



*Digital Repository Infrastructure Vision for European Research*

Directrizes para fornecedores de conteúdos

***Exposição de recursos textuais com o  
protocolo OAI-PMH***

Aplicação Piloto

Versão 1.0

Colaboradores

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

Tradução

Serviços de Documentação da Universidade do Minho



## ***Sobre o DRIVER***

### **O que é o DRIVER**

O DRIVER, “*Digital Repository Infrastructure Vision for European Research*”, é um projecto dinamizado por um consórcio financiado pela União Europeia (UE) e que visa a constituição de uma estrutura organizacional e tecnológica para implementar uma camada de dados pan-europeia que permita o uso avançado de recursos de conteúdos na área da investigação no ensino superior. O DRIVER desenvolve uma *infra-estrutura de serviços* (que não serão descritos neste documento) e uma *infra-estrutura de dados*. Ambas estão concebidas para instrumentar os recursos e serviços existentes na rede de repositórios.

### **DRIVER como infra-estrutura de dados**

A infra-estrutura de dados baseia-se em recursos alojados localmente, como publicações científicas recolhidas em repositórios digitais de instituições e organismos de investigação. Estes recursos serão recolhidos pelo DRIVER e agregados à escala europeia. Para poder garantir uma elevada qualidade da agregação, o DRIVER fornecerá os meios possíveis para a harmonizar e validar. O DRIVER respeitará a proveniência dos recursos mediante a sua “marcação” com informação do repositório local. Além disso, quando um recurso for descarregado o Driver irá remeter para o repositório local ao invés de o fornecer ele próprio. Os dados do DRIVER estarão disponíveis para reutilização via OAI-PMH por todos os parceiros da rede DRIVER de fornecedores de conteúdos.

### **Banco de ensaios DRIVER**

Na actual fase de testes, o projecto DRIVER lança os alicerces para uma rica e ambiciosa infra-estrutura pan-europeia de repositórios. A paisagem dos repositórios digitais é multifacetada no que concerne aos diferentes países, aos diferentes tipos de recursos como texto, dados ou multimédia, às diferentes plataformas tecnológicas, às diferentes políticas de metadados, etc. No entanto, existem pontos comuns que se aplicam a grande parte deste espectro: o tipo de recurso mais comum nos repositórios digitais é o texto e a principal forma de oferecer estes recursos textuais é o protocolo OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Por esse motivo, a actual fase de testes do projecto DRIVER centra-se nos recursos textuais que podem ser recolhidos através do protocolo OAI-PMH.



## ***Desafios***

### **O que esperam os investigadores**

Os investigadores e outros actuais utilizadores de sistemas de informação têm expectativas bastante elevadas quanto ao fornecimento de conteúdos digitais. A recuperação deve ser rápida, directa, acessível com poucos cliques e versátil. A cultura actual no panorama dos repositórios digitais não satisfaz completamente estas expectativas. Embora muitos serviços de valor acrescentado tenham sido implementados para pesquisar e recuperar registos bibliográficos (metadados), o recurso em si mesmo é, por vezes, ocultado por detrás de várias páginas intermédias, obscurecido por procedimentos de autorização, apresentado de forma incompleta ou totalmente irrecuperável. No entanto, para conseguir uma comunicação científica óptima seria necessário que o recurso fosse obtido apenas com um clique do rato. Para além disso, a recuperação simples do texto integral e dos metadados facilita o tratamento automático do conteúdo. Nem o registo bibliográfico recolhido nem o texto completo recuperado em separado – apenas a combinação de ambos – permitem o desenvolvimento de serviços avançados e integrados, como a pesquisa por assuntos combinada com a navegação através de classificações, análise de citações, etc.

### **O desafio do texto integral**

Favorecer o acesso directo a recursos textuais foi identificado como um grande desafio na fase de testes do DRIVER. Embora o consórcio DRIVER dedique todos os esforços possíveis para abordar este desafio do ponto de vista tecnológico através do processamento de dados agregados, os detentores de repositórios digitais podem apoiar localmente o DRIVER através do fornecimento de conteúdos de uma forma específica. As directrizes aqui apresentadas orientam para que os fornecedores de conteúdos locais saibam como disponibilizar os seus conteúdos.

### **O que se segue?**

A recuperação de texto integral com dados bibliográficos é um passo básico mas necessário para conseguir serviços ricos em informação baseados em repositórios digitais. Futuras directrizes, abordarão os passos seguintes relativamente a outros tipos de informação, como dados primários ou multimédia e objectos de informação mais complexos formados por vários recursos.



