Metadados para preservação digital e aplicação do modelo OAIS

Maria de Lurdes Saramago

Biblioteca do Departamento de Biologia Animal da Faculdade de Ciências da Universidade de Lisboa
Campo Grande, Ed. C2, 3º Piso
1700 Lisboa
Tel: 217500000
E-mail: lurdes.saramago@netvisao.pt

RESUMO

Caracterizam-se os metadados para preservação digital de longo prazo em descritivos, administrativos e estruturais e enfatiza-se a necessidade de proceder à sua correcta aplicação desde a criação, primeira fase do ciclo de vida dos recursos digitais.

Acrescenta-se ainda a título exemplificativo do uso desta boa prática um conjunto de metadados de utilização obrigatória para os criadores de modelos em CAD que depositam os seus recursos na AHDS (AHDS (Arts and Humanities Data Service).

Descreve-se em seguida um enquadramento conceptual que neste momento se encontra em grande expansão, baseado no modelo de referência OAIS (Open Archive Information System) criado para a NASA (National Aeronautics and Space Administration) pelo CCSDS (NASA Consultative Committee for Space Data Systems).

Definem-se, no âmbito deste modelo de referência, o papel dos principais intervenientes, que são os produtores de informação, a gestão e os utilizadores que interrogam o sistema.

PALAVRAS-CHAVE

Preservação digital, Metadados de preservação, OAIS

INTRODUÇÃO

Ao longo dos últimos anos assistimos ao crescimento acentuado da criação de recursos electrónicos não só provenientes de digitalização como nascidos digitais. Consequentemente surge o receio de, dada a enorme vulnerabilidade do ambiente digital, vir a perder-se toda a informação assim gerada. É impossível ter acesso aos recursos digitais ao longo do tempo sem assegurar a existência de sistemas de *hardware* e *software* compatíveis pois os recursos para serem lidos carecem de enquadramento tecnológico. Da mesma forma é também impossível recuperar os recursos sem a existência de um conjunto de metadados que os enquadre e os documente.

Ao mesmo tempo que surgem novos desafios no campo da preservação colocados pelo ambiente da Internet aparecem favorecidos os acessos e os contactos entre instituições congéneres. É fundamental ou mesmo condição de sobrevivência da memória colectiva pensar em termos de cooperação entre arquivos, bibliotecas, museus, grandes editores, produtores de informação em geral, criadores de *software*, etc.

Os altos custos a ultrapassar, por um lado, e a distribuição generalizada dos recursos em rede, por outro, facilitam a emergência de parcerias.

A título de exemplo podemos referir a utilização do modelo de referência OAIS (Open Archive Information System) criado sob os auspícios da NASA (National Aeronautics and Space Administration) pelo CCSDS (NASA Consultative Committee for Space Data Systems) [1].

Dadas as suas características, este modelo manifesta condições para poder ser aplicado por variadas instituições em parceria, não obstante as divergências existentes entre as potenciais comunidades envolvidas. Da mesma forma os metadados de preservação devem ser pensados em função das vocações e objectivos das comunidades, ou seja, partindo de uma base comum a vários parceiros, os conjuntos de metadados devem sofrer as necessárias adaptações.

A capacidade tecnológica e a disponibilidade de meios financeiros, é também condição fundamental dado que a preservação de recursos digitais movimenta enormes massas de recursos de todos os tipos e para os preservar é necessário recorrer a montantes bastante elevados.

METADADOS PARA PRESERVAÇÃO DIGITAL DE LONGO PRAZO

Definimos metadados para preservação de longo prazo como informação de apoio aos processos associados com a preservação digital de longo prazo. Neste contexto longo prazo define o espaço de tempo determinado pelo acesso continuado aos recursos digitais ou pelo menos à informação neles contida, indefinidamente.

Em ambiente digital, os recursos sofrem transformações, cujos resultados nem sempre são fáceis de controlar. Por este motivo deve ser criado um histórico da mudança ao longo do tempo. O objectivo principal é garantir que a sua autenticidade e integridade sejam recompostas.

