ARQUIVO DISTRITAL DO PORTO

PROJECTO DIGITARQ

RELATÓRIO E MÉTRICAS DE APLICAÇÃO DE OCR E ANOTAÇÃO

Relatório final

Porto

2004

Projecto **DigitArq – Produção, conversão e gestão de conteúdos digitais de Arquivo** financiado por:

Programa Operacional da Cultura

FEDER
União Europeia

SUMÁRIO

1. Descrição do fluxo de trabalho

O processo de conversão a OCR seguiu os seguintes passos:

1. Identificação das diferentes estruturas de páginas (*layouts*): foram identificadas 3 estruturas diferentes e uma variante referenciadas como A, C, C2, E.

2. Digitalização dos documentos e processo de OCR com a aplicação OMNIPAGE 9.0.

3. O OCR era realizado através da segmentação da cada página e de cada coluna quando tal se justificasse.

4. Revisão e correcção do texto assim obtido ainda através da aplicação Omnipage.

5. Anotação do texto corrigido aplicando etiquetas descritivas a segmentos específicos do texto. Essas etiquetas foram definidas provisoriamente numa base *ad-hoc* e destinavam-se a importar posteriormente o texto de forma automática para o Digitarq, i. e., para o DTD EAD definido. Esta anotação foi inicialmente realizada de forma manual através de um processador de texto simples e posteriormente através das aplicações **Xmetal** e **Authentic** de forma semi-automática (o operador seleccionava o texto e o tipo de etiqueta e a ferramenta aplicava-a).

6. Por vezes passavam erros na etapa 4 que eram posteriormente despistados na fase de anotação, o que obrigava a despender mais tempo nessa acção.

7. Para cada uma das fases anteriores o operador assentava numa grelha concebida para o efeito o tempo dispendido em cada etapa, discriminando esse tempo por registo e por página.

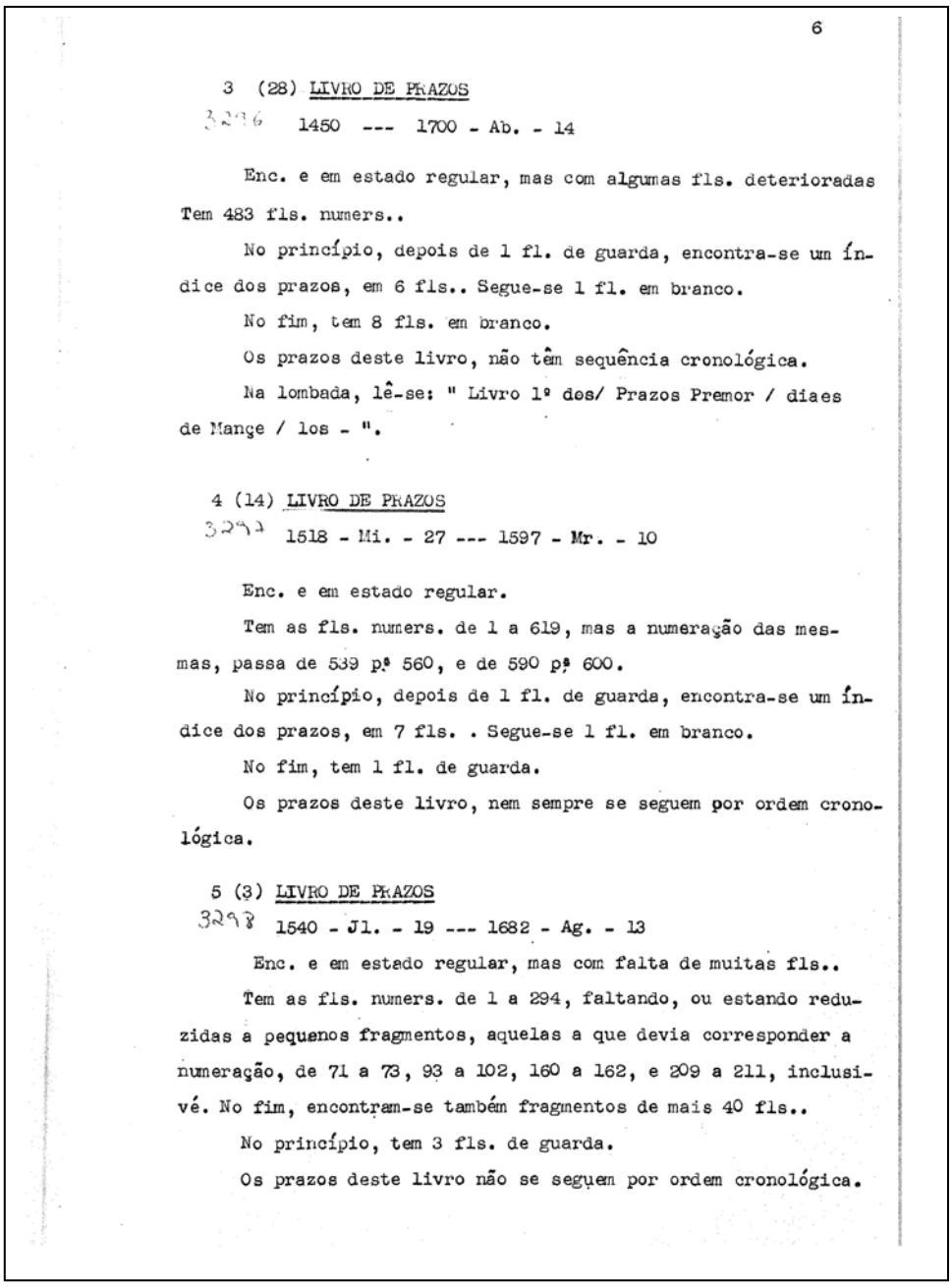
8. Foram afectos a este trabalho dois operadores a tempo inteiro, i. e., 7 horas de trabalho diário.