## ***Sobre as Directrizes***

### **Porquê utilizar as directrizes?**

O documento do DRIVER, *Directrizes para fornecedores de conteúdos: Exposição de recursos textuais com o protocolo OAI-PMH*, fornece orientação aos administradores dos novos repositórios na definição de políticas locais de gestão de dados, aos administradores de repositórios já existentes na tomada de medidas para serviços melhorados e aos programadores de plataformas de repositórios no acrescento de novas funcionalidades de suporte em futuras versões.

### **Como cumprir as directrizes**

Num futuro próximo, o DRIVER oferecerá aos repositórios locais (através de um interface web) um modo de aferir o grau de conformidade com as directrizes. O DRIVER também oferece suporte telefónico e assistência em linha (ver em baixo). Se os pontos *obrigatórios* das directrizes forem cumpridos, o repositório recebe o estatuto de fornecedor DRIVER *validado*. Se também forem cumpridos os pontos *recomendados*, o repositório recebe o estatuto de fornecedor DRIVER *certificado para o futuro*. Os repositórios DRIVER validados podem reutilizar dados do DRIVER para desenvolver serviços locais. Passam a integrar a rede de fornecedores de conteúdos DRIVER.

### **O que acontece se não for conforme?**

Não estar em conformidade com todos os pontos obrigatórios ou recomendados das directrizes não significa necessariamente que os conteúdos de um repositório não serão recolhidos ou agregados pelo DRIVER. Porém, em função dos serviços específicos oferecidos através da infra-estrutura DRIVER, é possível que o conteúdo destes repositórios simplesmente não seja recuperável. Por exemplo, um serviço de pesquisa, que pretenda listar apenas registos que ofereçam um apontador para o texto integral não pode processar todo o conteúdo de um repositório que ofereça registos unicamente de metadados ou que oculte os textos integrais através de procedimentos de autorização. Estas directrizes ajudarão a distinguir esses registos. As directrizes não determinarão, como é óbvio, que registos devem ser mantidos no repositório local.

### **Existe suporte?**

O DRIVER oferecerá suporte aos repositórios locais para que possam implementar as directrizes numa base individual. O suporte pode ser obtido através da internet<sup>1</sup>

---

<sup>1</sup> <http://www.driver-support.eu>



## Directrizes “Exposição de recursos textuais com o protocolo OAI-PMH”

Versão 1.0

ou pode ser pessoal<sup>2</sup>. O DRIVER está empenhado em qualquer solução possível que possa realizar-se através do processamento central de dados. Não obstante, o caminho sustentável, transparente e escalável para serviços melhorados passa pelos repositórios locais.

---

<sup>2</sup> Ver documento “*Advice for implementation of the DRIVER Guidelines*”



## ***Âmbito das Directrizes***

### **As directrizes são uma norma?**

Não. Embora o uso de normas como o OAI-PMH providencie certamente uma base sólida para criar uma rede como o DRIVER, são necessárias directrizes adicionais. O principal motivo é que as normas ainda dão lugar a interpretações e implementações locais. Sem isso, uma norma não poderia existir. Porém esta abertura pode-se converter num obstáculo para a obtenção de serviços de alta qualidade quando se combinam implementações divergentes.

### **As directrizes equivalem a regras de catalogação?**

Não. As directrizes são um instrumento para mapear (ou traduzir) os metadados utilizados no repositório para os metadados em Dublin Core, tal como são recolhidos pelo DRIVER. Não estão pensadas para serem utilizadas como instruções de introdução de dados na inserção de metadados nos sistemas de repositório locais.

### **As directrizes possuem instruções do nível de qualidade científica?**

Não. As directrizes não indicam que recursos possuem o nível de qualidade requerido no que respeita ao conteúdo científico e quais os que não. Assumiremos que esta distinção já foi feita ao nível dos repositórios, isto é, assumiremos que a qualidade dos recursos expostos através da recolha é suficientemente boa.

### **Quais são os principais componentes das directrizes?**

As directrizes focam basicamente três questões: colecções, metadados e implementação do protocolo OAI-PMH.

