Sistema de
Edições Eletrônicas
do Corpus Histórico do Português Tycho Brahe
Fundamentos, Diretrizes e Procedimentos

Memórias do Texto
[http://www.ime.usp.br/~tycho/participants/psousa/memorias]

Maria Clara Paixão de Sousa

Setembro, 2007

Versão para impressão do Manual em
http://www.ime.usp.br/~tycho/corpus/manual/prep/index.html

Índice Completo

0.1 Visão geral..... 5

0.2 O Processamento Eletrônico de Textos.....9
0.2.1 Conceitos Introdutórios.....9
0.2.2 A codificação eletrônica da informação..... 10
0.1.3 Breve tipologia dos textos eletrônicos..... 13

0.3 Controle do Processamento Eletrônico..... 15
0.3.1 Motivações para o processamento controlado.....15
0.3.2 O Hipertexto como forma de processamento controlado.....16
0.3.3 A linguagem de anotação extendida, ou XML..... 18

0.4 Etapas do Trabalho de Edição Eletrônica..... 19
0.3.1 Procedimentos de Edição – Visão Geral..... 19
1. Edições Conservadoras, ou Diplomáticas:..... 19
2. Edições Semi-interpretativas, ou Semi-Diplomáticas:..... 19
3. Edições Interpretativas, ou Modernizadas:..... 20
0.3.2 Procedimentos da Edição Eletrônica..... 20
A. Rescensão – Pesquisa e Catalogação..... 20
B. Cópia – Transcrição ou Transposição eletrônica 20
C. Preparação e Estabelecimento do Texto – Edição Interpretativa 21
D. Apresentação.....21
Notas da Seção.....23

I.1 Diretrizes.....24

I.2 Procedimentos..... 25

I.2.1 Arquivo Básico de Catalogação.....25

I.2.2 Anotação dos metadados..... 26
I.2.2.1 Esquema da anotação de cabeçalhos.....27
B. Guia das Nomenclaturas..... 28
I.2.2.2 Amostra: Cabeçalho e Ficha catalográfica de um texto do Corpus..... 30
I.2.3 Formação do Catálogo 32
I.2.3.1 Ativação do Catálogo – via local.....33
I.2.3.2 Ativação do Catálogo – via remota..... 33
Atualização.....33

I.3 Documentação (Catalogação Interna).....34

I.3.1 Estruturas de Diretório.....34

I.3.2 Tabela de nomeação dos arquivos.....35
I.3.3 Documentação de Anotação..... 36
I.3.3.1 Gramática da anotação XML no Corpus (DTD).....36

II.1 Diretrizes..... 37
II.1.1Visão Geral.....37
II.1.2 Normas para a transcrição de originais..... 37
II.1.2.1 Caracteres alfabéticos:..... 37
II.1.2.2 Abreviaturas, pontuação e organização espacial do texto:..... 37
II.1.2.3 Indicações de Dificuldades de Leitura e Aspecto do suporte material 37

II.2 Procedimentos.....38
II.2.1 Digitação.....38
II.2.2 Anotações de Estruturas de Texto (gross structure) 39
II.2.2.1 Anotação de Formatação.....39
II.2.2.2 Anotação de Divisões do Texto 39
II.2.2.3 Anotação de Elementos do Texto..... 39
II.2.2.4 Anotação de Dificuldades de Leitura e outros Comentários.....39

II.2.3 Exemplo: Transcrição de um original impresso 40
II.2.3.1 O original impresso (fac-simile) 40
II.2.3.2 A Transcrição 40

III.1 Diretrizes 41
III.1.1.Visão Geral 41
III.1.2. Normas de Edição 42
III.1.2.1 Caracteres alfabéticos 42
III.1.2.2 Abreviaturas, pontuação e organização espacial do texto 42
III.1.2.3 Grafia 43
III.1.2.4 Léxico e Morfosintaxe 43
III.1.2.5 Outras Interferências 43

III.2 Procedimentos 44
III.2.1 Anotação das Interferências de Edição 44
II.2.1.1 Anotação geral 45
II.2.1.2 Atributos para os tipos de interferências 45
III.2.2 Exemplo: edição interpretativa 46
III.2.2.1 Anotação das Interferências Editoriais no Texto-Base 46
II.2.2.2 Apresentações possíveis a partir da anotação de edição 46

III.3 Diretrizes para a transposição digital de outras edições 47
III.3.1 Exemplo: Transposição Digital de uma edição semi-diplomática 48
A. O Manuscrito Original (fac-simile) 48
..... 48
B. A Edição Semi-Diplomática 49
C. A Transposição Digital 49
III.3.2 Normas para a Transposição Digital de Outras Edições 51
III.3.1.1 Adaptação Facilitada no processador Open Office 51
Ferramentas para esse procedimento: 52

