Universidade do Minho
Instituto de Letras e Ciências Humanas

CEHUM
CENTRO DE ESTUDOS HUMANÍSTICOS
DA UNIVERSIDADE DO MINHO

A preservação de património científico nas Humanidades
A aplicação da *TEI – Text Encoding Initiative*

IDALETE DIAS • idalete@ilch.uminho.pt

Instituto de Letras e Ciências Humanas
Universidade do Minho
Centro de Estudos Humanísticos

Encontro Arquivos Científicos

FCSH – UNL • 4 de julho 2014

Índice

1) O Conceito de Metadados
   - Problemas e desafios
   - Tipologia de Metadados
2) A *Text Encoding Initiative* (TEI)
   - Funcionamento da TEI
   - Potencial da TEI aplicada a arquivos científicos (dados textuais; áudio; vídeo; gráficos; etc.)

O Conceito de METADADOS

- Os metadados têm várias funções e vários propósitos.
- Metadados = Informação estruturada que:
  - descreve o recurso de informação em termos de conteúdo, contexto e estrutura (metadados de preservação);
  - ajuda a localizar/identificar o recurso;
  - ajuda a extrair/recuperar, utilizar e gerir o recurso de informação;
  - esclarece questões de autenticidade, autoridade, acessibilidade, interoperabilidade, usabilidade;
  - identifica relações estruturais internas e entre objetos de informação.
- Quanto mais rica e consistente for a descrição do objeto de informação, mais facilmente essa estrutura pode ser pesquisada, manipulada, interrelacionada com outros objetos de informação, etc..

PROBLEMAS & DESAFIOS (1)

- Não existe um padrão de METADADOS capaz de atender a todas as necessidades de preservação dos diferentes tipos de recursos e contextos digitais.
- Existem vários esquemas e padrões de metadados desenvolvidos por comunidades diferentes.
  - Alguns destes padrões possuem um conjunto de dados nucleares comuns (título, data, criador, etc.)
  - TEI: *Text Encoding Initiative*
  - DC: *Dublin Core*
  - METS: *Metadata Encoding and Transmission Standard*

PROBLEMAS & DESAFIOS (2)

- Antecipar a informação que será necessária para atingir determinados objetivos de preservação.

- Identificar o(s) esquema(s) que melhor poderá/poderão corresponder:

  (i) às necessidades do gestor da informação;

  (ii) às propriedades do objeto de informação (*object-oriented*);

  (iii) às propriedades do repositório/acervo no qual o objeto está inserido;

  (iv) às necessidades dos utilizadores (*user-oriented*).

- Definir os tipos de METADADOS a incluir e o nível de granularidade para atingir os objetivos propostos.

Tipologia de METADADOS

- METADADOS administrativos
- METADADOS descritivos
- METADADOS de preservação
- METADADOS técnicos
- METADADOS de uso

METADADOS Administrativos (1)

- Identificação do objeto digital (e respetivo original) no repositório / coleção:
  - repositório, coleção, número de identificação
  - entidade responsável pela publicação / distribuição
  - direitos de autor /utilização /reprodução

- Menção de responsabilidade:
  - Quem criou o objeto digital e quando?
  - Quem é responsável pela preservação e manutenção do objeto digital?

METADADOS Administrativos (2)

- Historial do objeto a preservar:
  - História do objeto antes da sua aquisição (origem, proveniência)
  - Processo e termos da aquisição
- Registo de outras versões/reproduções do objeto
  - Existência de facsimile de documento manuscrito,…

METADADOS descritivos

- Identificação e descrição do(s) objeto(s) e/ou coleções
  - Descrição do objeto original (ex. manuscrito):
    - tipo textual;
    - localização geográfica;
    - descrição do documento: suporte físico; dimensões; condição/estado do documento;
    - elementos não textuais relacionados com o ato da escrita (esquemas, fórmulas, gráficos, desenhos, etc.)
    - parafernália
    - sobrescrito (correspondência)
    - resumo do conteúdo

METADADOS de Preservação

- registar todo o processo de criação de um objeto digital a partir do objeto original.
- documentar:
  - (i) os processos de preservação digital de longo prazo, garantindo que os conteúdos digitais possam ser consultados e interpretados no futuro;
  - (ii) decisões e ações/técnicas de preservação e edição adotadas – métodos e estratégias tomadas para a preservação;
  - (iii) efeitos da conversão de dados.

