# ARQUIVO DISTRITAL DO PORTO

# PROJECTO DIGITARQ

# OCR STATISTICS

2004-05-31

# 1 Workflow description

1. Identification of different pages layouts. Three layouts were identified: A, C, E.

2. Digitisation of the finding aids and OCR aplication with Omnipage Pro 12.0. The pages were segmented by columns when aplicable .

3. Revision and correction of the text still with Omnipage Pro 12.0

5. Anotation of the text already corrected apllying descriptive tags to specific segments of the text. Those tags –xml based- were intended to allow subsequent automatic importation process to EAD structure. This anotation was performed with Xmetal and Authentic tools.

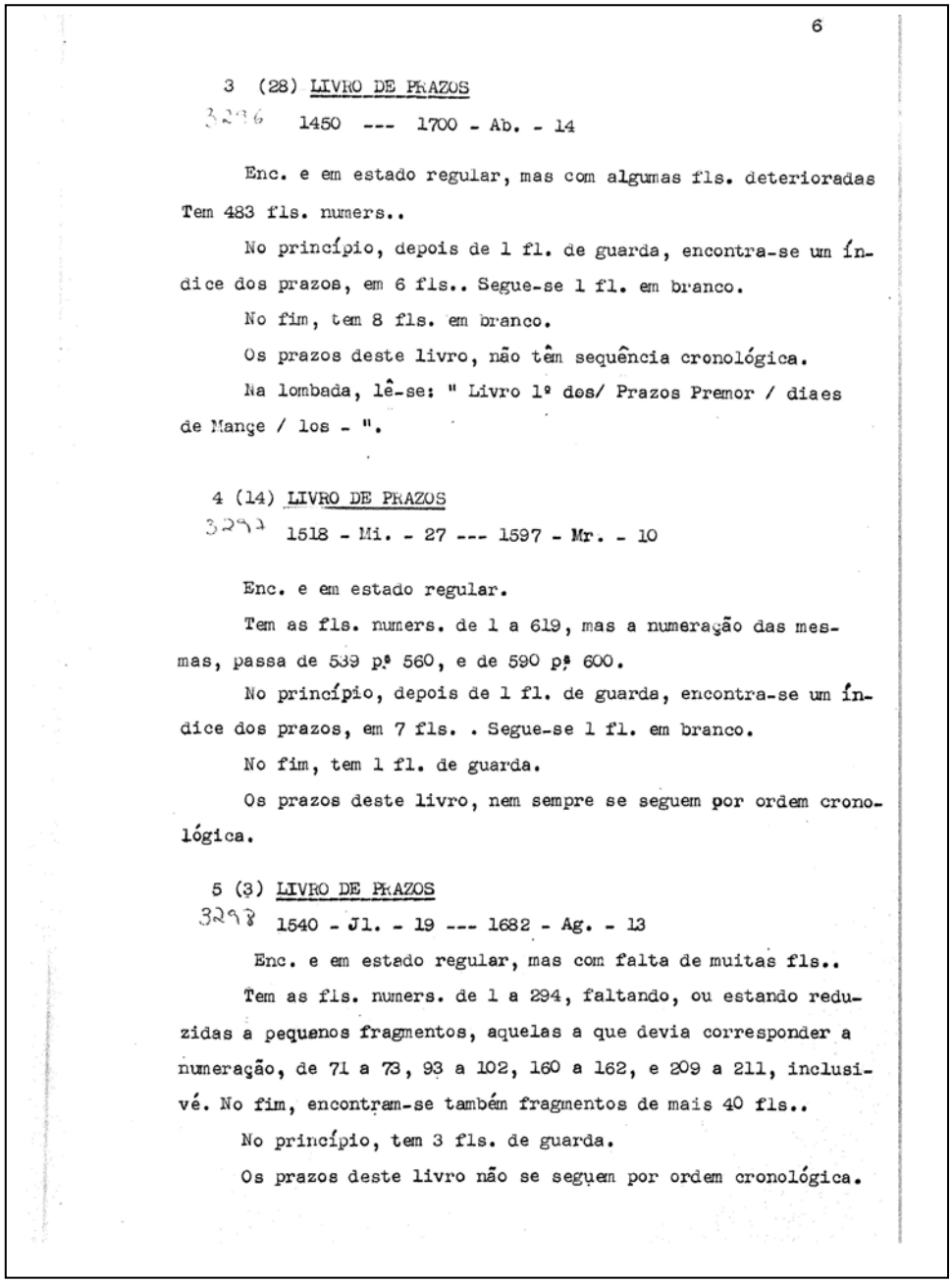
6. The time spent in each of the previous steps was anotated by the operator in a grid specially constructed for the purpose. The time was discriminated by record and by page.