IV.1 Fundamentos 53
IV.1.1. Visão Geral 53
IV.1.2. Exemplo: apresentação de um texto editado eletronicamente 53
IV.1.3. O Hipertexto Crítico 56

IV.2 Procedimentos 56

IV.2.1 Procedimentos de Ativação 57
IV.2.1.1 Ativação das transformações – via local 58
IV.2.1.2 Ativação das transformações – via remota 58
IV.2.1.3 Associação Direta nos Documentos 59

IV.2.2 Programações de Transformação 60

0. Introdução

O sistema de edição utilizado no *Corpus Histórico do Português Tycho Brahe* foi desenvolvido de modo a trazer para o âmbito do trabalho de edição filológica alguns avanços no processamento eletrônico de textos. Entende-se aqui por *edição filológica* a edição especializada que procura trazer a público o texto de difícil lição – tipicamente, o texto antigo.

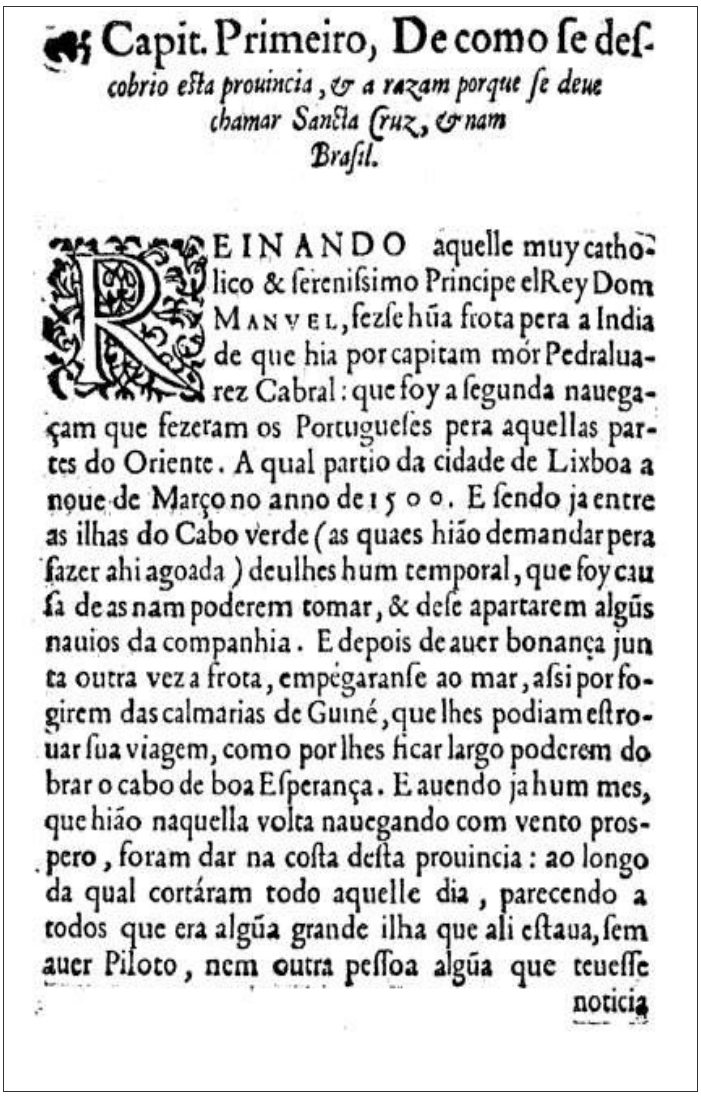
Focaliza-se a edição filológica dirigida a um público especializado: o lingüista, ou estudioso da história da língua. Neste tipo de edição, a finalidade da pesquisa lingüística confere à edição a responsabilidade pela máxima fidelidade em relação aos textos originais. A isso deve-se conjugar um segundo objetivo: a preparação dos textos para análise automática por ferramentas de análise lingüística (etiquetador morfológico e analisador sintático).

O sistema de edição aqui apresentado se propõe a atender a essas duas demandas (qualidade filológica e adequação para as ferramentas). Os fundamentos e motivações desse trabalho são expostos nesta Introdução.