METADADOS técnicos

- Documentação de tecnologias utilizadas para criar /aceder ao objeto digital:
  - *hardware* e *software* utilizados (versões, etc.)
  - migração de conteúdo
  - processo(s) de digitalização: formato, resolução, esquemas de compressão, …

METADADOS de Uso

- Gestão de direitos autorais e propriedade intelectual
- Autenticidade do recurso digital

Text Encoding Initiative (*TEI*)
Introdução

"The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation." (*Página oficial da TEI*)

- Association for Computers in the Humanities, the Association for Literary and Linguistic Computing, and the Association for Computational Linguistics

- **Objetivo:**
  criação de diretrizes para a preservação digital de documentos nas Humanidades (*TEI Guidelines for Electronic Text Encoding and Interchange*)

- **Página *Web*:** http://www.tei-c.org/

Text Encoding Initiative (*TEI*)
Vantagens (1)

- Fornece propostas de anotação para vários tipos textuais:
  - facilita o trabalho de preservação desde a criação dos metadados à anotação do contéudo do objeto a preservar;

  - facilita o intercâmbio de informações em suporte digital:
    - **anotação descritiva:**

      - classifica os componentes do texto de acordo com a sua função na estrutura lógica do documento;

      - o mesmo documento pode ser processado, sem nenhuma alteração, por vários processadores diferentes;

      - cada processador/programa/aplicação pode tratar apenas as partes relevantes para determinado *output*/objetivo.

Text Encoding Initiative (*TEI*)
Vantagens (2)

➢ facilita o intercâmbio de informações em suporte digital:

❖ **A TEI baseia-se na metalinguagem de anotação XML (*eXtensible Markup Language*):**

- Conjunto de regras sintáticas de anotação/etiquetagem

- XML concebe os documentos como sendo instâncias de determinado tipo textual:

- Se os documentos pertencem a determinados tipos, é possível **definir** a sua estrutura formal → Definição do Tipo de Documento / *Document Type Definition* (**DTD**).
- A partir da definição da estrutura formal do tipo de documento é possível **validar**, i.e., controlar se o documento obedece à estrutura formal definida.

- A XML é independente de plataformas e aplicações → **interoperabilidade**

XML – Extensible Markup Language

"Currently, a number of evolving technologies are influencing the way scholarly communication is carried out. The "extensible markup language" (XML) data format is gaining ground as is the establishment of institutional repositories as part of a digital preservation strategy. The relevance of the standardized XML data format lies in its proclaimed non-proprietary, self-describing features. Storing digital objects as XML files has been recognized as a real possibility for both long-term storage and access to the data they represent. Apart from migration of the files to more recent formats and emulation of extinct applications and operating systems, XML has been presented as a possible approach to prevent the files from becoming an uninterpretable clump."

(Van Nispen, *et al*., 2005)

A estrutura de um documento TEI (1)

- **TEI header / Cabeçalho TEI:**
  - documenta o processo de criação do documento digital a partir do objecto original.
  - **<teiHeader>** Descrição do documento eletrónico e do documento original **</teiHeader>**

- **TEI text / Texto TEI:**
  - **<text>** o texto anotado **</text>**

A estrutura de um documento TEI (2)

```
<TEI.2>

<teiHeader> Cabeçalho do documento </teiHeader>

<text>
        <front> matéria preliminar </front>*
        <body> corpo do texto </body>
        <back> matéria posposta ao corpo </back>*
</text>

</TEI.2>
```

* Elementos opcionais

TEI Header/Cabeçalho TEI

O <teiHeader> é composto por 4 partes:

1) <fileDesc> → File Description

- Descrição do documento eletrónico e a sua relação com o texto original.

2) <encodingDesc> → Encoding Description*

- Descrição das práticas de anotação e edição implementadas.
- Registo de problemas e soluções de codificação, transcrição, etc.

3) **<profileDesc> → Profile Description***

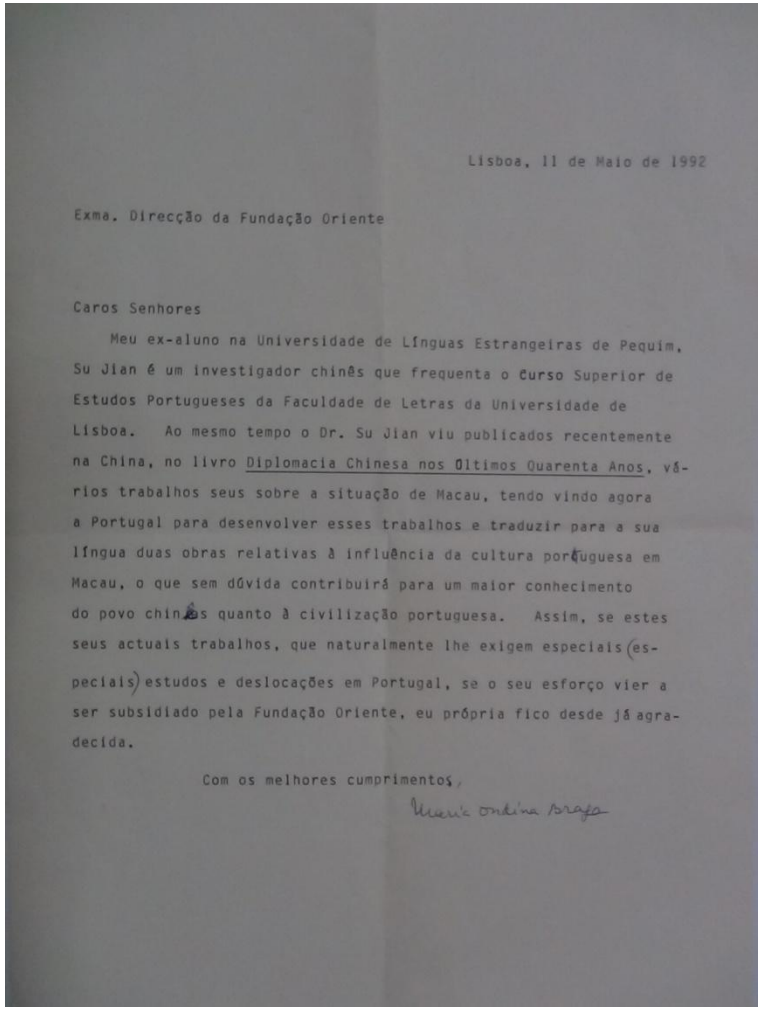
- Descrição de aspetos do texto eletrónico úteis para fins investigativos: língua(s), variedade linguística, registo(s); contexto de criação; intervenientes; *setting*.