Two operators with archival background were assigned full-time to this task (7 hour work/day).

# 2 Page Layouts

## *MODEL A*

1 IMAGE





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..

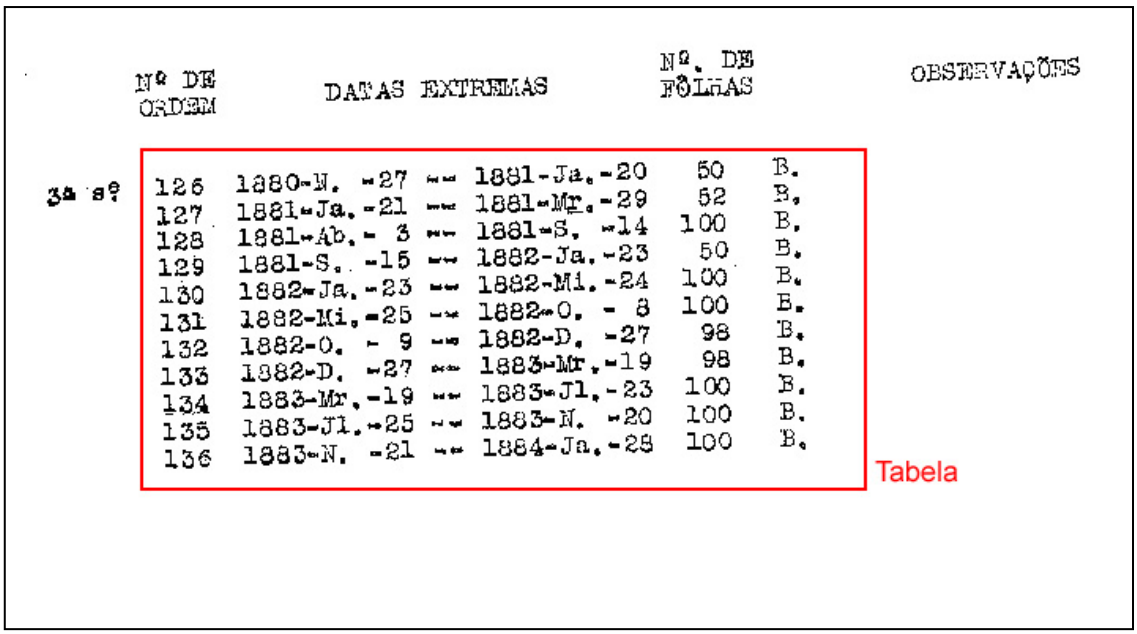
Tem as fls. numers. de 1 a 294, faltando, ou estando reduzidas a pequenos fragmentos, aquelas a que devia corresponder a numeração, de 71 a 73, 93 a 102, 160 a 162, e 209 a 211, inclusivé. No fim, encontram-se também fragmentos de mais 40 fls..

No princípio, tem 3 fls. de guarda.

Os prazos deste livro não se seguem por ordem cronológica.

