



Digital Repository Infrastructure Vision for European Research

Directrices para proveedores de contenido

*Exposición de recursos textuales con el
protocolo OAI-PMH*

Aplicación piloto

Versión 1.0

Colaboradores

Martin Feijen, Wolfram Horstmann, Friedrich Summann, Muriel Foulonneau, Karen
Van Godtsenhoven, Patrick Hochstenbach, Paolo Manghi, Bill Hubbard



Acerca del proyecto DRIVER

Qué es DRIVER

Las siglas DRIVER corresponden a *Digital Repository Infrastructure Vision for European Research*, proyecto mediante el cual un consorcio financiado por la Comunidad Europea está creando un marco de trabajo organizativo y tecnológico para implementar una capa paneuropea de datos que permita el uso avanzado de los recursos de contenido en el ámbito de la investigación y la educación superior. DRIVER desarrolla una *infraestructura de servicios* (de la que no se hablará en este documento) y una *infraestructura de datos*. Ambas están diseñadas para orquestar los recursos y los servicios existentes en la red de repositorios.

DRIVER como infraestructura de datos

La infraestructura de datos se sustenta sobre los recursos alojados localmente, como publicaciones científicas recopiladas en repositorios digitales de instituciones y organismos de investigación. Estos recursos se recolectan con DRIVER y se agregan a escala europea. Para poder garantizar una calidad óptima, DRIVER facilitará los medios posibles para armonizar y validar la agregación. DRIVER respetará la procedencia de los recursos mediante su “marcación” con información del repositorio local. DRIVER seguirá apuntando al repositorio local cuando se descargue un recurso en vez de suministrarlo. Los datos de DRIVER estarán disponibles para que todos los socios de la red DRIVER de proveedores de contenido puedan reutilizarlos mediante el protocolo OAI-PMH.

Banco de pruebas de DRIVER

La fase actual de pruebas del proyecto DRIVER sienta las bases de una ambiciosa y rica en contenidos infraestructura paneuropea de repositorios. La red de repositorios digitales es polifacética en lo que respecta a los distintos países, los distintos recursos (texto, datos o multimedia), las diferentes plataformas tecnológicas, las distintas políticas de metadatos, etc. Aun así, existen puntos en común en gran parte de este contexto: el texto es el tipo de recurso más suministrado por los repositorios digitales y el mejor modo de ofrecer estos recursos textuales es el protocolo OAI-PMH (*Open-Archives-Initiative Protocol for Metadata-Harvesting*). Por lo tanto, la fase actual de pruebas del proyecto DRIVER se centra en los recursos textuales que pueden recolectarse con el protocolo OAI-PMH.



Retos

Qué esperan los investigadores

Las expectativas de investigadores y otros usuarios actuales de los sistemas de información digital respecto al suministro de contenido digital son realmente altas. La recuperación debe ser rápida, directa, a una distancia de pocas pulsaciones del ratón y versátil. La cultura actual de la red de repositorios digitales no satisface estas expectativas. Aunque es cierto que se han establecido servicios valiosos de búsqueda y recuperación de registros bibliográficos (metadatos), el recurso en sí a veces está oculto tras varias páginas intermedias, escondido entre procedimientos de autorización, con presentación incompleta o totalmente irrecuperable. No obstante, para conseguir una comunicación científica óptima en sería preciso que el recurso se obtuviera con una única pulsación del ratón. Además, la recuperación sencilla del texto completo y de los metadatos, facilita el tratamiento automático del contenido. Ni el registro bibliográfico recolectado ni el texto completo recuperado por separado (aunque sí cuando ambos se combinan) permiten el desarrollo de servicios avanzados integrados, como búsquedas por materias combinadas con la navegación a través de clasificaciones, análisis de citas, etc.

El reto del texto completo

El fomento del acceso directo a los recursos textuales se ha convertido en un gran reto dentro del banco de pruebas del proyecto DRIVER. Mientras el consorcio de DRIVER dedica todo su esfuerzo a enfrentarse a este reto desde un punto de vista tecnológico (procesando los datos agregados), los servidores de repositorios digitales pueden ofrecer soporte a DRIVER localmente mediante el suministro de contenido de modo específico. Las directrices presentadas aquí pueden servir para guiar a los proveedores locales de contenido a la hora de ofrecer su contenido.

El siguiente paso

La recuperación de texto completo con datos bibliográficos supone un paso fundamental y necesario para conseguir servicios ricos en información basados en repositorios digitales. En futuros documentos de directrices, se tratarán otros tipos de información, como los datos primarios o multimedia, y se profundizará en objetos de información más complejos formados por varios recursos.



