



Universidade
do Minho

Seminário
Lic. Ciência da Informação



Três anos depois... *...uma reflexão sobre o projecto DigitArq*

Miguel Ferreira
mferreira@dsi.uminho.pt



2006-03-27



Conteúdo

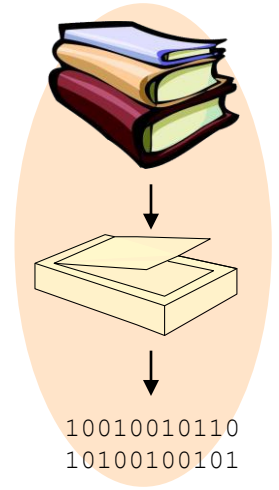
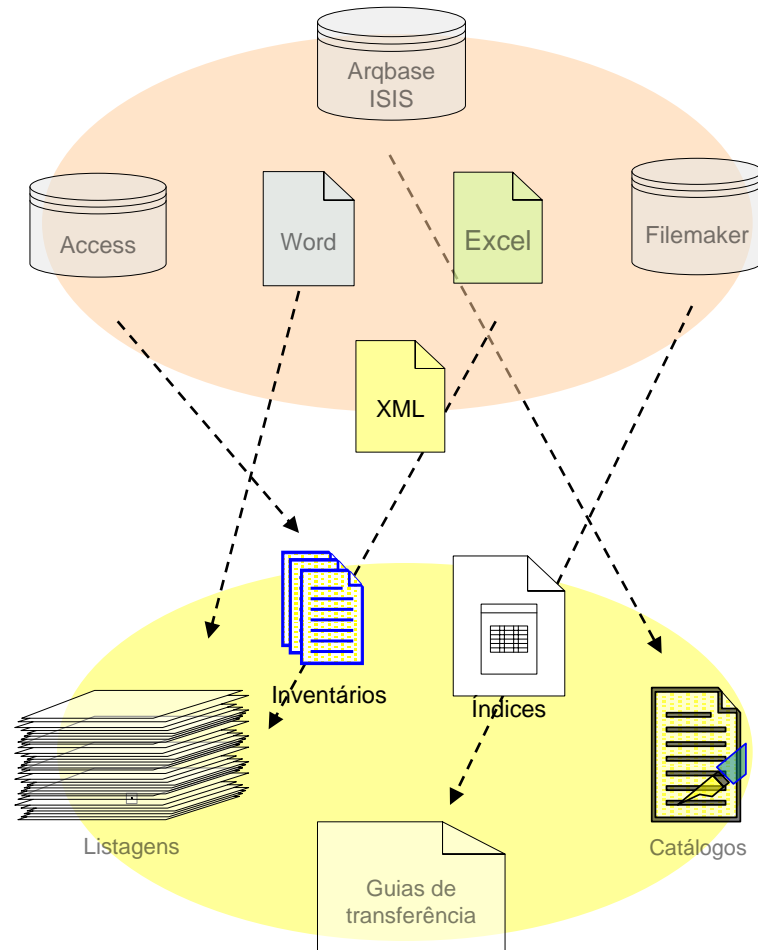
- **Enquadramento e objectivos** do projecto
- Introdução à **descrição arquivística**
- Fase 1: **Migração** aux. pesq. digitais
- Fase 2: **Migração** aux. pesq. em papel
- Fase 3: Módulo de **descrição**
- Fase 4: Software de **aquisições**
- Fase 5: Módulo de **acesso Web**
- Fase 6: Gestão de **objectos digitais**
- **Metainformação** usada pelo GOD
- Notas quanto ao uso de **XML/EAD**
- Algumas reflexões sobre o **passado...**
- Algumas reflexões sobre o **futuro...**



Enquadramento do projecto



Arquivista



Utente



Objectivos do projecto

- **Eliminação** do papel
- **Centralização** da informação
- Utilização de **normas** internacionais
 - *International Standard Archival Description* - ISAD(g)
 - *Encoded Archival Description* - EAD/XML
- Gestão de **descrições** arquivísticas
- Gestão de **objectos digitais**
- Permitir o **acesso** via **Web** ao acervo do arquivo



Parâmetros do projecto

- **Equipa** de desenvolvimento
 - 3 informáticos
 - 2 arquivistas
- **Duração**
 - 9 meses
- Equipa de **gestão**
 - 1 gestor financeiro (arquivista)
 - 1 gestor de projecto (arquivista)
 - 1 coordenador informático