9. Para cada uma destas estruturas identificadas foi definido um conjunto de procedimentos e de etiquetas de anotação do qual daremos a seguir exemplos.

2. Estruturas de páginas

MODELO A

1- IMAGEM



6

3 (28) LIVRO DE PRAZOS

3296 1450 --- 1700 - Ab. - 14

Enc. e em estado regular, mas com algumas fls. deterioradas Tem 483 fls. numers..

No princípio, depois de 1 fl. de guarda, encontra-se um índice dos prazos, em 6 fls.. Segue-se 1 fl. em branco.

No fim, tem 8 fls. em branco.

Os prazos deste livro, não têm sequência cronológica.

Na lombada, lê-se: " Livro 1º dos/ Prazos Premor / diaes de Mange / los - ".

4 (14) LIVRO DE PRAZOS

3297 1518 - Mi. - 27 --- 1597 - Mr. - 10

Enc. e em estado regular.

Tem as fls. numers. de 1 a 619, mas a numeração das mesmas, passa de 539 pª 560, e de 590 pª 600.

No princípio, depois de 1 fl. de guarda, encontra-se um índice dos prazos, em 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 cronológica.

5 (3) LIVRO DE PRAZOS

3298 1540 - Jl. - 19 --- 1682 - Ag. - 13

Enc. e em estado regular, mas com falta de muitas fls..

Tem as fls. numers. de 1 a 294, faltando, ou estando reduzidas a pequenos fragmentos, aquelas a que devia corresponder a numeração, de 71 a 73, 93 a 102, 160 a 162, e 209 a 211, inclusivé. No fim, encontram-se também fragmentos de mais 40 fls..

No princípio, tem 3 fls. de guarda.

Os prazos deste livro não se seguem por ordem cronológica.

2- DESCRIÇÃO

Neste tipo de documentos, a informação apresentada, embora possua alguma forma/estrutura, não é consistente, ou seja, no corpo de cada registo, a informação presente pode variar bastante.

3- PROCEDIMENTO

Correcção de erros e anotação manual em XML.