Acerca de las directrices

Por qué es necesario redactar directrices

El documento de DRIVER, *Directrices para proveedores de contenido: Exposición de recursos textuales con el protocolo OAI-PMH*, sirve para guiar a los administradores de los nuevos repositorios en la definición de las políticas de administración de datos, a los administradores de los repositorios existentes en los pasos que se deben seguir para conseguir servicios mejorados, y también a los desarrolladores de plataformas de repositorios para la incorporación de funciones adicionales en versiones futuras.

Cumplir las directrices

En un futuro próximo, DRIVER ofrecerá a los repositorios locales un modo de comprobar (mediante una interfaz web) el grado de conformidad con las directrices. DRIVER también ofrece soporte telefónico y asistencia en línea (véase más abajo). Si se cumplen los puntos *obligatorios* de las directrices, el repositorio recibe el estado de proveedor de DRIVER *validado*. Si también se cumplen los puntos *recomendados*, el repositorio recibe el estado de proveedor de DRIVER *con futuro*. Los repositorios *validados* de DRIVER pueden reutilizar datos de DRIVER para desarrollar servicios locales. Pasan a formar parte de la red DRIVER de proveedores de contenido.

Qué ocurre si no hay conformidad

No ser conforme con todos los puntos obligatorios o recomendados de las directrices, no significa necesariamente que el contenido de un repositorio no vaya a ser recolectado o agregado por DRIVER. Sin embargo, en función de los servicios específicos ofrecidos a través de la infraestructura de DRIVER, es posible que el contenido de estos repositorios simplemente no sea recuperable. Un servicio de búsquedas, por ejemplo, que prometa mostrar únicamente los registros que proporcionen un vínculo de texto completo no puede procesar todo el contenido de un repositorio que ofrezca registros únicamente de metadatos o que oculte textos completos mediante procedimientos de autorización. Estas directrices pueden ayudar a diferenciar dichos registros. Las directrices, por supuesto, *no* indicarán qué registros se deben conservar en el repositorio local.



Directrices de “Exposición de recursos textuales con el protocolo OAI-PMH”

Versión 1.1

Qué apoyo se ofrece

DRIVER ofrecerá soporte a los repositorios locales para que puedan implementar las directrices de forma individual. Puede obtener soporte en línea (Internet)¹ o de forma personalizada². DRIVER tiene un compromiso con las posibles soluciones que puedan conseguirse mediante el procesamiento central de datos. No obstante, el camino sostenible, transparente y escalable hacia los servicios mejorados pasa por los repositorios locales.

¹ <http://www.driver-support.eu>

² Véase el documento “Escenario de implementación de las directrices del proyecto DRIVER”



Ámbito de las directrices

¿Equivalen estas directrices a un estándar?

No. Aunque el uso de estándares como el protocolo OAI-PMH ciertamente supone una base sólida para crear una red como DRIVER, se necesitan directrices adicionales. El motivo principal es que los estándares están abiertos a interpretaciones e implementaciones locales. Sin ello, un estándar no podría existir. Pero esta apertura se puede convertir en un problema a la hora de conseguir servicios de calidad si se combinan implementaciones divergentes.

¿Las directrices son lo mismo que las reglas de catalogación?

No. Las directrices son una herramienta para asignar (o convertir) los metadatos empleados en el repositorio en metadatos de Dublin Core tal como los recolecta DRIVER. No están pensadas para utilizarlas como instrucciones de introducción de datos en la operación de inserción de metadatos en el sistema de repositorios.

¿Contienen las directrices instrucciones de calidad científica?

No. Las directrices no indican qué recursos han logrado alcanzar el nivel de calidad requerido en lo que respecta al contenido científico y cuáles no. Asumiremos que esta distinción ya se ha realizado en el nivel de repositorios, es decir, daremos por hecho que la calidad de los recursos expuestos en la recolección son lo suficientemente aceptables.

¿Cuáles son los componentes principales de las directrices?

Básicamente, las directrices se centran en tres asuntos: colecciones, metadatos y la implementación del protocolo OAI-PMH.