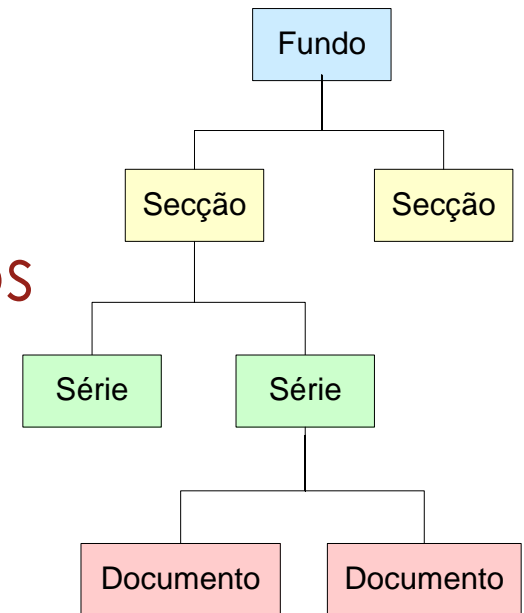
Auxiliares de pesquisa

- Metainformação descritiva
- Permite o acesso à informação
- Descrita segundo as normas ISAD(g) e EAD
 - Crosswalks em <http://www.loc.gov/ead/ag/agappb.html>
- Princípio da **proveniência**
 - *Respect des fonds*
 - Agregação de documentos com a mesma proveniência
 - A base da **ciência arquivística** actual



Organização da metainformação

- Estrutura **hierárquica**
- Descrição do **mais geral** para o **mais específico**
- Diferentes **níveis** descritivos
- **Elementos** descritivos:
 - Referência
 - título
 - datas extremas
 - condições de acesso
 - âmbito e conteúdo
 - ...



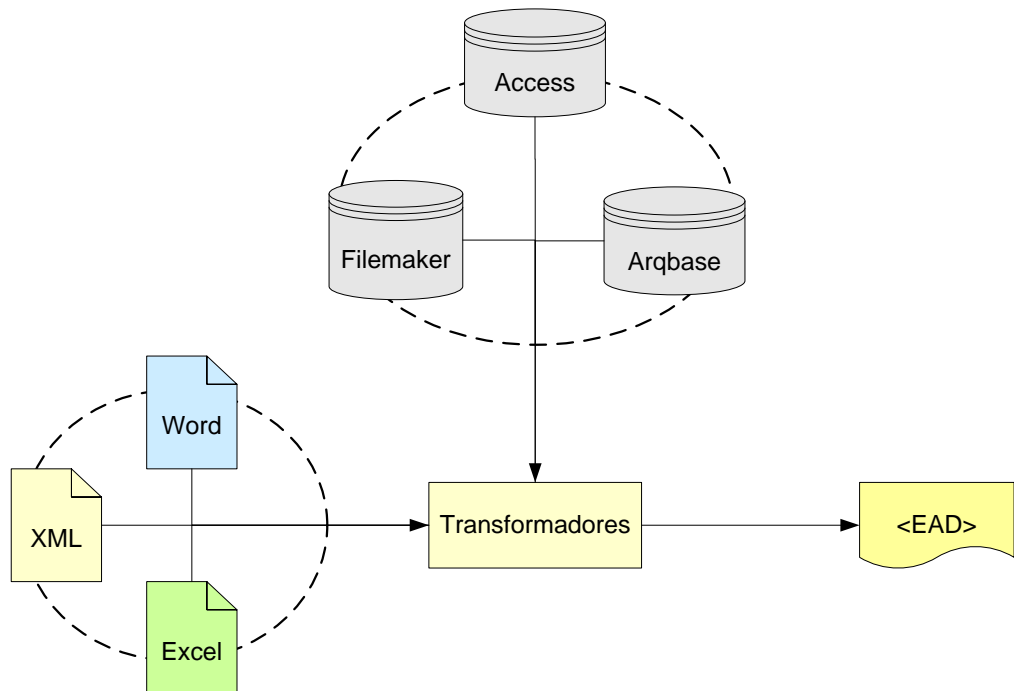


Fase 1: Homogeneização de AUX.PESQ digitais

1. **Exportação** das BD para texto (e.g CSV, XML)