4- TEXTO DEPOIS DE SUBMETIDO A OCR SEM CORRECÇÃO

3 (28) LIVRO DE PRAZOS
1450 --- 1700 - Ab. - 14
Enc. e em estado regular, mas com algumas fls. deterioradas Tem 483 fls. nuners..
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.
Da lombada, lé*-se: Livro 19 dos/ Prazos Prernor / dixes de Mane / los -

4 (14) LIVRO DE PRAZOS
1518 - 1,1i. - 27 --- 1597 - Mr. - 10
Enc. e em estado regular.
Tem as Vs, numers, de 1 a 619, mas a numeração das mesmas, passa de 539 p? 560, e de 590 pj 600.
>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 f1. de guarda.
Os prazos deste livro, nem sempre se seguem por ordem cronologica.

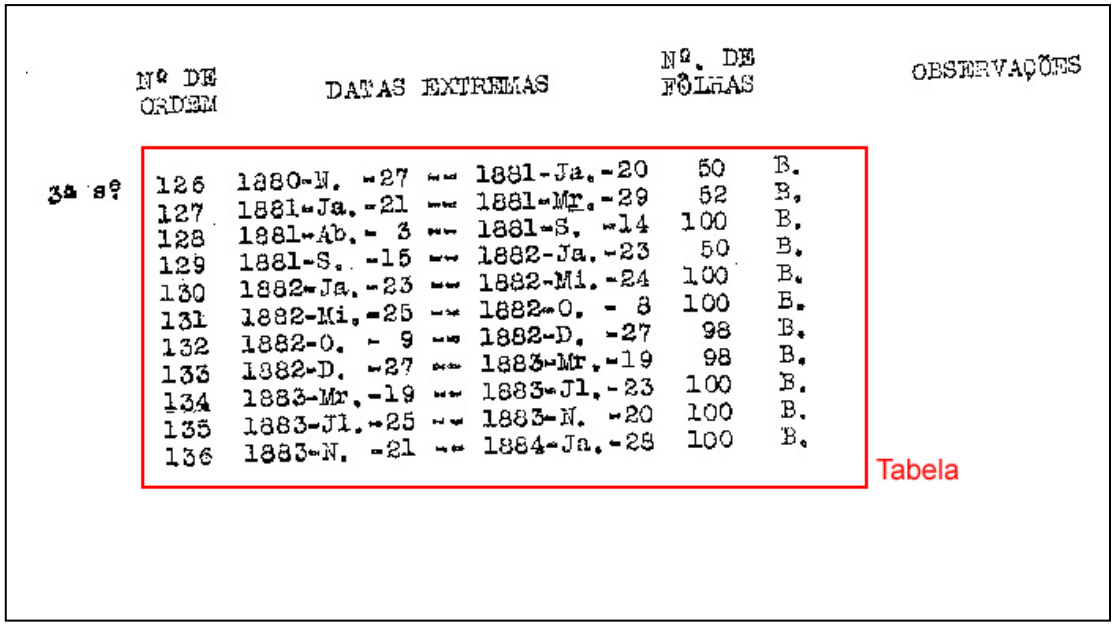
5 (3) LIVRO DE PRAZOS
,351540-J1.-19---1682-Ag. 0"- 13
Enc. e em estado regular, mas com falta de muitas fls..
Tem as fis. numers. de 1 a 294, faltando, ou estando reduzida s a pequenos

5. TEXTO ANOTADO

```
<registo>
<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…
</registo>
```

MODELO C

1 IMAGEM



	Nº DE ORDEM	DATAS EXTREMAS			Nº. DE FÔLHAS	OBSERVAÇÕES
3ª s?	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.-28	100	B.

Tabela

2 DESCRIÇÃO

Texto estruturado e repetitivo com informação resumida a datas, dimensão e n.º ordem.
Há uma variante a esta estrutura denominada de C2 em que os meses são representados em caracteres numéricos e entre as datas aparece uma tablatura.