- En lo que respecta a las colecciones del repositorio, es obligatorio utilizar “sets” (agrupaciones) que definan las colecciones accesibles a texto completo. Si todos los recursos del repositorio son textuales y no sólo los metadatos, sino también el texto completo es accesible sin autorización, el uso de *sets* es optativo.
- En lo que respecta al protocolo OAI-PMH, se han definido algunas características obligatorias y otras características recomendadas para solucionar los problemas que puedan surgir en las distintas implementaciones del repositorio local.
- En lo que respecta a los metadatos, se han definido algunas características obligatorias y otras recomendadas para solucionar las dificultades



Directrices de “Exposición de recursos textuales con el protocolo OAI-PMH”

Versión 1.1

semánticas que puedan surgir en las diferentes interpretaciones de DUBLIN CORE.

¿Quién ha creado estas directrices?

Las directrices de DRIVER no surgen de la nada. Han sido recopiladas por profesionales con años de experiencia en la construcción y el mantenimiento de redes similares de repositorios intervinculados, como HAL (Francia), DARE (Países Bajos), DINI (Alemania) o SHERPA (Reino Unido), a partir de la experiencia de proveedores de servicios con gran trayectoria, como BASE, y con organizaciones comunitarias, como el grupo OAI de prácticas recomendadas.

¿Qué son exactamente los recursos textuales?

En esta fase del proyecto DRIVER, nos centraremos en los recursos textuales. Utilizamos las siguientes definiciones de trabajo:

recurso textual = artículos científicos, tesis doctorales, ensayos, libros electrónicos y formatos similares para actividades de investigación científica

acceso abierto = acceso sin necesidad de ningún medio de pago, licencias, control de acceso con contraseña, control de acceso mediante ip, etc.

Muchos repositorios se utilizan para depositar distintos tipos de recursos, por ejemplo, artículos, libros, fotografías, vídeos, datasets o material de aprendizaje. Estos recursos tienen registros de datos que los describen. Normalmente, los recursos se encuentran en formato digital (aunque no siempre es así) y los archivos digitales se suelen almacenar en una base de datos que forma parte del sistema del repositorio (aunque no siempre). El acceso a los recursos suele ser abierto (aunque no siempre).

En el proyecto DRIVER, nos centramos en un subconjunto del vasto dominio de recursos de los repositorios europeos: nos centramos en los recursos textuales en formato digital de acceso abierto.

Los estudios indican que, de este modo, podremos cubrir más del 80 % de todos los recursos disponibles.

Por este motivo, la primera directriz obligatoria de la Sección A dice así: “el repositorio contiene recursos textuales digitales”. Esto no significa que el repositorio no pueda incluir otros materiales o elementos no digitales también. La afirmación hace hincapié en que DRIVER se centra en los recursos textuales.



Directrices de “Exposición de recursos textuales con el protocolo OAI-PMH”
Versión 1.1

Puede consultar una lista completa de los recursos textuales en el elemento dc:type de las directrices de metadatos del anexo 1.

¿Qué son exactamente los *sets*?

Los *sets* (agrupaciones) son un componente estándar del protocolo OAI-PMH y se utilizan para acotar (filtrar) partes concretas de un repositorio. Si el repositorio también contiene elementos no textuales, no digitales, elementos de acceso de pago o únicamente elementos de metadatos, puede utilizar el mecanismo de *sets* para filtrar los elementos al suministrar el contenido a DRIVER.



Recursos adicionales

¿Qué más se debe tener en cuenta?

Los recursos existentes se han utilizado como información para redactar estas directrices y se ha prestado especial atención para evitar soluciones especiales. Así, se podría decir que las directrices de DRIVER sacan el máximo partido de la experiencia práctica y de otras directrices existentes a nivel internacional.

- DRIVER se ha diseñado siguiendo la estructura de redes distribuidas de proveedores de contenido operativas, concretamente la red DARE de los Países Bajos. Las directrices de DARE sirven como modelo para DRIVER. En vez de facilitar múltiples referencias a otras directrices repartidas por todo el mundo, DRIVER proporciona las directrices de DARE como un único documento (anexos). Concretamente, hay dos secciones fundamentales:
 - El documento USING SIMPLE DUBLIN CORE TO DESCRIBE EPRINTS, de Andy Powell, Michael Day y Peter Cliff, UKOLN, Universidad de Bath (versión 1.2), que se ha adaptado para cumplir algunos de los requisitos específicos de DARE. Está disponible como “Uso de DRIVER de Dublin Core” (versión 2, noviembre del 2006, véase el anexo 1)
 - La versión 2.0 del protocolo OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*), que también se ha adaptado a los requisitos específicos de DARE y está disponible como “Uso en DRIVER de las directrices de OAI-PMH” (versión 2, diciembre del 2006, véase el anexo 2)
- El certificado DINI *Document and Publication Services 2007* (Servicios de documentos y publicaciones del 2007) (versión 2, septiembre del 2006)³ expone de forma fiable qué se necesita tener en cuenta cuando se trabaja con un repositorio. Dado que DRIVER estudia los repositorios desde el punto de vista de un agregador, las directrices de DRIVER no cubren los aspectos descritos en el certificado DINI, que está diseñado como guía de operación local general de un repositorio. Pero las directrices de DRIVER se basan en la afirmación de que los criterios del certificado DINI se tienen en cuenta cuando se trabaja con un repositorio.