Os procedimentos no processo de edição eletrônica refletem fundamentalmente os procedimentos do trabalho de edição em geral; entrentanto, o processamento eletrônico confere singularidades técnicas a algumas etapas do trabalho. As as diretrizes e os procedimentos segidos no sistema de edição são descritos nas seções I a IV:

- a catalogação dos textos (I)
- a transcrição dos textos (II);
- a codificação da interferência editorial sobre os textos (III),
- a apresentação dos textos (IV)

0.1 Visão geral



Capit. Primeiro, De como ſe deſcobrio eſta prouincia, & a razam porque ſe deue chamar Sancta Cruz, & nam Braſil.

REINANDO aquelle muy catholico & ſereniſsimo Principe elRey Dom MANVEL, fezſe hũa frota pera a India de que hia por capitam mór Pedraluarez Cabral: que foy a ſegunda nauegaçam que fezeram os Portugueſes pera aquellas partes do Oriente. A qual partio da cidade de Lixboa a noue de Março no anno de 1500. E ſendo ja entre as ilhas do Cabo verde (as quaes hião demandar pera fazer ahi agoada) deulhes hum temporal, que foy cauſa de as nam poderem tomar, & deſe apartarem algũs nauios da companhia. E depois de auer bonança junta outra vez a frota, empẽgaranſe ao mar, aſsi por fogirem das calmarias de Guiné, que lhes podiam eſtrouar ſua viagem, como por lhes ficar largo poderem dobrar o cabo de boa Eſperança. E auendo ja hum mes, que hião naquella volta nauegando com vento prospero, foram dar na coſta deſta prouincia: ao longo da qual cortáram todo aquelle dia, parecendo a todos que era algũa grande ilha que ali eſtaua, ſem auer Piloto, nem outra peſſoa algũa que teueſſe noticia

A forma mais fiel de se reproduzir um texto antigo no meio digital é sem dúvida o facsimile. Entretanto, para pesquisas lingüísticas é necessário trabalhar o texto como seqüências de caracteres (não como imagens).

No trabalho de preparação de textos para o Corpus Tycho Brahe, a saída de curto prazo escolhida foi a transcrição dos originais, enquanto se pesquisam formas de adequação do reconhecimento automático (tanto via OCRs aprimorados como via sistemas de correção posterior)

A saída de curto prazo foi a transcrição dos originais, enquanto se pesquisam formas de adequação do reconhecimento automático (tanto via OCRs aprimorados como via sistemas de correção posterior).

Capit. Primeiro, De como ſe deſ-
cobrio esta prouincia, & a razam porque ſe deue
chamar Sancta Cruz, e nam
Braſil.

REINANDO aquelle muy catho-
lico & ſereniſsimo Principe elRey Dom
MANVEL, fezſe hũa frota pera a India
de que hia por capitam mór Pedralua-
rez Cabral: que foy a ſegunda nauega-
çam que fezeram os Portugueſes pera aquellas par-
tes do Oriente. A qual partio da cidade de Lixboa a
noue de Março no anno de 1500. E ſendo ja entre
as ilhas do Cabo verde (as quaes hião demandar pera
fazer ahi agoada) deulhes hum temporal, que foy cau
ſa de as nam poderem tomar, & deſe apartarem algũs
nauios da companhia. E depois de auer bonança jun
ta outra vez a frota, empegaranſe ao mar, aſsi por fo-
girem das calmarias de Guiné, que lhes podiam eſtro-
uar sua viagem, como por lhes ficar largo poderem do
brar o cabo de boa Eſperança. E auendo jahum mes,
quehião naquella volta nauegando com vento pros-
pero, foram dar na coſta deſta prouincia: ao longo
da qual cortáram todo aquelle dia, parecendo a
todos que era algũa grande ilha que ali eſtaua, ſem
auer Piloto, nem outra peſſoa algũa que teueſſe
noticia

A transcrição deve ser fidedigna ao original, para satisfazer os objetivos lingüísticos das pesquisas.

- As normas de transcrição seguidas nesse sistema estão expostas na seção II.1 deste manual; o sistema seguido para obter a apresentação da transcrição do texto está em IV.

Entretanto, as características gráficas e grafemáticas dos textos mais antigos (preservadas nas transcrições conservadoras) dificultam o processamento automático posterior (anotação morfológica).

Para cumprir o objetivo de processamento automático, portanto, o texto original deve ser editado.

Capítulo Primeiro, De como se descobriu
esta província,
e a razão porque se deve
chamar Santa Cruz, e não
Brasil.