2. **Transformação** para EAD

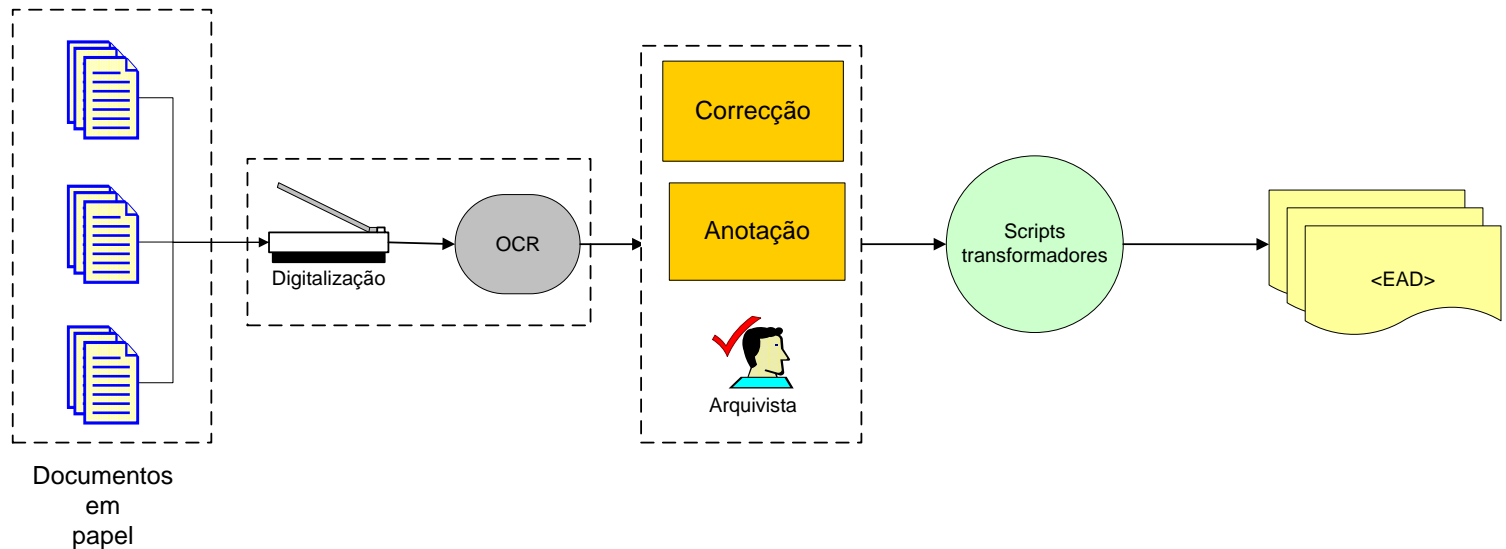
- Scripts *Perl* (expressões regulares)
- XSLT
- XML-DT
- ...





Fase 2: Homogeneização de AUX.PESQ em papel

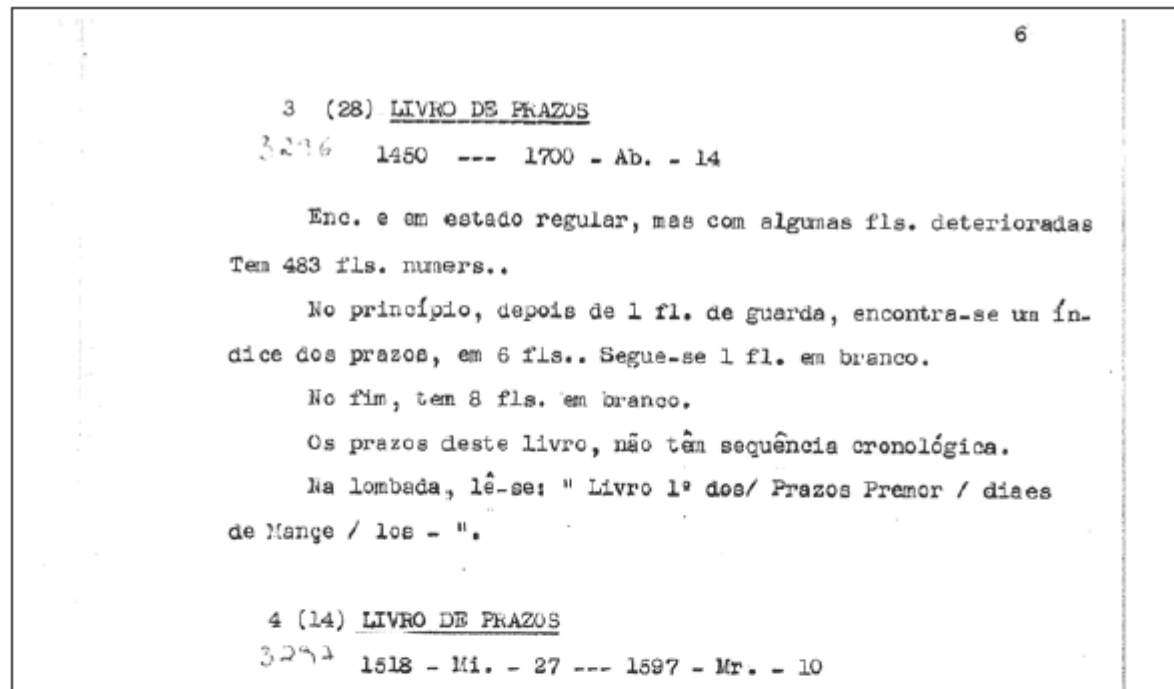
1. **Digitalização**
2. Identificação de **modelos** (3 modelos)
3. **Reconhecimento** de caracteres
4. **Correcção** (processo manual)
5. **Anotação** (processo manual)
6. **Conversão** para EAD recorrendo a *scripts Perl*





MODELO A

- Muito **pouco estruturado**
- **Informação** altamente **variável**
- Anotação **XML** de acordo com um *schema*