¿Existe una solución que resuelva varios problemas a la vez?

Sí. En DARE, se ha demostrado útil implementar un “contenedor XML” para cada recurso que permita la recolección de recursos con OAI-PMH, proporcione un vínculo inequívoco con el recurso (no mediante una página de acceso), admita el indexado de texto completo y permita representar documentos complejos

³ <http://www.dini.de/documents/dini-zertifikat2007-en.pdf>



compuestos de varios archivos PDF (anexo 3). El contenedor XML está basado en el Lenguaje de Declaración de Elementos Digitales (*Digital Item Declaration Language; MPEG21-DIDL*)⁴. También se han desarrollado otras soluciones basadas en DIDL (por ejemplo, aDORe⁵, o los perfiles METS⁶) y otras que se publicarán en el futuro (por ejemplo, ORE⁷).

APARTADO A

Recursos textuales

obligatorio

- El repositorio contiene recursos textuales digitales (consultar la explicación en la página 5).
- Los recursos textuales están en formatos ampliamente utilizados y extendidos (PDF, TXT, RTF, DOC, TeX, etc.).
- Los recursos textuales son de acceso abierto, están disponibles directamente en el repositorio para cualquier usuario del mundo, y sin ninguna restricción (autorización, pago).
- Los recursos textuales están descritos mediante registros de metadatos.
- Los recursos textuales y de metadatos se vinculan entre sí de tal modo, que un usuario final pueda acceder al recurso textual mediante el identificador (normalmente una URL) del registro de metadatos.
- La URL de un recurso codificada en el registro de metadatos siempre se puede localizar y nunca se cambia ni se reasigna.
- Un identificador único identifica el registro de metadatos y el recurso textual (no hay punteros a sistemas externos, como un sistema bibliotecario nacional o un editor).

⁴ <http://xml.coverpages.org/mpeg21-didl.html>

⁵ <http://african.lanl.gov/aDORe/projects/adoreArchive/>

⁶ <http://www.loc.gov/standards/mets/mets-profiles.html>

⁷ <http://www.openarchives.org/ore/>



recomendado

- Verificación transparente de la integridad de un recurso textual.
- Medidas de control de calidad (del contenido científico) de los recursos textuales expuestos para limitarlos a, por ejemplo, los recursos textuales incluidos en el informe científico anual (o equivalente).
- La URL de un recurso codificada en el registro de metadatos se basa en un esquema de identificadores persistentes (como DOI, URN, ARK, etc.).



APARTADO B

Metadatos

obligatorio

- Los metadatos se estructuran según la norma Unqualified Dublin Core (ISO 15836:2003).
- Los elementos individuales de DC se utilizan según lo dispuesto en el documento “Uso en DRIVER de Dublin Core” (anexo 1).

recomendado

- Los metadatos se estructuran según esquemas más completos, como Dublin Core Qualified o MODS.
- El idioma de los metadatos queda a la elección del proveedor de contenido. Se recomienda utilizar el inglés.
- El idioma recomendado para los resúmenes (los resúmenes son optativos) del artículo es el inglés.



APARTADO C

Implementación de OAI-PMH

obligatorio

- El repositorio debe ser conforme con OAI y con la especificación “Uso en DRIVER de OAI-PMH” (anexo 2).
- Debe existir un identificador de repositorio y debe utilizarse el esquema de identificador OAI (anexo 2).
- Si (y sólo si) el repositorio contiene recursos que no sean obligatorios en el APARTADO A debe definirse un *set* OAI (consultar explicación en la página 5) que identifique la colección de recursos textuales digitales con acceso directo (anexo 2).

recomendado

- Previsiones de modificación de URL base (anexo 2).
- Respuesta completa a Identify, incluido el uso de la declaración Description (anexo 2).
- Uso del contenedor XML de DIDL (anexo 3).