- No que respeita às colecções do repositório, é obrigatório utilizar “*sets*” (conjuntos) que definam as colecções com texto integral. Se todos os recursos do repositório forem textuais, incluam não só os metadados, mas também o texto integral e todos os recursos forem acessíveis sem autorização, o uso de *sets* é opcional.
- No que respeita ao protocolo OAI-PMH, foram definidas algumas características obrigatórias e outras recomendadas para solucionar os problemas que surjam nas diferentes implementações no repositório local.
- No que respeita aos metadados, foram definidas algumas características obrigatórias e outras recomendadas para solucionar as dificuldades semânticas que surjam de diferentes interpretações do Dublin Core.



Directrizes “Exposição de recursos textuais com o protocolo OAI-PMH”  
Versão 1.0

### Quem criou as directrizes?

As directrizes do DRIVER não surgem do nada. Foram compiladas por profissionais com anos de experiência na construção e manutenção de redes similares de repositórios interligados, como *HAL* (França), *DARE* (Holanda), *DINI* (Alemanha), *SHERPA* (Reino Unido), e envolvem a competência de fornecedores de serviços experientes, como o *BASE*, e organizações comunitárias, como o grupo *OAI Best-Practice*.

### O que se entende por recursos textuais?

Nesta fase do projecto DRIVER estamos centrados nos recursos textuais. Como definições de trabalho utilizamos as seguintes:

*recurso textual = artigos científicos, teses de doutoramento, documentos de trabalho, livros electrónicos e resultados similares de actividades de investigação científica*

*acesso livre = acesso sem qualquer forma de pagamento, licenciamento, controlo de acesso com password, controlo de acesso mediante IP, etc*

Muitos repositórios são utilizados para depositar diferentes tipos de recursos, por exemplo, artigos, livros, fotografias, vídeos, conjuntos de dados (*data sets*) ou recursos de aprendizagem. Estes recursos possuem registos de metadados que os descrevem. Normalmente, os recursos encontram-se em formato digital (mas nem sempre) e estes ficheiros digitais são usualmente armazenados numa base de dados que é parte do sistema do repositório (mas nem sempre). O acesso aos recursos é geralmente livre (mas nem sempre).

No projecto DRIVER, concentramo-nos num subconjunto do vasto domínio de recursos dos repositórios europeus: focamo-nos nos recursos textuais em formato digital de acesso livre. Estudos indicam que deste modo poderemos cobrir mais de 80% de todos os recursos disponíveis. Por este motivo, a primeira directriz obrigatória da Secção A refere: “o repositório contém recursos textuais digitais”. Isto não significa que o repositório não possa incluir outros materiais ou itens não-digitais. A afirmação é uma expressão do foco do DRIVER nos recursos textuais.

Pode consultar uma lista completa dos recursos textuais no elemento *dc:type* nas directrizes de metadados no Anexo 1.



## Directrizes “Exposição de recursos textuais com o protocolo OAI-PMH”

Versão 1.0

### **O que são sets?**

*Sets* (conjuntos) são um componente normalizado do protocolo OAI-PMH e são utilizados para apontar (filtrar) partes específicas de um repositório. Se o repositório também contém itens não textuais, ou não digitais, ou itens de acesso pago ou registos exclusivamente com metadados, pode utilizar o mecanismo de *sets* para filtrar esses itens quando disponibilizar os conteúdos ao DRIVER.



## ***Mais Recursos***

### **Que mais se deve considerar?**

Foram utilizados recursos existentes como contributo para elaborar estas directrizes e prestou-se particular atenção para evitar soluções especiais. Assim, poder-se-á afirmar que as directrizes do DRIVER exploram ao máximo a experiência prática e outras directrizes existentes a nível internacional.

- O DRIVER foi desenhado seguindo a estrutura de redes operacionais e distribuídas de fornecedores de conteúdos, concretamente da rede DARE na Holanda. As directrizes DARE servem como modelo para o DRIVER. Ao invés de indicar múltiplas referencias a outras directrizes espalhadas por todo o mundo, o DRIVER apresenta as directrizes do DARE como um único documento (Anexos). Concretamente, há duas secções essenciais:
  - O documento USING SIMPLE DUBLIN CORE TO DESCRIBE EPRINTS, de Andy Powell, Michael Dayy Peter Cliff, UKOLN, Universidade de Bath (versão 1.2), que foi adaptado para cumprir alguns requisitos específicos do DARE. Está disponível como “Directrizes Específicas de Metadados DRIVER” (versão 2, Novembro de 2006, ver anexo 1)
  - A versão 2.0 do protocolo OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), que também foi adaptada aos requisitos específicos do DARE e que está disponível como “*Directrizes Específicas para o uso de OAI-PMH*” (Versão 2, Dezembro de 2006, ver Anexo 2)
- O DINI-Certificate “*Document and Publication Services 2007*” (Versão 2, Setembro de 2006)<sup>3</sup> expõe de forma fiável o que se deve considerar quando se trabalha num repositório. Dado que o DRIVER aborda os repositórios do ponto de vista de um agregador, as directrizes do DRIVER não cobrem os aspectos descritos no certificado DINI, que está desenhado como guia geral da operação local de um repositório. Mas as directrizes do DRIVER baseiam-se na suposição de que os critérios do certificado DINI são tidos em conta no funcionamento de um repositório.