REINANDO aquele muito católico
e sereníssimo Príncipe el-Rei Dom
Manuel, fez-se uma frota para a Índia
de que ia por capitão mór Pedro Álvares
Cabral: que foi a segunda navegação
que fizeram os Portugueses para aquelas partes
do Oriente. A qual partiu da cidade de Lisboa a
nove de Março no ano de 1500. E sendo já entre
as ilhas do Cabo Verde (as quais iam demandar para
fazer aí aguada) deu-lhes um temporal, que foi causa
de as não poderem tomar, e de se apartarem alguns
navios da companhia. E depois de haver bonança junta
outra vez a frota, empegaram-se ao mar, assim por fugirem
das calmarias de Guiné, que lhes podiam estorvar
sua viagem, como por lhes ficar largo poderem dobrar
o cabo de Boa Esperança. E havendo já um mês,
que iam naquela volta navegando com vento próspero,
foram dar na costa desta província: ao longo
da qual cortaram todo aquele dia, parecendo a
todos que era alguma grande ilha que ali estava, sem
haver Piloto, nem outra pessoa alguma que tivesse
[notícia]

A edição dos textos inclui a modernização das grafias e a normalização dos aspectos grafemáticos, tornando-o assim adequado para o processamento automático.

Neste processo, entretanto, não desejamos perder as características do texto original, importantes para o estudo histórico da língua.

O desenvolvimento do sistema de edições críticas eletrônicas teve como objetivo solucionar essas demandas conflitantes.

- As normas de edição seguidas nesse sistema estão expostas na seção III.1 deste manual; o sistema seguido para obter a apresentação da versão editada do texto está em IV.

```
<section_title>
 <ed_mark id="ed_mark_373">
    <ed id="e_373">Capítulo</ed>
    <or id="o_373">Capit.</or></ed_mark> Primeiro, De como
 <ed_mark id="ed_mark_374">
    <ed id="e_374">se</ed>
    <or id="o_374">ſe</or></ed_mark>
 <ed_mark id="v_375">
   <ed id="e_375">descobriu<nl/></ed>
   <or id="o_375">deſ-<nl/>cobrio</or></ed_mark> esta
 <ed_mark id="ed_mark_376">
   <ed id="e_376">província</ed>
   <or id="o_376">prouincia</or></ed_mark>,
 <ed_mark id="ed_mark_377">
   <ed id="e_377">e</ed>
   <or id="o_377">&</or></ed_mark> a
 <ed_mark id="ed_mark_378">
   <ed id="e_378">razão</ed>
   <or id="o_378">razam</or></ed_mark> porque
 <ed_mark id="ed_mark_379">
   <ed id="e_379">se</ed>
   <or id="o_379">ſe</or></ed_mark>
 <ed_mark id="ed_mark_380">
   <ed id="e_380">deve</ed>
   <or id="o_380">deue</or></ed_mark><nl/> chamar
 <ed_mark id="ed_mark_381">
   <ed id="e_381">Santa</ed>
   <or id="o_381">Sancta</or></ed_mark> Cruz, e
 <ed_mark id="ed_mark_382">
   <ed id="e_382">não</ed>
   <or id="o_382">nam</or></ed_mark><nl/>
 <ed_mark id="ed_mark_383">
   <ed id="e_383">Brasil</ed>
   <or id="o_383">Braſil</or></ed_mark><nl/>
</section_title>
```

A edição é realizada sob forma de uma anotação em XML (eXtensible Markup Language) sobre a transcrição do texto. Nesta anotação, são codificados os itens originais e as interferências do editor. Cada uma dessas categorias pode ser mais tarde selecionada isoladamente do arquivo (via XSLT), atendendo portanto o objetivo de facilitar o processamento ao mesmo tempo em que se preservam as informações históricas.

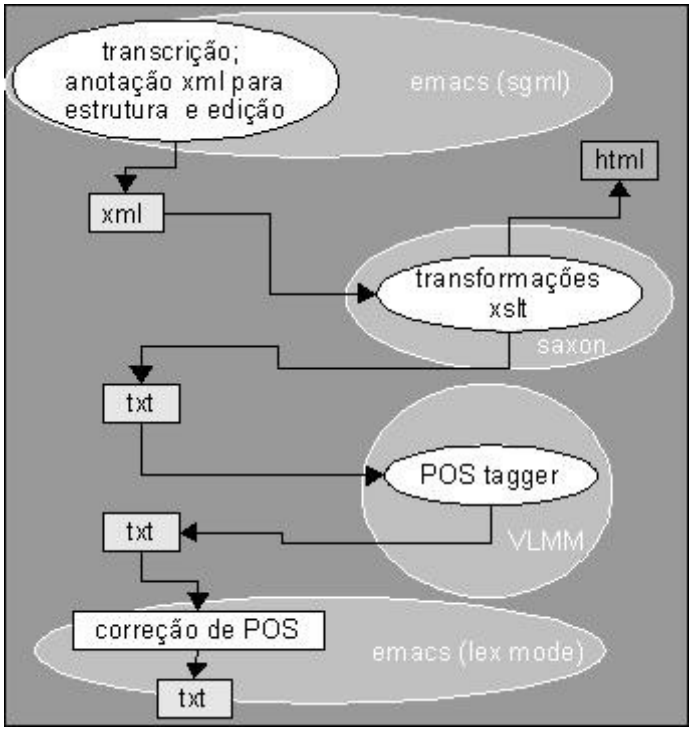
Ao lado se vê um trecho da anotação XML que gera as edições original e modernizada. Atualmente, a anotação é realizada manualmente, no processador Emacs.