De igual maneira, as tecnologias de acesso aos recursos digitais rapidamente se tornam obsoletas e por isso haverá que encapsular informação acerca dos suportes de

armazenamento, hardware, sistema operativo e respectivas aplicações utilizadas durante o ciclo de vida dos recursos.

Os metadados de preservação devem conter informação técnica e administrativa sobre decisões e acções de preservação, registar os efeitos das estratégias de conversão de dados, assegurar a autenticidade dos recursos digitais ao longo do tempo, registar informação acerca de gestão de colecções e de direitos e ainda fornecer informação acerca dos próprios metadados.

Não é demais referir que existem recursos para os quais é necessário assegurar o seu *look and feel*, ou seja, as suas características externas de usabilidade. É fundamental que existam metadados que documentem exaustivamente qualquer das estratégias envolvidas.

Do estrito ponto de vista da preservação digital de longo prazo, apesar de todos os objectivos apresentados serem indispensáveis à boa gestão dos recursos antes e depois de depositados num repositório, interessa-nos, particularmente, o objectivo de recuperar informação acerca dos requisitos e condições, técnicas ou formais de preservação de longo prazo.

Segundo o "OCLC/RLG Working Group on preservation metadata" em 2002 [2] as características fundamentais dos metadados de preservação são as seguintes:

- Abrangência, i.e., devem ser constituídos por todos os requisitos de informação necessários à gestão de um repositório desde a sua inclusão até à sua disponibilização e acesso.

- Estruturação, ou seja, devem apresentar uma descrição de alto nível dos componentes chave do sistema e das suas funcionalidades. Este ponto vem complementar o primeiro.

- Aplicação alargada, i.e., os metadados de preservação devem poder aplicar-se a um leque variado de tipos de recursos digitais, de actividades e de instituições. Uma estrutura de metadados de preservação representa o consenso de um grupo de trabalho e deve ser imparcial sobre assuntos relacionados com as opções de estratégias de preservação.

Consideramos que os metadados de preservação são de três tipos :

(a) descritivos

(b) administrativos

(c) estruturais

A incidência sobre os últimos dois é relevante, pois é neste espaço que vamos encontrar as descrições dos métodos e das estratégias tomadas para preservação. Os metadados descritivos destinam-se fundamentalmente às fases de acesso e estão para os recursos digitais como os formatos MARC (Machine Readable Cataloguing) para os recursos bibliográficos tradicionais.

Do ponto de vista da preservação propriamente dita, ou seja, o objectivo deste trabalho, são os metadados administrativos aqueles que têm um peso mais importante pois documentam actos de gestão ao longo do tempo, desde a ingestão no repositório. Os metadados estruturais complementam a informação administrativa pois acrescentam o enquadramento tecnológico indispensável à boa recuperação dos recursos.

A inclusão de metadados de preservação deve acompanhar todo o ciclo de vida do recurso digital, ou seja :

- Criação

- Selecção

- Identificação persistente

- Descrição e acesso

- Armazenamento

- Preservação

Referindo-nos à fase da criação, é importante que os repositórios alertem os criadores para a inserção dos metadados necessários logo nessa primeira fase.

Enquanto o trabalho está em mãos é muito mais fácil recordar os passos dados para a construção do trabalho. A documentação produzida ajudará tanto os próprios membros da equipa de trabalho como no futuro será uma componente vital no processo de preservação a longo prazo. É desnecessária uma documentação exaustiva de todo o processo criativo mas fundamental documentar algumas fases do processo dado que cada trabalho pode conter um número alargado de recursos e ficheiros.

Encontramos nas instruções da AHDS (Arts and Humanities Data Service) [3] para os produtores de modelos em CAD um bom exemplo de um conjunto de metadados que devem ser incluídos na fase de criação.