3 PROCEDIMENTO

Correcção ortográfica com anotação *ad-hoc.*

4 TEXTO DEPOIS DE SUBMETIDO A OCR SEM CORRECÇÃO

```
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.
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.
```

5. TEXTO ANOTADO

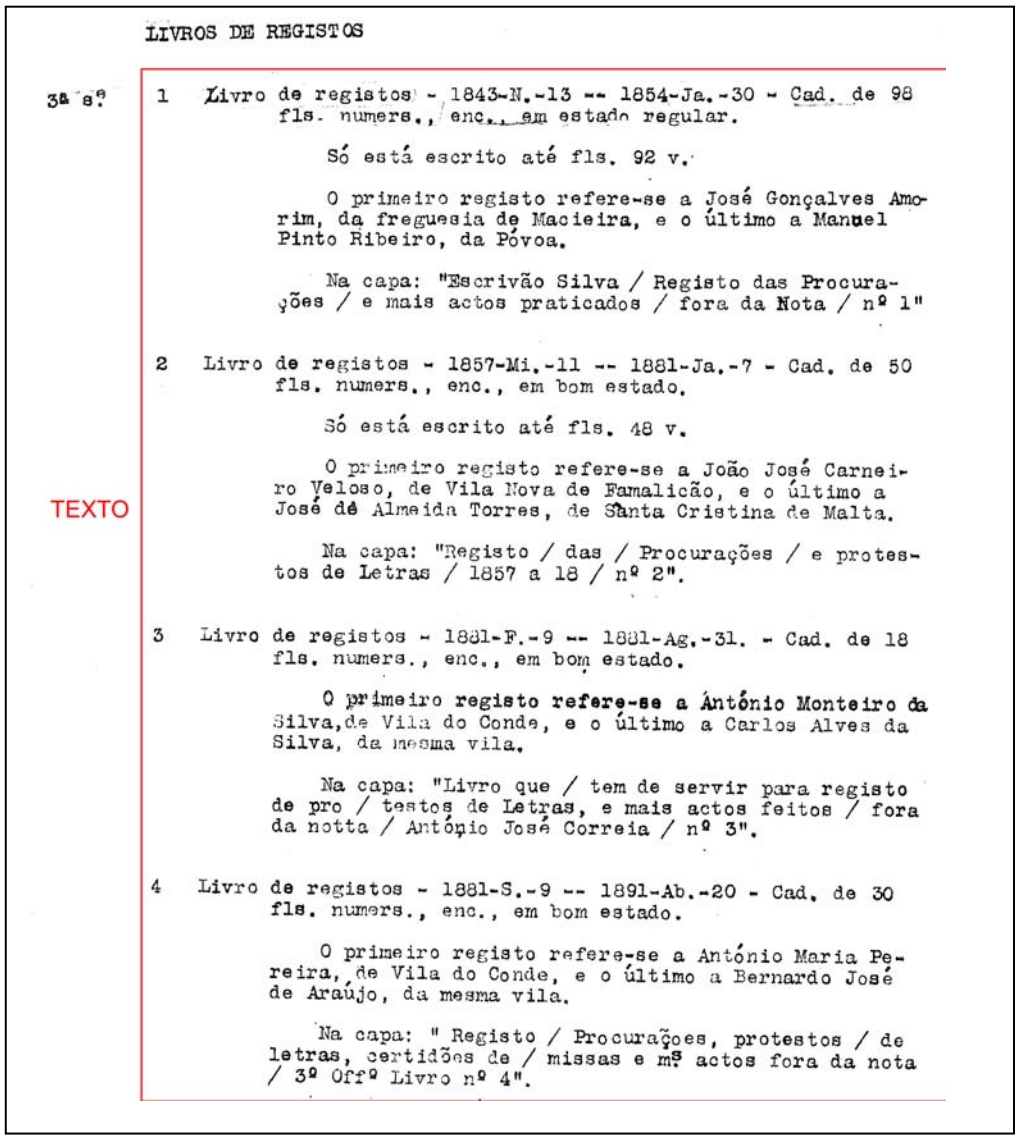
Etiquetas utilizadas:

Etiqueta	Descrição
MOD C	Início de descrições no modelo de documentos de tipo C.
F <TEXTO>	Início de fundo.
SR <NUM>	Abre uma nova série.
SSR <NUM>	Abre uma nova subsérie.
COTA <TEXTO>	Define a cota dos registos subsequentes.
COTAAUTO <TEXTO> <NUM>	Define a cota dos registos subsequentes fazendo crescer o valor de NUM e mantendo TEXTO fixo.
TI <TEXTO>	Define o título dos documentos subsequentes
SA <NUM>	Série antiga. Nova sequência numérica (define o número dos milhares no código das peças).

```
MODC

F CN-CNVCD02

SR 001

TI Livros de Notas

COTA 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.

COTAAUTO 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 E

1 IMAGEM



2 DESCRIÇÃO

Texto muito informativo com um padrão de disposição de informação sem variações significativas. Barras a separar linhas de título provocando ruído no reconhecimento.

3 PROCEDIMENTO

Correcção ortográfica com anotação *ad-hoc*.

4 TEXTO DEPOIS DE SUBMETIDO A OCR SEM CORRECÇÃO