4) **<revisionDesc> → Revision Description***

- Descrição do historial das alterações e correções editoriais efetuadas ao texto eletrónico (responsável, data, motivo que originou alterações).

*** Elementos opcionais**

TEI Header: File Description



Lisboa, 11 de Maio de 1992

Exma. Direcção da Fundação Oriente

Caros Senhores

Meu ex-aluno na Universidade de Línguas Estrangeiras de Pequim, Su Jian é um investigador chinês que frequenta o Curso Superior de Estudos Portugueses da Faculdade de Letras da Universidade de Lisboa. Ao mesmo tempo o Dr. Su Jian viu publicados recentemente na China, no livro Diplomacia Chinesa nos Últimos Quarenta Anos, vários trabalhos seus sobre a situação de Macau, tendo vindo agora a Portugal para desenvolver esses trabalhos e traduzir para a sua língua duas obras relativas à influência da cultura portuguesa em Macau, o que sem dúvida contribuirá para um maior conhecimento do povo chinês quanto à civilização portuguesa. Assim, se estes seus actuais trabalhos, que naturalmente lhe exigem especiais (especiais) estudos e deslocações em Portugal, se o seu esforço vier a ser subsidiado pela Fundação Oriente, eu própria fico desde já agradecida.

Com os melhores cumprimentos,

Maria Ondina Braga

```
<teiHeader>

    <fileDesc>
        <titleStmt> (...) </titleStmt>

        <publicationStmt> (...)
        </publicationStmt>

        <sourceDesc>
            <letDesc>
                (...)
            </letDesc>
        </sourceDesc>
    </fileDesc>

</teiHeader>
```

Digital Archive of Letters in Flanders (DALF)

Carta da escritora bracarense Maria Ondina Braga dirigida à Direcção da Fundação Oriente
Museu Nogueira da Silva • Braga

TEI Header: File Description

- **Elemento obrigatório do <teiHeader>.**
- **Fornece dados bibliográficos sobre:**
  - o texto eletrónico;
  - a fonte/documento original.
- **As informações do <fileDesc> podem ser utilizadas:**
  - para catalogar, identificar e recuperar o documento;
  - por investigadores que pretendam consultar o documento eletrónico e/ou documento original (facsimile).

O <fileDesc> pode ser comparado à página de rosto de um documento impresso.

Projetos Internacionais: Arquivos Digitais

- **Arquivo de texto:**

»» *Vincent van Gogh - The Letters:* http://vangoghletters.org/vg/

- **Arquivos de imagem:**

»» Ad Access: http://library.duke.edu/digitalcollections/adaccess/

- **Arquivo de vídeo:**

»» Shoah Visual History: http://vhaonline.usc.edu/

- **Arquivo de áudio:**

»» Historical voices: http://www.historicalvoices.org/

REFERÊNCIAS

- Giordano, Richard (1995), “The TEI Header and the documentation of Electronic Texts”. In: Nancy Ide/Jean Véronis, *Text Encoding Initiative. Background and Context*, Kluwer Academic Publishers, Dordrecht/Boston/London, 75-83.

- Qin, Jian and Li, Kai (2013), “How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure. Paper presented at International Conference on Dublin Core and Metadata Applications, Lisbon, Portugal, September 2-6, 2013.

- Van Nispen, Annelies & Rutger Kramer & René Van Horik (2005), “The eXtensible Past. The Relevance of the XML Data Format for Access to Historical Datasets and a Strategy for Digital Preservation”, in *D-Lib Magazine* – Vol. 11 No. 2.
  Disponível em : http://www.dlib.org/dlib/february05/vannispen/02vannispen.html

- World Wide Web Consortium (W3C):
  http://www.w3.org/