MODELO A: original, reconhecimento e anotação

		6
	3 (28) <u>LIVRO DE PRAZOS</u>	
	3296 1450 --- 1700 - Ab. - 14	
	3 (28) LIVRO DE PRAZOS 1450 --- 1700 - Ab. - 14	
Tem	Enc. e em estado regular, mas com algumas fls. deterioradas Tem 483 fls. nunsers..	
dic	No principio, depois de 1 fl. de guarda, encontra-se um Ln_ dice dos prazos, em 6 fis.. Segue-se 1 fl. em branco. No fim7 tacq 8 fls. 'er-ri branco. Os prazos deste livro, não tem sequência cronológica.	
	<pre> <registro> <unitid> 3 </unitid> (28) <unittitle>LIVRO DE PRAZOS </unittitle> <unitdate_inicial>1450 </unitdate_inicial>--- <unitdate_final>1700 - Ab. - 14 </unitdate_final> <phystech>Enc. e em estado regular, mas com algumas fls. deterioradas</phystech> etc... </registro> </pre>	
	<p>>o principio, depois de 1 fl. de guarda, encontra-se um in~ dice dos prazos, ma 7 fls. . Segue-se 1 fl. em branco. No fim, tem 1 fl. de guarda. Os prazos deste livro, nem sempre se seguem por ordem cronologica.</p>	
mas	5 (3) LIVRO DE PRAZOS ,351540-J1.-19---1682-Ag. 0" - 13	
dic	Enc. e em estado regular, mas com falta de muitas fls.. Tem as fis. numers. de 1 a 294, faltando, ou estando reduzida s a pequenos fragmentosq aquelas a que devia corresponder a	
lóg	numeração , de 7& a 73, 93 a 102, 160 a 162, e 209 a 211, inclusi ve, No fim, encontram-se tambéjn fragmentos de mais 40 fls..	
	No principio, tem 3 Vs. de guarda.	
	3. Os prazos deste livro não se seguem por ordem cronológica	



Modelo B

- Bem estruturado – em tabela
- Elementos de informação estáveis
- Anotação baseada em estados
 - A abertura de uma etiqueta define um novo estado de interpretação para o *parser*

	Nº DE ORDEM	DATAS EXTREMAS		Nº. DE FOLHAS	OBSERVAÇÕES
34 89	126	1880-N. -27	-- 1881-Ja. -20	50	B.
	127	1881-Ja. -21	-- 1881-Mr. -29	52	B.
	128	1881-Ab. - 3	-- 1881-S. -14	100	B.
	129	1881-S. -15	-- 1882-Ja. -23	50	B.
	130	1882-Ja. -23	-- 1882-Mi. -24	100	B.
	131	1882-Mi. -25	-- 1882-O. - 8	100	B.
	132	1882-O. - 9	-- 1882-D. -27	98	B.
	133	1882-D. -27	-- 1883-Mr. -19	98	B.
	134	1883-Mr. -19	-- 1883-Jl. -23	100	B.
	135	1883-Jl. -25	-- 1883-N. -20	100	B.
	136	1883-N. -21	-- 1884-Ja. -25	100	B.

Tabela



Modelo B: original, reconhecimento e anotação

Nº DE ORDEM	DATA	EXTREMAS	Nº. DE FOLHAS	OBSERVAÇÕES
34 89	126	1880-N. -27 -- 1881-Ja.-20	50	B.
	127	1		
	128	1		
	129	1		
	130	1		
	131	J		
	132	J		
	133	J		
	134	J		
	135	J		
	136	J		
		<u>MODC</u>		
		<u>F</u> CN-CNVCD02		
		<u>SR</u> 001		
		<u>TI</u> Livros de Notas		
		<u>COTA</u> I/102/12 CX 12		
	126	1880-14. .27 -r 1881-Ja.-20	50	B,
	127-	1881 Ja. - 21 .».. 1881-Mr. - 29	52	B.
	128	1881-Ab. - 3 -- 1881-S. -14	100	B.
	12.9	1881-S.. -15 -- 1882-Ja. -23	50	B.
	130	1882-Ja. -23 -d 1882-Mi.-24	100	B.
		<u>COTAAUTO</u> I/102/12 CX 13		
	131	188-25 -- 1882-0. - 8	100	B,
	132	1882-0. - 9 1882-D. -27	98	B,
	133	1882-D. -27 -r 1887-Mr.-19	98	B.
	13.4.	1883-Mr, -19 -- 1883-J1.-23	100	B.
	135	1883-J1.-25 N 1883-N. -20	100	B.
	136	1883-17. -21 -- 1884-Ja.-28	100	B.



Modelo C

- **Minimamente estruturado**
- Elementos de **informação pouco variáveis**
- **Anotação não baseada em estados**
 - Uma etiqueta por cada elementos de informação

LIVROS DE REGISTOS	
3a s?	<p>1 Livro de registos - 1843-N.-13 -- 1854-Ja.-30 - Cad. de 98 fls. numera., enc., em estado regular.</p> <p>Só está escrito até fls. 92 v.</p> <p>O primeiro registo refere-se a José Gonçalves Amorim, da freguesia de Macieira, e o último a Manuel Pinto Ribeiro, da Póvoa.</p> <p>Na capa: "Escrivão Silva / Registo das Procurações / e mais actos praticados / fora da Nota / nº 1"</p>
TEXTO	<p>2 Livro de registos - 1857-Mi.-11 -- 1881-Ja.-7 - Cad. de 50 fls. numera., enc., em bom estado.</p> <p>Só está escrito até fls. 48 v.</p> <p>O primeiro registo refere-se a João José Carneiro Veloso, de Vila Nova de Famalicão, e o último a José de Almeida Torres, de Santa Cristina de Malta.</p> <p>Na capa: "Registo / das / Procurações / e protes-</p>



Modelo C: original, reconhecimento e anotação

LIVROS DE REGISTOS

3a s? 1 Livro de registos - 1843-N.-13 -- 1854-Ja.-30 - Cad. de 98 fls. numer., enc., em estado regular.