```
1      livro de registos,' -,1843-N.-13 -- 1854-Ja, -30 w Cad., de 98 fie-
nimers„í enç„,_,,__ffiz estado regular,

Só esta. escrito até fls, 92 v,'

0 primeiro registo refere-sie a José Gonçalves Amorim, da freguesia de
Macieira, e o último a Mandei Pinto Ribeiro, da Póvoa.

Na capa: "Escrivão Silva / Registo das Procura~ees / e anais actos
praticados / fora da Nota / nQ 1"


2 Livro de registos - 1857-Mi,-11      1831-Ja,-7  -  Cad,  de  50  fie.
numera„ enc., em bom estado.

Só estó, escrito até fls. 48 v,

o pr'i!neiro registo refere-se a João José Carnei~ ro Veloso, de Vila
Nova de Famalicão, e o último a José d4 Almeida Torres, de Santa Cristina
de Malta.

Na capa: "Registo / das / Procurações / e protes
tos de Letras / 1857 a 18 / ng 2",
3      Livro de registos ae l881-F, 9 ~- 1831-Ag, -31, .. Cad, de 18 fie,
numera„, eno,, em bota estado,
0      imeiro registo refere-se a Ántánio Monteiro da, ~ilva,de Vila do
Conde, e o último a Carlos Alves da Silva, da )nesma vila,
Na capa: "Livro que / tem de servir para registo de pro / testos de
Letras, e mais actos feitos / fora da notta / , nto io José: Correia / nQ
3",


4      Livro de registos - 1881-S,-9 -- 1891-Ab,-20 - Cad, de 30
flsf   numera., enc,, em bom estado.
0 primeiro registo refere-se a António Maria Pe. re ira, (~e Vila do
Conde, e o último a Bernardo José de Araújo, da mesma vila.
Na capa: " Registo / Procuraçoes, protestos / de letras, certidões de /
missas e rns actos fora da nota / 30 0ff Q Livro ng 4".
```

5. TEXTO ANOTADO

Etiquetas utilizadas:

Etiqueta	Descrição
MOD E	Início de descrições no modelo de documentos de tipo E.
#	Separador de registo.
F <TEXTO>	Início de fundo.
SR <NUM>	Abre uma nova série.
SSR <NUM>	Abre uma nova subsérie.
COTA <TEXTO>	Define a cota dos registos subsequentes.
COTAAUTO <TEXTO> <NUM>	Define a cota dos registos subsequentes fazendo crescer o valor de NUM e mantendo TEXTO fixo.
ID	Identificação do registo.
TI <TEXTO>	Define o título do documento.
CARF <TEXTO>	Características físicas.
OBS <TEXTO>	Nota ou observações.
AMBCONT <TEXTO>	Âmbito e conteúdo.
TIPROP <TEXTO>	Título próprio.
TUI <TEXTO>	Tipo de unidade de instalação.
DIM <TEXTO>	Dimensões físicas.
UIPAI <TEXTO>	Unidade de instalação que contém o documento: <localizacao>$cota $uipai</localizacao>.
DATASPRED <TEXTO>	Datas predominantes.
SA <NUM>	Série antiga. Nova sequência numérica (define o número dos milhares no código das peças).