Deste modo, para cada projecto devem ser fornecidos metadados que contenham:

- Uma descrição alargada, em diagonal, de todo o projecto

- Métodos e normas usados no projecto

- Descrição individualizada dos modelos no projecto

Para cada projecto deve ser fornecida uma lista dos ficheiros criados que deve incluir :

- Nome do ficheiro

- Data de criação ou da última actualização

- Formato dos dados e número da versão utilizada

- Descrição do conteúdo

- *Copyright* associado

De igual modo para cada modelo os criadores devem

também apresentar informação sobre :

-Título do projecto

- Número de referência

- Criador

- Título do modelo CAD

- Software CAD

- Ficheiros usados

E ainda informação para algumas bases de dados associadas, que deve incluir :

- Título do projecto

- Referência do projecto

- Base de dados c/ versão e tipo

- Título das tabelas ou ficheiros, assim como número de referência

- Campos da tabela

- Título do ficheiro CAD que está associado à BD

- Formato do ficheiro

- Data de criação da BD

Sem a ajuda dos metadados do criador, antes da aceitação para depósito, não seria possível reconstruir os modelos posteriormente.

Cabe desta forma a cada repositório encaminhar os criadores de conteúdos da sua comunidade para uma conduta baseada no seguimento de boas práticas e verificar no acto de depósito se estas foram cumpridas.

De entre os sistemas de metadados de âmbito mais genérico e que procuram adaptar-se à preservação digital, o esquema de metadados METS (Metadata Encoding and Transmission Standards) [4] é uma norma para codificação de metadados descritivos, administrativos e estruturais de recursos digitais que utiliza a linguagem XML. esta norma é mantida pelo Network Development and MARC Standards Office da Biblioteca do Congresso e tem sido desenvolvida como uma iniciativa da DLF [5] (Digital Library Federation).

Dependendo da sua utilização, um documento METS pode ser usado como SIP (Submission Information Package), como AIP (Archival Information Package) ou mesmo como DIP (Dissemination Information Package no âmbito do modelo de referência OAIS.

O MODELO DE REFERÊNCIA OAIS (OPEN ARCHIVE INFORMATION SYSYTEM)

O modelo de referência OAIS (Open Archival Information System Reference Model) [1], foi desenvolvido pelo Consultative Committee for Space Data Systems (CCSDS) no âmbito da NASA. Este modelo, é uma norma ISO com o nº 14721:2002 que descreve um enquadramento conceptual para um repositório digital genérico, aberto a todas as comunidades com garantias de confiabilidade. Da norma consta também um léxico próprio que viabiliza a comunicação entre as comunidades e os repositórios.

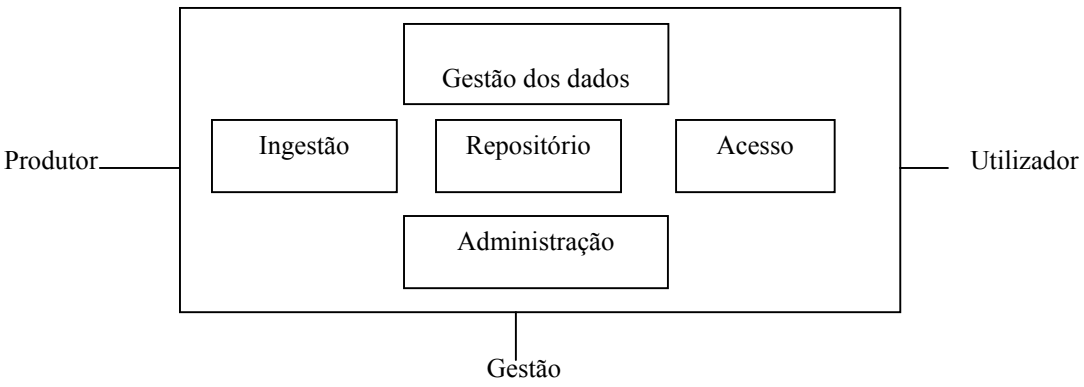
Tal como referido na norma o modelo OAIS consiste numa organização de pessoas e sistemas que aceitaram a responsabilidade de preservar a informação e torná-la disponível a uma designada comunidade.