Só está escrito até fls. 92 v.

O primeiro registo refere-se a José Gonçalves Amorim da freguesia de Macieira, e o último a Manuel

1 livro de registos, -, 1843-N.-13 -- 1854-Ja, -30 w Cad., de 98 fls. numer., í enc., , , ffiz estado regular,

Só esta. escrito até fls, 92 v, '

0 pr
Maci MODE
F NOT-CNVCD02

Na c
prat SR 010
COTA I/23/23 CX 12

2 Li
nume ID 1
TI livro de registos
DATAS 1843-N.-13 -- 1854-Ja-30
CARF Estado regular,

Só e DIM 98 fls numeradas
TUI Caderno

o pr
Nove OBS Só está escrito até fls, 92 v.
de M AMBCONT O primeiro registo refere-sie a José Gonçalves Amorim, da freguesia de Macieira, e o último a Mandeir Pinto Ribeiro, da Póvoa.

Na c TIPROP Na capa: "Escrivão Silva / Registo das Procuções / e anais
tos actos praticados / fora da Nota / nQ 1"

3 #
nume ID 2
TI Livro de registos

0 DATAS 1857-Mi,-11 -- 1831-Ja,-7

Cond TUI Cad



Fase 2: Algumas conclusões

$$t_{conversion} = t_{recognition+correction} + t_{anotation}$$

Equation 1: Simplified formula for calculating the conversion time.

Model	Statistics per Page			Statistics per Record		
	$t_{rec+corr}$ (minutes)	$t_{annotation}$ (minutes)	$t_{conversion}$ (minutes)	$t_{rec+corr}$ (minutes)	$t_{annotation}$ (minutes)	$t_{conversion}$ (minutes)
Model B						
Model C						
1	13,5	19,5	33,0	5,8	8,4	14,3
2	17,7	14,5	32,1	7,5	6,1	13,6
3	14,9	10,9	25,8	6,5	4,8	11,3
4	11,3	14,5	25,8	4,4	5,7	10,0
5	11,4	14,6	26,0	4,0	5,2	9,2
6	7,2	7,8	15,0	2,8	3,1	5,9
Average	12,7	13,6	26,3	5,2	5,5	10,7

Table 3: Conversion statistics for finding aids of model C.