## *MODEL C*

1 IMAGE

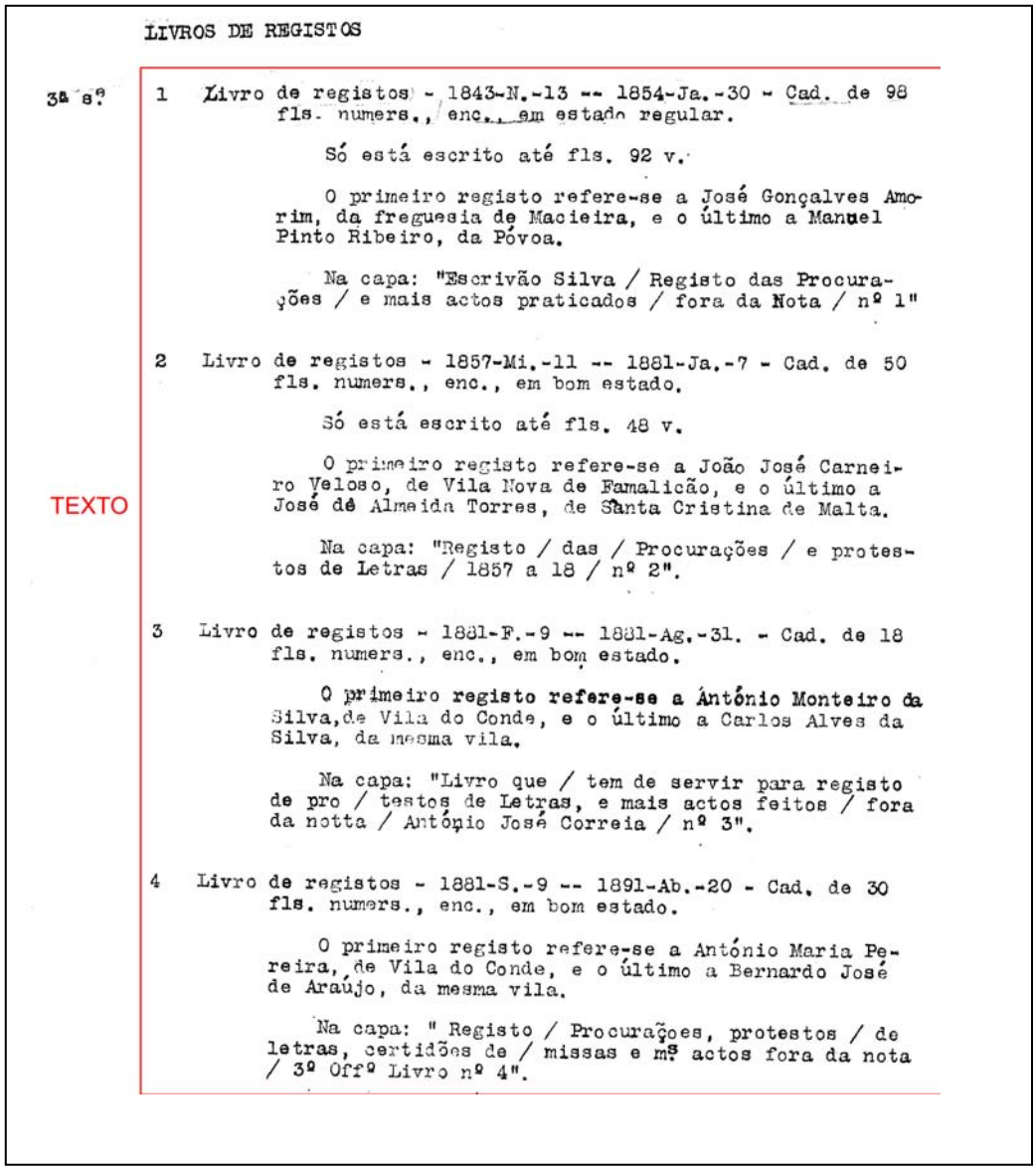


| | Nº DE ORDEM | DATAS EXTREMAS | | | Nº. DE FÔLHAS | OBSERVAÇÕES |
|---|---|---|---|---|---|---|
| 3ª s? | 126 | 1880-N. -27 | -- | 1881-Ja.-20 | 50 | B. |
| | 127 | 1881-Ja.-21 | -- | 1881-Mr.-29 | 52 | B. |
| | 128 | 1881-Ab.- 3 | -- | 1881-S. -14 | 100 | B. |
| | 129 | 1881-S. -15 | -- | 1882-Ja.-23 | 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

## *MODEL E*

1 IMAGE



LIVROS DE REGISTOS

3ª s?

TEXTO

1 Livro de registos - 1843-N.-13 -- 1854-Ja.-30 - Cad. de 98 fls. numers., 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 Pinto Ribeiro, da Póvoa.

Na capa: "Escrivão Silva / Registo das Procurações / e mais actos praticados / fora da Nota / nº 1"

2 Livro de registos - 1857-Mi.-11 -- 1881-Ja.-7 - Cad. de 50 fls. numers., enc., em bom estado.

Só está escrito até fls. 48 v.

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.

Na capa: "Registo / das / Procurações / e protestos de Letras / 1857 a 18 / nº 2".

3 Livro de registos - 1881-F.-9 -- 1881-Ag.-31. - Cad. de 18 fls. numers., enc., em bom estado.

O primeiro registo refere-se a António Monteiro da Silva, de Vila do Conde, e o último a Carlos Alves da Silva, da mesma vila.