Este modelo existe em sistema aberto e destina-se a um alargado leque de repositórios, os quais estão por sua vez ao serviço das mais variadas comunidades tanto de âmbito nacional e patrimonial como académico ou outras.

Deverá existir em simultâneo um modelo de informação onde se encontram descritos os requisitos de metadados de preservação de longo prazo.

O ambiente do modelo OAIS

O modelo OAIS opera num ambiente constituído pela interacção de produtores, utilizadores, gestão e o repositório em si mesmo.



Ambiente OAIS [1]

O papel dos principais intervenientes :

- O produtor fornece a informação a preservar;

- A gestão estabelece a política do OAIS e monitoriza-a passo a passo;

- O utilizador interage com o OAIS e obtém a informação que procura.

Toda a informação submetida a um OAIS por um produtor e toda a difusão estabelecida a partir do OAIS a um utilizador ocorre numa ou mais sessões discretas através de pacotes de informação.

Um pacote de informação é um envelope conceptual onde estão encapsulados informação de conteúdo e metadados de preservação.

No ambiente de gestão é necessário percorrer alguns passos :

- Negociar com os produtores e detentores de informação e respectivos direitos, a garantia de autorizações de preservação dos recursos a longo prazo e a também a possibilidade da sua disponibilização aos utilizadores finais.

- Assegurar que a informação a preservar é compreensível por si só, na comunidade designada, ou seja, que a comunidade, através de interfaces adequados, será capaz de compreender a informação sem a necessidade de recorrer à assistência de terceiros.

- Seguir políticas e procedimentos documentados que assegurem que a informação é preservada contra quaisquer contingências e assegurar a disseminação da informação com cópias autênticas a partir do original ou similares ao original.

- Assegurar que a informação preservada está disponível para a comunidade designada.

- Trabalhar em conjunto com a comunidade de produtores/fornecedores de informação do repositório, aconselhando a utilização de boas práticas na criação dos recursos digitais.

- Verificar a qualidade dos metadados. Idealmente quaisquer metadados que acompanhem o recurso quando este é submetido ao repositório devem ser verificados e, se necessário melhorados para que a manutenção de longo prazo seja assegurada e ao mesmo tempo que o acesso continuado seja permitido.

- Estabelecer identificadores únicos e persistentes para os recursos.

O modelo de informação

Tal como já referimos, os recursos encontram-se expressos através de diversos tipos de dados. É a interpretação destes, em simultâneo com a representação da informação que é necessário adicionar que vai permitir a posterior reconstituição da informação.

A Representação da Informação

Num nível muito baixo a representação da informação está contida um fluxo de *bits*.

A representação da informação indica se um fluxo de *bits* representa um parágrafo de texto, um ficheiro de som, uma imagem, etc. Contudo o conhecimento do formato do ficheiro descrito no fluxo de *bits* pode não ser suficiente para interpretar o seu conteúdo.

A representação da informação pode assumir duas formas:

1) informação estrutural

2) informação semântica.

A informação estrutural interpreta os *bits* organizando-os por tipos de dados, grupos de tipos de dados e outros significados de alto nível. Esta deve incluir especificação do formato dos dados e uma possível descrição do ambiente do *hardware* e do *software* em que os dados foram criados e que se torna necessária para o acesso posterior.

A informação semântica, por outro lado, acrescenta significado à estrutura dos dados, identificada através da informação estrutural. P. ex. a informação estrutural pode identificar um fluxo de caracteres de texto ASCII enquanto a informação semântica pode indicar que esse texto se encontra escrito em língua inglesa.