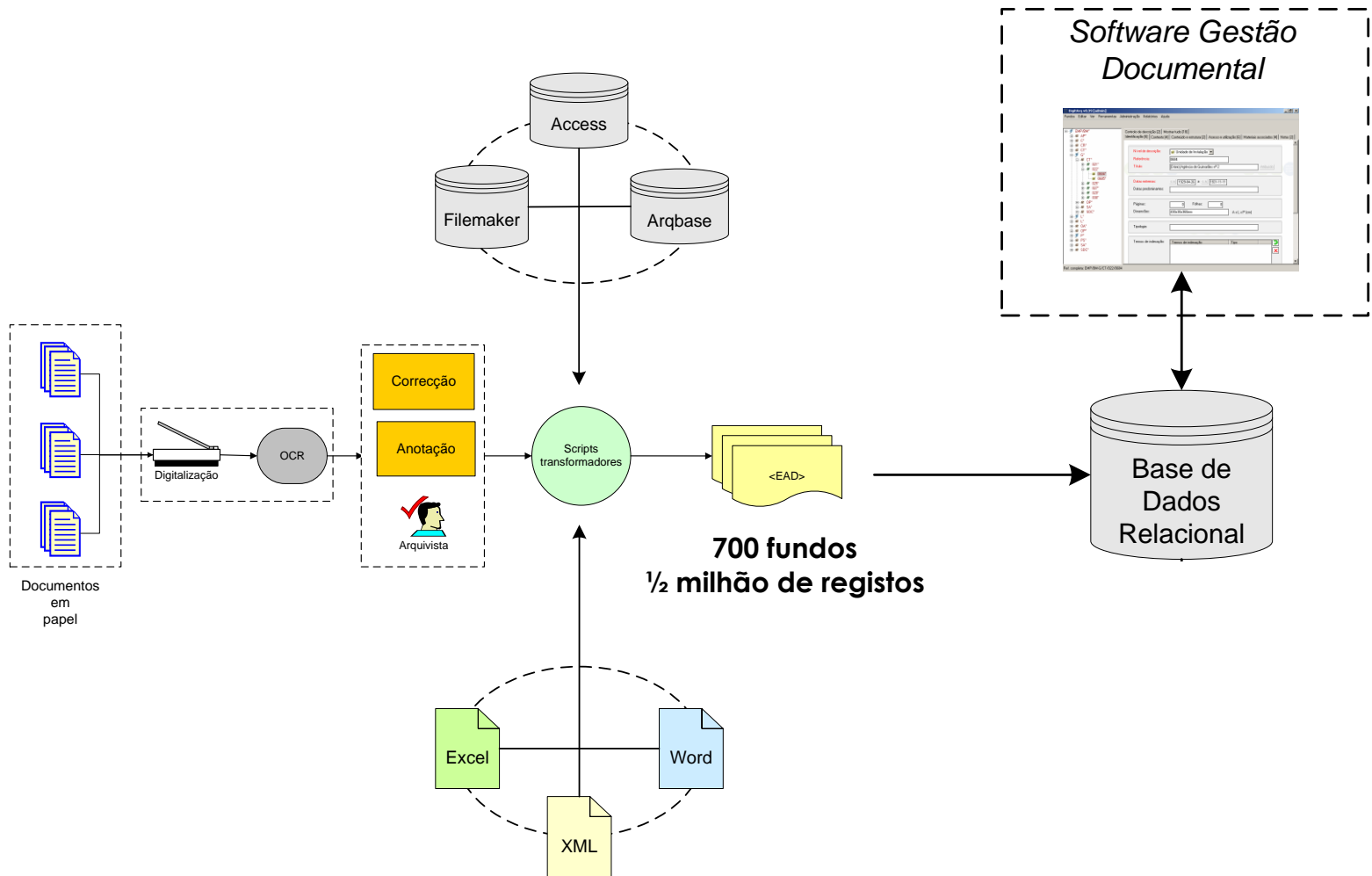


Fase 2: Algumas conclusões

- **Tempo** médio de conversão
 - reconhecimento + correcção + anotação
 - 22,3 min/pag
- Sem dados comparativos em relação à **transcrição manual**
- Na opinião dos técnicos de arquivo...
 - Apesar da necessidade de aprender **novas ferramentas**
 - **Menos aborrecido** e fatigante
 - **Menos erros** humanos



Fase 3: Módulo de descrição





Fase 3: Módulo de descrição

- **Referências** relativas
 - Evita erros durante a descrição
- Restrições à **hierarquia**
 - Evita erros durante a descrição
- **Detecção de erros** nas descrições
 - Datas omissas, datas inicial/final trocadas, campos obrigatórios não preenchidos, ...
- **Inferência** a partir de níveis inferiores
 - Datas extremas, nº de unidades de instalação, ...
- **Drag & drop**
 - Simplifica a descrição de fundos regulares
- Trabalho **cooperativo**
 - Várias pessoas podem trabalhar no mesmo fundo em simultâneo



Fase 3: Módulo de descrição

- **Relatórios**
 - 13 relatórios para auxílio da gestão
- **Registo de actividades**
 - Produção de estatísticas
 - Quantos registos são produzidos por dia
 - N° de registos por fundo
 - Grau de completude das descrições
- **Vocabulários controlados** (2 níveis)
- Registo de **autoridade** (EAC – Encoded Archival Context)
- **Importação/Exportação**
 - EAD
 - CSV
 - DScibe CALM Natural Format



Fase 3: Módulo de descrição

DigitArq v0.39 [admin]
Fundos Editar Ver Ferramentas Administração Relatórios Ajuda

Controlo de descrição [2] | Mostrar tudo [18]
Identificação [6] | Contexto [4] | Conteúdo e estrutura [2] | Acesso e utilização [6] | Materiais associados [4] | Notas [2]

Nível de descrição: Unidade de Instalação

Referência:

Título:

Datas extremas: c.a. a c.a.

Datas predominantes:

Páginas: **Folhas:**

Dimensões: A x L x P (cm)

Tipologia:

Termos de indexação:

Termos de indexação	Tipo	
		<input type="button" value="v"/>
		<input type="button" value="x"/>

Ref. completa: EMP/BM-G/CT/022/0684



Fase 4: Módulo de aquisições

- Versão simplificada do módulo de **descrição**
 - **Menos elementos** descritivos
 - **Menos funcionalidades**
 - **Emissão de documentos** para validar as incorporações
- **Modelos** de fundos
 - Paroquiais
 - Notariais
 - Judiciais
- **Disseminação** junto das organizações que enviam documentação para o Arquivo



Fase 4: Módulo de aquisições

The screenshot displays the 'Software de Aquisições 2.1 [mferreira]' application interface. On the left, a tree view shows a hierarchy of folders: EMP/BM, AP, [novo]*, 018, 091, 067, 099, 006, 005, 004, SA, CR, CT, QA, P, C, OP, L, G, SDC, PS, and L. The main window shows a 'Pré-visualizar relatório...' dialog box with a 'MainReport' tab. The report content includes the 'digit Arq' logo, the title 'Auto de Entrega', and two paragraphs of text. The first paragraph describes the incorporation of documentation from the University of Minho into the Aveiro District Archive. The second paragraph states that the documentation will be under the custody of the Aveiro District Archive and subject to its internal regulations. The report footer shows 'Current Page No: 1', 'Total Page No: 1', and 'Zoom Factor: 100%'. The Windows taskbar at the bottom shows the Start button, several open applications, and the system tray with the time 15:37 on Wednesday.