```
MODE
F NOT-CNVCD02
SR 010
COTA I/23/23 CX 12

ID 1
TI livro de registos
DATAS 1843-N.-13 -- 1854-Ja-30
CARF Estado regular,
DIM 98 fls numeradas
TUI Caderno
OBS Só está escrito até fls, 92 v.
AMBCONT 0 primeiro registo refere-sie a José Gonçalves Amorim, da
freguesia de Macieira, e o último a Mandei Pinto Ribeiro, da Póvoa.
TIPROP Na capa: "Escrivão Silva / Registo das Procurações / e anais
actos praticados / fora da Nota / nQ 1"
#
ID 2
TI Livro de registos
DATAS 1857-Mi,-11 -- 1831-Ja,-7
TUI Cad
DIM 50 fls. Numeradas
CARF Enc., em bom estado.
OBS Só estó, escrito até fls. 48 v,
AMBCONT o pr'i!neiro registo refere-se a João José Carnei~ ro
Veloso, de Vila Nova de Famalicão, e o último a José d4 Almeida
Torres, de Santa Cristina de Malta.
TIPROP Na capa: "Registo / das / Procurações / e protes
tos de Letras / 1857 a 18 / ng 2",
#
```

3. Estatísticas

1. Os colaboradores preenchiam uma grelha com a estrutura abaixo representada sempre que procediam a tarefas no OCR, revisão/correcção e anotação, de forma a contabilizar de forma precisa o tempo dispendido em cada uma das actividades. Assim tornou-se possível recolher dados discriminados sobre o processo na sua totalidade.

Fundo	Recon. / Anotação	Data	Hora inicial	Hora final	Funcionário

Posteriormente estes dados eram refinados e lançados noutra grelha abaixo representada onde os resultados eram discriminados de acordo com a metodologia de anotação utilizada indicando o total de registos e de folhas tratadas.

FUNDOS	R	A (manual)	A (correcção de texto)	A (XMetaL)	A (Authentic)	TOTAL	N.º de fls.	N.º de registos

2. A fórmula de contabilização prevista era a seguinte:

$$t_{conversão} = t_{digitalização} + t_{correcção} + t_{anotação}$$

em que t=tempo

Não foi no entanto utilizada a variável *tDigitalização* pelo facto que este processo se ter realizado muito antes do início do processo de OCR. Assim sendo esta variável foi transformada em: $t_{conversão} = t_{ocr\_correcção} + t_{anotação}$