Na capa: "Livro que / tem de servir para registo de pro / testos de Letras, e mais actos feitos / fora da notta / António José Correia / nº 3".

4 Livro de registos - 1881-S.-9 -- 1891-Ab.-20 - Cad. de 30 fls. numers., enc., em bom estado.

O primeiro registo refere-se a António Maria Pereira, de Vila do Conde, e o último a Bernardo José de Araújo, da mesma vila.

Na capa: " Registo / Procurações, protestos / de letras, certidões de / missas e mª actos fora da nota / 3º Offº Livro nº 4".

# 3 Statistics

1. The following grid was completed by the colaborators as the work was progressing.

| Fond | OCR/Anotation | Date | Time begin | Time End | Employee |
|---|---|---|---|---|---|
| | | | | | |
| | | | | | |

2. The formula used was the following:

$$t_{conversion} = t_{ocr\_correction} + t_{anotation}$$

t=time

3. The universe of aplication consisted of a set of 35 finding aids reported to 35 fonds (records groups) containing different levels of description

4. The final results are presented in the following table:

| # ref. Finding aids | Time (minuts) | # records | # pages | Time/record (minutes) | Time/page (minutes) |
|---|---|---|---|---|---|
| 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 |
| **totals** | **45.703** | **23.332** | **2.350** | **128,71** | **873,32** |

- **Average time spent (minutes) by page ➔24,95**
- **Average time spent (minutes) by record ➔3,68**

5. The following table presents the results discriminated by OCR/correction and Anotation spent time. The % are also represented in the columns. The universe of aplication was reduced in this case to 26 finding aids representing 26 fonds (records groups).

| | OCR/Correction (minutes) | % | Anotation (minutes) | % | *Total time (minutes)* |
|---|---|---|---|---|---|
| 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* |

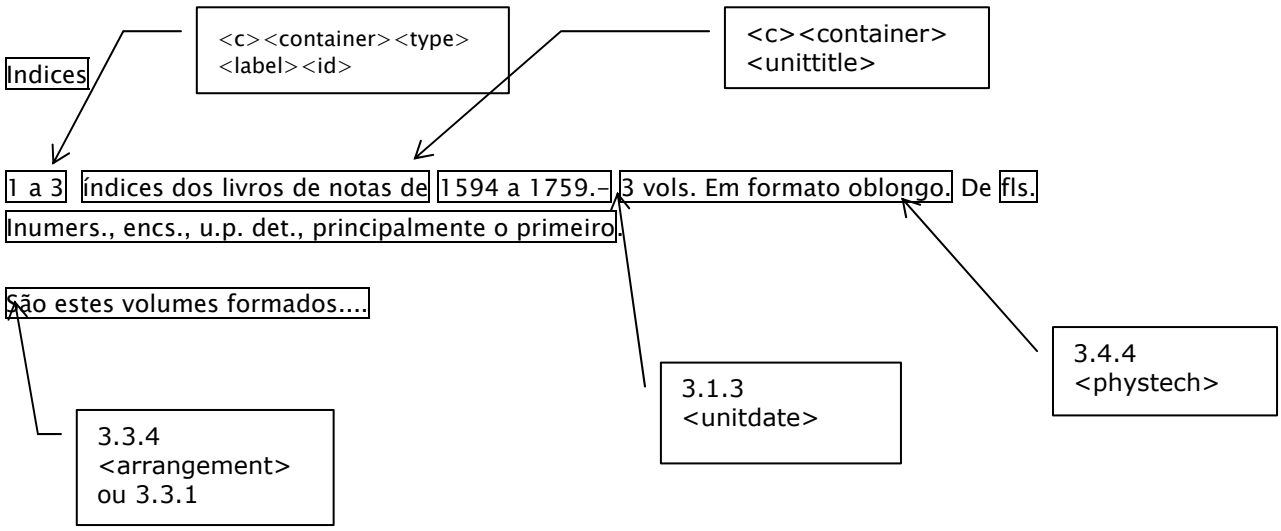
| | | | | | |
|---|---|---|---|---|---|
| 14 | 470 | 26,11 | 1330 | 73,89 | *1800* |
| 15 | 420 | 40,00 | 630 | 60,00 | *1050* |
| 16 | 200 | 17,39 | 950 | 82,61 | *1150* |
| 17 | 1020 | 47,89 | 1110 | 52,11 | *2130* |
| 18 | 100 | 32,26 | 210 | 67,74 | *310* |
| 19 | 1020 | 60,71 | 660 | 39,29 | *1680* |
| 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. We can see that in 17 cases more time was spent in anotation. In 8 cases the opposite happenned as more time was spent in correction time. In one case the time spent was equal.