Software de Aquisições 2.1 [mferreira]
Fundos Editar Ver Ferramentas Relatórios Ajuda

EMP/BM
AP
[novo]*
018
091
067
099
006
005
004
SA
CR
CT
QA
P
C
OP
L
G
SDC
PS
L

Descrição [5]
Nível de descrição: Série

Pré-visualizar relatório...
MainReport

digit Arq
Software de Aquisições

Auto de Entrega

Ao(s) vinte e três dia(s) do mês de Dezembro de dois mil e cinco, no Arquivo Distrital de Aveiro perante Miguel Ferreira e António Sousa, procedeu-se à incorporação da documentação proveniente de Universidade do Minho, conforme o constante na guia de remessa anexa que, rubricada e autenticada por estes representantes, fica a fazer parte integrante deste auto.

O referido conjunto documental ficará sob a custódia do Arquivo Distrital de Aveiro e a sua utilização sujeita aos regulamentos internos, podendo ser objecto do necessário tratamento arquivístico no que respeita à conservação, acessibilidade e sua comunicação.

Current Page No: 1 Total Page No: 1 Zoom Factor: 100%

Ref. completa: EMP/BM/AP/018

Start 00-studio 01-paper 3 Microsoft O... Plano Tecnológ... Adobe Acrobat... Software de A... Pré-visualiza... 15:37 terça-feira



Fase 5: Módulo de acesso Web

Selecione os níveis de descrição ⓘ

ADP2004P9999OD1

- Proposta
 - 000001.tif
 - 000002.tif
 - 000003.tif
- expropriação1
 - 000004.tif
 - 000005.tif
 - 000006.tif
 - 000007.tif
 - 000008.tif
 - 000009.tif
- faturas
- Correspondência
- Verbas
- Relatórios da Brigada
- Expropriação2
- Homologação
- Fase de Atravanc
 - 000000.tif

Meta Informação da Imagem Matriz:

Esquema de Cor:	Gray
Resolução Espacial:	314
Profundidade Bits:	8
Largura:	5756
Altura:	2710
Tamanho:	5225

Localização física I/16/2 - 25.14



Fase 6: Gestão de objectos digitais

- Arquivo de **reproduções digitais**
- Gestão de **metainformação** associada aos objectos digitais
- Gestão de **perfis** de digitalização
 - Geração de derivadas para publicação em-linha
- **Transferência** de objectos para **CD**
 - Fora de linha
 - Gestão do espaço de armazenamento
- Funcionalidades básicas de **preservação**
 - Monitorização da **integridade** dos objectos
 - Avisos para **refrescamento** de suporte



Fase 6: Gestão de objectos digitais

- Metainformação **descritiva**
 - Associação a registos do módulo de descrição (ISADg/ÉAD)
- Metainformação **administrativa**
 - Library of Congress Core Metadata elements
- Metainformação **técnica**
 - NISO Z39.87 – 2002 (Technical metadata for digital images)
- Metainformação **de preservação**
 - CEDARS
- Metainformação **estrutural**
 - O METS foi abandonado por ser demasiado complexo para os objectivos do ADP
 - No entanto os objectos digitais são estruturados



Fase 6: Gestão de objectos digitais

The screenshot shows a software application window titled "Pré-visualizar relatório..." (Preview report...). The window contains a report titled "Informação de Preservação" (Preservation Information) for the object "ADP2004P9999OD1". The report includes the following details:

Referência:	AC/FFH-SAALN/OPT-ZIPRT_SRQ/016/17.055.001
Data Integração:	15-03-2004
Localização:	040329_1204; 040329_1231
Operador:	FB
Nº Imagens:	116
Data Migração:	15-03-2009
Data Revisão:	06-02-2006
Historial:	
Norma Preservação:	CEDARS Project - Metadata for Digital Preservation: The Cedars Project Outline Specification
Método Preservação:	

The report also features a logo at the top left with the letters 'D' and 'P' in green circles, and a blue circle with the text 'Associação Digital'.

At the bottom of the window, the status bar shows: "Current Page No: 1", "Total Page No: 1", and "Zoom Factor: 100%".