A correcção foi adjunta ao OCR visto que era realizada na mesma acção, imediatamente após o OCR e, ainda, através da mesma ferramenta informática utilizada (o OMNIPAGE 9.0).

3. Os resultados são representados na seguinte tabela:

# ref. , de ID de núcleos documentais	tempo (minutos)	# registos	# folhas	tempo/registo	tempo/folha
1	5070	1828	234	2,77	21,67
2	5490	991	226	5,54	24,29
3	390	220	17	1,77	22,94
4	1010	308	53	3,28	19,06
5	230	24	6	9,58	38,33
6	2640	185	80	14,27	33,00
7	1800	132	56	13,64	32,14
8	1420	126	55	11,27	25,82
9	1030	102	40	10,10	25,75
10	2460	507	113	4,85	21,77
11	910	99	35	9,19	26,00
12	4070	764	150	5,33	27,13
13	690	116	46	5,95	15,00
14	1800	1861	97	0,97	18,56
15	1050	710	133	1,48	7,89
16	1150	1078	116	1,07	9,91
17	2130	1321	124	1,61	17,18
18	310	686	80	0,45	3,88
19	1680	1420	60	1,18	28,00
20	1430	1171	27	1,22	52,96
21	1100	322	13	3,42	84,62
22	885	122	9	7,25	98,33
23	240	611	54	0,39	4,44
24	120	631	81	0,19	1,48
25	930	1140	52	0,82	17,88
26	840	285	36	2,95	23,33
27	886	1382	25	0,64	35,44
28	835	691	24	1,21	34,79
29	350	1668	107	0,21	3,27
30	174	748	38	0,23	4,58
31	160	183	9	0,87	17,78
32	873	580	21	1,51	41,57
33	205	395	34	0,52	6,03
34	605	514	58	1,18	10,43
35	740	411	41	1,80	18,05
totais	**45.703**	**23.332**	**2.350**	**128,71**	**873,32**

- **Média de tempo dispendido em minutos por folha ➔24,95**
- **Média de tempo dispendido em minutos por registo ➔3,68**

4. Apresentam-se, finalmente, discriminadamente os tempos contabilizados para as actividades de OCR/correcção e de anotação. O somatório destes dois tempos retorna o tempo total dispendido representado na coluna "tempo" da tabela anterior. Para maior facilidade de leitura indicam-se as percentagens de tempo relativas a cada actividade.
Apenas foram contabilizados dados relativos a 26 núcleos documentais respeitantes à segunda fase do processo.

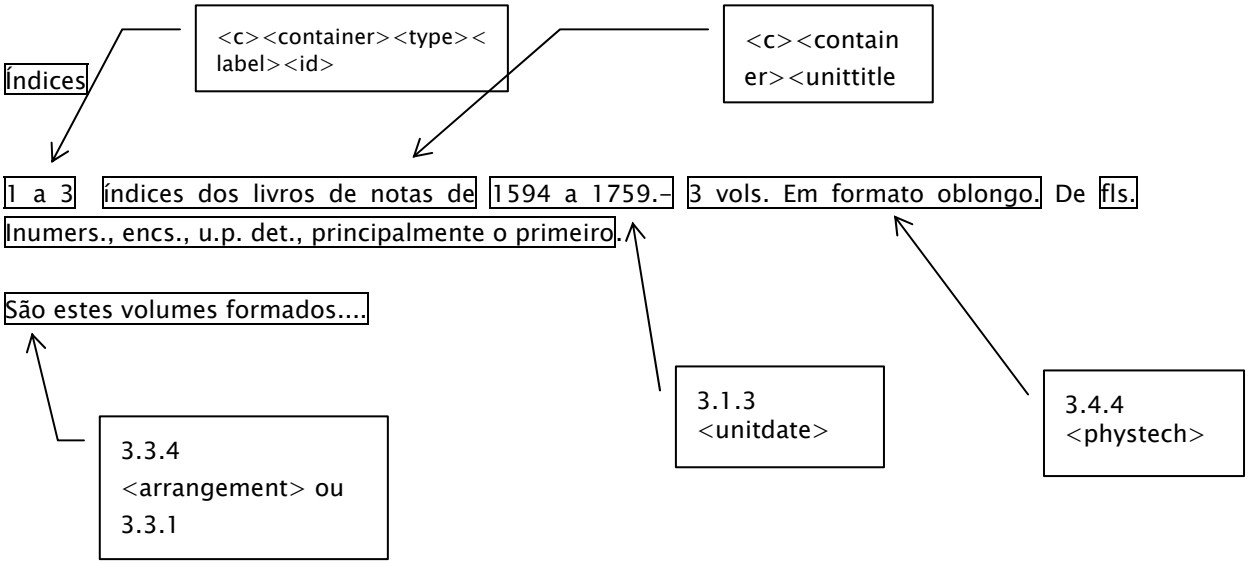
	OCR/Correcção	%	Anotação	%	*Tempo total*
