemas de información existentes, con el fin de nutrir a nuestra plataforma de publicación abierta con los datos y las relaciones necesarias. Posteriormente, la información extraída es almacenada para su consulta enlazada, explotación o publicación en el portal de OpenData o a través de la capa REST que expone las entidades principales.

Esta capa intermedia permite la extracción de información en distintos formatos de una forma sencilla, ofreciendo una serie de conectores especializados para los orígenes más habituales, siendo extremadamente simple incorporar nuevos orígenes. De esta manera, no es necesario instalar herramientas adicionales cuya complejidad es en muchos casos bastante elevada.

6.8. Catalogación

Se podrían contemplar varias posibilidades, desde sitios propios de cada Universidad, de las comunidades autónomas, de la CRUE, nacional... siempre siguiendo unas reglas o principios de armonización. Las herramientas de federación o alojamiento y exportación de catálogos que proporciona datos.gob.es parecen la solución más indicada, tras la fructíferea colaboración entre la CRUE-TIC y Red.es.

7. Conclusiones

Nos encontramos en los inicios de un probable camino hacia el gobierno abierto, empezando por la transparencia, que se practica idealmente mediante la apertura de datos. Esta nueva forma de relación entre las instituciones, al menos las públicas, y la ciudadanía está siendo poco a poco demandado por ésta, pero sobre todo ha sido identificada como necesaria, y beneficiosa, en algunas instancias políticas supranacionales o por algunas administraciones nacionales o locales pioneras.

Las universidades, como administraciones públicas que también son, no debieran quedar rezagadas en este camino, complementario a la apertura del conocimiento que está en su naturaleza, y que también está adquiriendo nuevas formas (acceso abierto, cursos abiertos en línea, software de fuentes abiertas, etc).

La autonomía universitaria no debería perjudicar este importante avance, y para ello es necesario un esfuerzo de coordinación que está en el corazón de todas las iniciativas de apertura en las administraciones, en las que se requiere armonización y cumplimiento de normas de interoperabilidad. Esta armonización debiera ser especialmente estrecha en el caso de las universidades, pues daría más valor a la reutilización de sus datos y

redundaría en mayor eficiencia.

La coordinación, que se quiere promover desde la sectorial TIC de la CRUE, alcanza también a las soluciones técnicas, adaptadas a cada caso, de forma que no se tengan que replicar los esfuerzos de análisis y diseño de soluciones, y se favorezca la colaboración, que en definitiva es la aspiración más elevada del gobierno abierto.