Metainformação administrativa [GOD]

- LC Core Metadata elements
 - Data de produção
 - Data de integração
 - Direitos e permissões de acesso
 - Documentação de apoio
 - Entidade produtora
 - Checksum
 - Dimensão do objecto (bytes)
- <http://www.loc.gov/standards/metadata.html>



Metainformação técnica [GOD]

- NISO Z39.87 – 2002
 - Ambiente tecnológico (SO, dispositivo de captura, fabricante do dispositivo, software associado, ...)
 - Algoritmo de compressão
 - Nível de compressão
 - Espaço de cores (RGB, CMYK, BW,...)
 - Iluminação (câmaras digitais)
 - Resolução espacial (dots per inch)
 - Largura e altura da imagem
 - Profundidade de bits
 - Mime type
 - ...
- http://www.niso.org/standards/resources/Z39_87_trial_use.pdf



Metainformação de preservação [GOD]

- CEDARS

- Metainformação de preservação
 - **Histórico** de acções de reformatação
 - Método de reformatação
 - Informação sobre o **transformador**
 - Plataforma
 - Parâmetros
 - Dispositivo de visualização
 - Estrutura do objecto digital
- <http://www.leeds.ac.uk/cedars/metadata.html>



Metainformação descritiva

- Encoded Archival Description (EAD)
- Desenvolvido ao longo da década de 90
- Formato digital para auxiliares de pesquisa
- Partiu da prática descritiva de vários arquivos da altura
- Resultou num modelo flexível
 - Quase todos os elementos são opcionais



EAD/XML: Um exemplo

- Bando do Minho
 - [EMP-BM.xml](#)



EAD/XML: Algumas decisões

- **Datas** extremas
 - Formato interno vs 2 elementos distintos
 - Datas de comprimento fixo no formato **YYYY-MM-DD** (ISO 8601)
 - Zeros para **datas incompletas**

Hipótese 1:

```
<unitdate> 1436/1441 </unitdate>
```

Hipótese 2:

```
<unidade datechar='initial'> 1436-00-00 </unitdate>
```

```
<unidade datechar='final'> 1441-00-00 </unitdate>
```



EAD/XML: Algumas decisões

- Componentes numerados vs não numerados
 - Elemento agregador dos elementos de cada nível de descrição

Hipótese 1:

```
<c1>, <c2>, ..., <c12>
```

Hipótese 2:

```
<c>  
  <c>  
  </c>  
</c>
```



EAD/XML: Algumas decisões

- Level vs OtherLevel
 - Atributos do elemento <c>

Hipótese 1:

```
<c level='fonds' >  
<c level='series' >  
<c level='subfonds' >  
<c level='recordgrp' >  
<c level='otherlevel' >  
...
```

Hipótese 2:

```
<c level='otherlevel' otherlevel='F' >  
<c level='otherlevel' otherlevel='SF' >  
<c level='otherlevel' otherlevel='SR' >  
<c level='otherlevel' otherlevel='SSR' >  
...
```



EAD/XML: Algumas decisões

- Conteúdos mistos
 - Foram **excluídos** do projecto
 - **Complexidade** adicional ao nível da **interface**
 - **Complexidade** ao nível do modelo da **base de dados**

Exemplo:

```
<scopecontent>
```

```
The founder of this institution,
```

```
<person>John Marshal</person>, was born  
in <date>August 1904</date> and soon...
```

```
</scopecontent>
```



Algumas reflexões... sobre a Migração de Aux. Pesq.

- Melhor **documentação** do processo de **migração**
 - Partilha do conhecimento adquirido
- Utilização de **ferramentas genéricas** baseadas em regras/padrões
 - *ADL_mapper*
 - <http://www.alexandria.ucsb.edu/mm/>
 - Altova MapForce
- A **falta de tempo** e de experiência fez “obrigou” a que assim fosse



Algumas reflexões... sobre a Módulo de Descrição

- **Performance** do modelo de dados
 - Optimização do modelo relacional
 - Utilização de outro tipo de BD
 - BD Orientadas a Objectos
 - BD XML-Nativas
- Melhor implementação de **vocabulários controlados**
- Maior **integração** com o módulo de Gestão de Objectos Digitais
 - No futuro poderão vir a ser uma **aplicação única**



- Pesquisa assistida
 - Condução do utilizador através de um conjunto de painéis que vão **filtrando o espaço de procura**
 - Tentativa de **mimar o diálogo** entre o arquivista e o utente
- Melhoria do componente de **visualização de imagens**
 - Livros digitais
 - Download em PDF e outros formatos à escolha



Rumo ao futuro... o DigitArq 2.0

- Maior disseminação da ferramenta
 - Centro Português de Fotografia
 - Outros arquivos nacionais
- Consulta Real em Ambiente Virtual (CRAV)
 - Comércio electrónico
 - Disponibilização em-linha de grande parte dos serviços do ADP
 - Emissão de **certidões**, **reserva** de documentos, pedido de **pesquisa** por arquivista, compra de **reproduções**
- Arquivo Digital (Torre do Tombo)
 - Repositório com funcionalidades de **preservação** digital
 - Poderá vir a substituir o GOD



Universidade
do Minho

Seminário
Lic. Ciência da Informação



Questões?

Miguel Ferreira
mferreira@dsi.uminho.pt

2006-03-27