1	1650	32,54	3420	67,46	*5070*
2	3340	60,84	2150	39,16	*5490*
3	120	30,77	270	69,23	*390*
4	360	35,64	650	64,36	*1010*
5	50	21,74	180	78,26	*230*
6	1080	40,91	1560	59,09	*2640*
7	990	55,00	810	45,00	*1800*
8	820	57,75	600	42,25	*1420*
9	450	43,69	580	56,31	*1030*
10	1320	53,66	1140	46,34	*2460*
11	400	43,96	510	56,04	*910*
12	2590	63,64	1480	36,36	*4070*
13	330	47,83	360	52,17	*690*
14	470	26,11	1330	73,89	*1800*
15	420	40,00	630	60,00	*1050*
16	200	17,39	950	82,61	*1150*
17	1020	47,89	1110	52,11	*2130*
18	100	32,26	210	67,74	*310*
19	1020	60,71	660	39,29	*1680*
20	1050	73,43	380	26,57	*1430*
21	420	38,18	680	61,82	*1100*
22	265	29,94	620	70,06	*885*
23	90	37,50	150	62,50	*240*
24	70	58,33	50	41,67	*120*
25	390	41,94	540	58,06	*930*
26	420	50,00	420	50,00	*840*

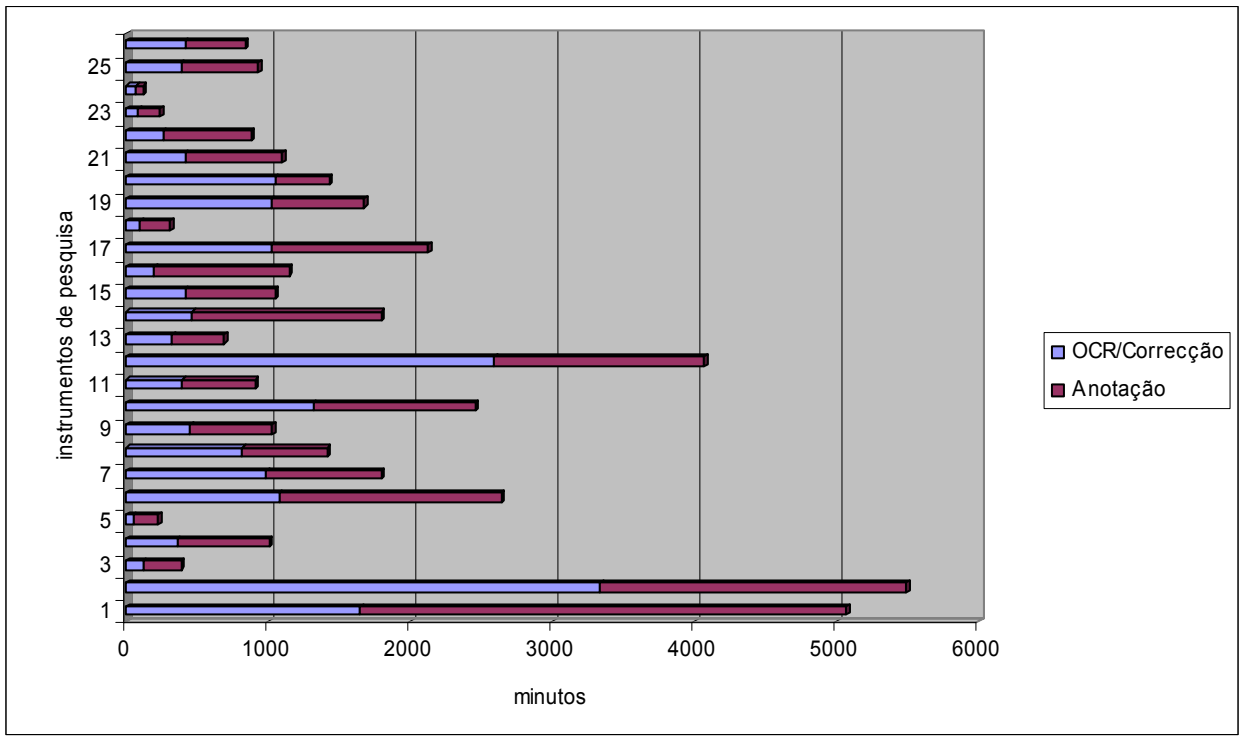
5. Sintetizando os resultados verifica-se que em

65,38	17	mais tempo de anotação
30,77	8	mais tempo de correcção
3,85	1	igual

Em 65% dos casos o tempo de anotação foi superior ao de correcção. Em 30% dos casos passou-se o inverso e num caso apenas os tempos foram idênticos.
Conclui-se assim que a anotação era o processo que mais tempo consumiu, o que é explicável pelo facto de esta ser aplicada de forma muito detalhada. Por exemplo, em conjuntos de texto onde residia informação repetível por vários campos descritivos existentes nas normas de descrição utilizadas, esta era isolada nos diferentes segmentos textuais que assim eram anotados com as marcações correspondentes. Ao contrário das recomendações exaradas no *EAD Cookbook* da SAA (Society of American Archivists) que apontam no sentido de que sempre que haja uma situação de aglomerados textuais repartível por diversos campos de descrição, se deverá utilizar a etiqueta <odd> (other descriptive data), optámos por efectivamente realizar essa repartição de forma a tornar as descrições resultantes mais precisas e coerentes.

Exemplo de anotação:





Tempo dispendido na anotação e correcção.