No ambiente do modelo OAIS a representação da informação encontra-se ela própria em formato digital e por esse motivo deve acrescentar-se informação adicional para interpretar o fluxo de *bits* da representação da informação, é por este motivo, necessária a existência de uma terceira camada de representação da informação, etc.

O modelo de referência OAIS recomenda que o resultado da rede de representação termine com a elaboração de um documento físico que dê por finda a construção da rede e dê início ao processo de interpretação.

O Objecto de Informação

Ao conjunto formado pelos objectos-dados e pela representação da informação que os acompanha chamamos objectos de informação.

É de realçar que os objectos-dados e os seus metadados são, pelo menos do ponto de vista lógico, objectos separados, mesmo que os metadados estejam inseridos no objecto, o que pode ser p. ex. o caso de um documento HTML.

Num ambiente digital isto implica uma sequência de *bits*, combinada com todos os dados necessários a torná-la compreensível. Existem quatro classes de objectos de informação que, em conjunto, formam pacotes de informação:

- Informação acerca do conteúdo

- Informação descritiva para preservação

- Informação para empacotamento

- Informação descritiva

Pacotes de Informação

Os pacotes de informação podem ser de três tipos:

1) Pacote de informação para submissão - é enviado do produtor da informação para o depósito.

2) Pacote de informação para depósito - preparado para ser armazenado pelo depósito.

3) Pacote de informação para difusão - enviado ao utilizador em resposta e uma pesquisa já em contexto de acesso.

Num contexto de preservação de metadados, a informação relevante encontra-se no pacote de informação para depósito, dado que este é o pacote que se destina à preservação de longo prazo.

Por sua vez, o pacote de informação para depósito é uma agregação de quatro tipos de objectos de informação:

1) Informação acerca do conteúdo - consiste na informação que o repositório tem a obrigação de preservar em conjunto com a informação de representação.

2) Informação descritiva para a preservação - contém informação necessária para gerir a preservação da informação sobre o conteúdo, com que está associada. Esta informação divide-se em quatro tipos :

- Informação acerca da referência – enumera e descreve os identificadores destinados à informação sobre o conteúdo de tal maneira que se tornem inequívocos, interna e externamente ao depósito (p. ex : ISBN, URN)

- Informação acerca da proveniência - documenta a história da informação sobre o conteúdo (p. ex. origem, histórico de custódia, acções e efeitos da preservação)

- Informação acerca do contexto - documenta as relações entre a informação sobre o conteúdo e o seu ambiente (p. ex. razões pelas quais foi criado, relações com outras informações de conteúdo, etc.)

- Informação acerca da reparabilidade : documenta mecanismos de reparabilidade e autenticação usados para assegurar que o conteúdo da informação não foi alterado de forma não documentada (p. ex. *checksums* ou assinaturas digitais)

3) Informação para empacotamento - envolve o objecto digital e os metadados associados numa unidade ou pacote.

4) Informação descritiva – destina-se a facilitar o acesso à informação sobre o conteúdo através das ferramentas de pesquisa e recuperação. A informação descritiva serve de *input* das ajudas à localização de depósitos e deriva tipicamente da informação sobre o conteúdo ou da informação descritiva para preservação.

O modelo OAIS representa uma descrição de alto nível dos tipos de informação gerados e geridos num contexto global de sistema de depósito digital. Não transmite pressupostos acerca do tipo de recursos digitais manuseados no depósito nem acerca das especificações tecnológicas empregadas pelo depósito para atingir os seus objectivos de preservação e acesso de longo prazo.

Deste modo o modelo fornece uma estrutura útil de desenvolvimento de metadados para a preservação que vai ao encontro dos requisitos necessários a uma actividade de preservação digital de âmbito alargado.

O modelo de referência OAIS é, neste momento, a base de trabalho das instituições de maior renome internacional na área da preservação digital, através de projectos já desenvolvidos em comunidades que detêm todo o tipo de recursos digitais, tais como: NEDLIB (Networked European Deposit Library), CEDARS (CURL Exemplars in Digital Archives), PANDORA (Preserving and Accessing Networked Documentary Resources of Australia) e OCLC/RLG (Online Computer Library Center/ Research Libraries Group).