| | | |
|---|---|---|
| 65,38 | 17 | > anotation time |
| 30,77 | 8 | < correction time |
| 3,85 | 1 | = correction time |

The process of anotation was the more time consuming. This can be explained by the fact that is was a very detailed process in which blocks of text that might be included in a specific EAD tag, were isolated and marked with that specific tag. The tag <odd> (other descriptive data) was never used.

Anotation example



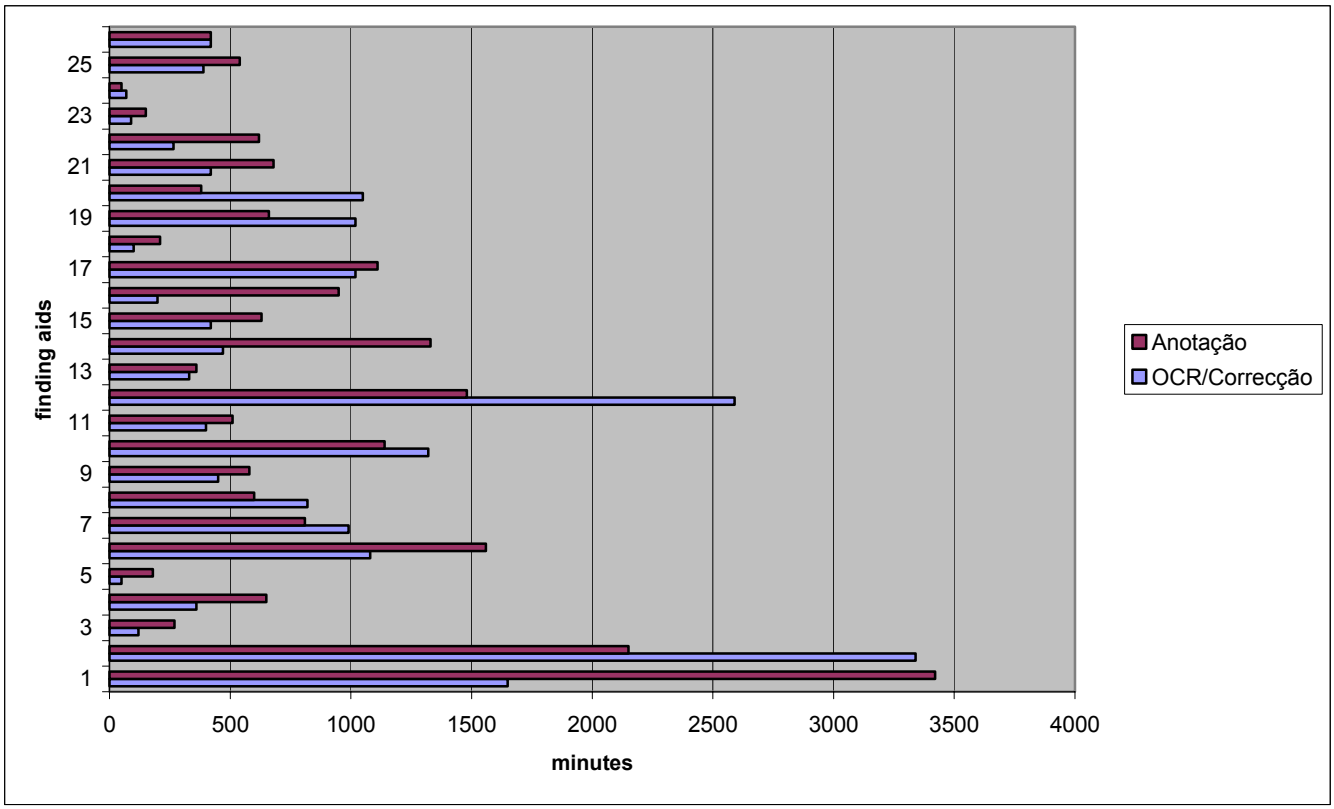



*Figure 1 – Time Spent in OCR/Correction and Anotation*