### **Existe alguma solução que resolva de imediato vários problemas?**

Sim, no contexto do DARE comprovou-se a utilidade de implementar um “XML-Container” para cada recurso que permita a recolha de recursos com OAI-PMH, proporcione um apontador inequívoco para o recurso (não mediante uma página de acesso), suporte a indexação de texto integral e permita a representação de

---

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



## Directrizes “Exposição de recursos textuais com o protocolo OAI-PMH”

Versão 1.0

documentos complexos compostos por vários ficheiros PDF (Anexo 3). O *XML-Container* é baseado na Digital Item Declaration Language (MPEG21-DIDL)<sup>4</sup>. Outras soluções baseadas em DIDL também foram desenvolvidas (ex. aDORe<sup>5</sup> e os perfis METS<sup>6</sup>) e outras a publicar no futuro (e.g. ORE<sup>7</sup>).

---

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

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

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

<sup>7</sup> <http://www.openarchives.org/ore/>



## PARTE A

### Recursos Textuais

#### Obrigatório

- O repositório contém recursos digitais textuais (ver explicação na página 6).
- Os recursos textuais estão em formatos amplamente utilizados e difundidos (PDF, TXT, RTF, DOC, TeX etc.).
- Os recursos textuais estão em acesso livre, disponíveis directamente do repositório para qualquer utilizador sem restrições como autorizações ou pagamento.
- Os recursos textuais são descritos por registos de metadados.
- Os recursos textuais e de metadados estão ligados entre si de tal modo, que um utilizador final possa aceder ao recurso textual através do identificador (normalmente um URL) no registo de metadados.
- O URL de um recurso inscrito no registo de metadados está permanentemente acessível e nunca se altera ou se atribui a outro recurso.
- Um identificador único identifica o registo de metadados e o recurso textual (não há apontadores para sistemas externos, como um sistema nacional de bibliotecas ou uma editora).

#### Recomendado

- Verificação transparente da integridade de um recurso textual.
- Medidas de controlo de qualidade (do conteúdo científico) dos recursos textuais expostos para limitá-los a, por exemplo, os recursos textuais incluídos no relatório científico anual (ou equivalente).
- O URL de um recurso inscrito no registo de metadados baseia-se num esquema de identificadores persistentes como: DOIs, URNs, ARKs.



## **PARTE B**

### **Metadados**

#### **Obrigatório**

- Os metadados são estruturados como Dublin Core não qualificado (ISO 15836:2003).
- Os elementos individuais de DC devem ser usados de acordo com o Documento “Directrizes Específicas de Metadados DRIVER” (Anexo 1).

#### **Recomendado**

- Os metadados são estruturados de acordo com esquemas mais completos como Dublin Core qualificado ou MODS.
- A escolha do idioma dos metadados fica ao critério do fornecedor de conteúdos. O idioma recomendado é o inglês.
- O idioma recomendado para o resumo (a inclusão de resumo é opcional) do artigo é o inglês.



## PARTE C

### Implementação OAI-PMH

#### Obrigatório

- O repositório deve estar em conformidade com o *OAI* e de acordo com as especificações “*Directrizes Específicas para uso do protocolo OAI-PMH*” (Anexo 2).
- Deve existir um identificador de repositório e deve utilizar-se o esquema de identificador OAI (Anexo 2).
- Se (e apenas se) o repositório contém outros recursos para além dos que são obrigatórios na PARTE A, deve ser definido um set OAI (ver explicação na página 5) que identifique a colecção de recursos textuais digitais com acesso directo (Anexo 2).

#### Recomendado

- Disposições para a alteração do URL Base (Anexo 2).
- Resposta ao *Identify* completa, incluindo o uso opcional da declaração *Description* (Anexo 2).
- Uso do DIDL *XML-container* (Anexo 3).