A comunidade CEDARS é a que maior esforço desenvolveu e como já referimos segue este modelo de perto.

A Aplicação do Modelo OAIS pela Comunidade CEDARS

O projecto CEDARS desenvolve-se no Reino Unido, patrocinado pelo JISC (Joint Information System Committee) [7] através do programa "eLib – The electronic libraries programme" sob proposta do consórcio de bibliotecas universitárias CURL (Consortium of University Research Libraries)[8].

Apresentamos a estrutura dos requisitos de metadados para um pacote de informação para depósito em aplicação nesta comunidade [9] :

Informação descritiva para preservação

- Informação sobre a referência

- Descrição do recurso

- Metadados existentes

- Registos existentes

Informação sobre o contexto

- Informação sobre outros objectos de informação

Informação sobre a proveniência

- História da origem
- Informação da gestão
- Gestão de direitos

Informação sobre a autenticidade

- Indicadores de autenticação

Informação sobre o conteúdo

- Informação sobre a representação
- Objecto-dados

Em ambiente CEDARS o conjunto de metadados está preparado para cumprir, além da sua principal missão de preservação, o acesso aos conteúdos do depósito e inclui metadados descritivos, administrativos, técnicos e legais.

Os metadados são definidos para aplicar a um leque alargado de objectos digitais na expectativa de que uma biblioteca digital contenha colecções constituídas por variados tipos de recursos.

Da mesma forma as especificações devem ser independentes do nível de granularidade aos quais os metadados estão associados.

CONSIDERAÇÕES FINAIS

O desafio da preservação digital é enfrentado pelas diferentes comunidades através da criação de um grande número de *standards*. Alguns deles já manifestaram a expressão da sua qualidade, outros tantos estarão em preparação acompanhando a evolução das próprias comunidades.

No que diz respeito às metodologias de implementação, estas devem ter em especial atenção a confiabilidade do repositório e do produto final que será a herança digital.

O enquadramento genérico proporcionado pelo modelo OAIS aberto a qualquer comunidade proporciona interoperabilidade entre sistemas e é um grande passo para a concretização da possibilidade de aproveitar a convergência de objectivos comuns e ao mesmo tempo respeitar a diversidade entre comunidades.

NOTAS

1. CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS – Reference Model for an Open archive Information System (OAIS), Blue Book (2002). [Em linha] [Consult. 22 Jan. 2003] URL : http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf

2. OCLC/RLG WORKING GROUP ON PRESERVATION METADATA – Preservation metadata and the OAIS Information Model : a metadata framework to support the preservation of digital objects : a report (2002). [Em linha] [Consult. 22 Jan. 2003] URL : http://oclc.org/research/pmwg/

3. EITELJORG II, Harrison *et al*. - Archaeology Data Service CAD : A Guide to Good Practice : AHDS. (2002)). [Em linha] [Consult. 22 Jan. 2003] URL : http://ads.ahds.ac.uk/project/goodguides/cad/

4. METS Metadata Encoding & Transmission Standard : Official Web Site. [Em linha] [Consult. 27 Jan. 2003] URL : http://www.loc.gov/standards/mets/

5. Digital Library Federation. [Em linha] [Consult. 27 Jan. 2003] URL : http://www.diglib.org/

6. CEDARS Guide to Preservation Metadata (2002)

7. The Joint Information Systems Committee. [Em linha] [Consult. 27 Jan. 2003] URL : http://www.jisc.ac.uk/

8. Consortium of University Research Libraries. [Em linha] [Consult. 27 Jan. 2003] URL : http://www.curl.ac.uk/

9. Cedars Guide to Preservation Metadata. . [Em linha] [Consult. 27 Jan. 2003] URL : http://www.leeds.ac.uk/cedars/guideto/metadata/guidetometadata.pdf