- Os procedimentos para a anotação das estruturas de texto estão na seção II.2 deste manual; os procedimentos para a anotação das estruturas de edição estão em III.2.

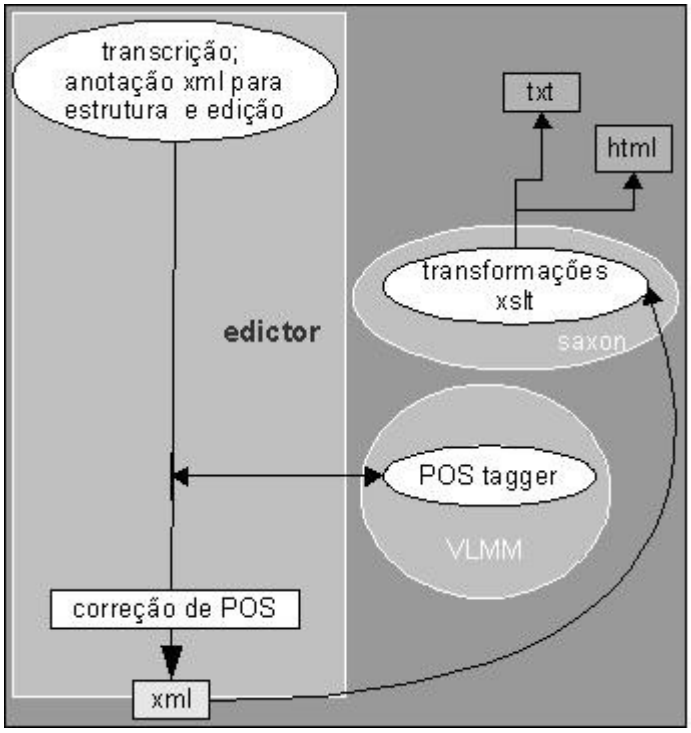
A etapa posterior é a anotação morfológica (POS) automática, realizada com base numa das versões geradas via XSLT (cf. Seção IV para esse procedimento):

transcrição	edição	anotação POS
REINANDO aquelle muy catho- lico & ſereniſsimo Principe elRey Dom MANVEL , fezſe hũa frota pera a India de que hia por capitam mór Pedralua- rez Cabral	Reinando aquele mui católico e sereníssimo Príncipe el-Rei Dom Manuel, fez-se uma frota para a Índia de que ia por capitão mór Pedro Álvares Cabral	Reinando/VB-AN aquele/D mui/ADV católico/ADJ e/CONJ sereníssimo/ADJ-S Príncipe/NPR el-Rei/NPR Dom/NPR Manuel/NPR ,/, fez/VB-D se/SE uma/D-UM-F frota/N para/P a/D-F Índia/NPR de/P que/WPRO ia/VB-D por/P capitão/N mór/ADJ-R Pedro/NPR Álvares/NPR Cabral/NPR

O sistema atual de anotação funciona em módulos separados, sendo cada tarefa desempenhada num ambiente de processamento diferente (cf. Fluxo de procedimentos).



Em breve, teremos à disposição uma ferramenta que permitirá trabalhar com maior segurança e em menor variedade de ambientes, o E-Dictor (Paixão de Sousa & Kepler, em curso).



0.2 O Processamento Eletrônico de Textos

0.2.1 Conceitos Introdutórios

Esta seção apresenta uma breve explanação introdutória em torno da seguinte pergunta:

O que é o processamento eletrônico de textos,
e como ele se diferencia das demais formas de processamento de textos?

Antes, cabe diferenciar os seguintes conceitos:

Texto Manuscrito *versus* Texto Produzido Manualmente;
Texto Impresso *versus* Texto Produzido Mecanicamente;
Texto Digital/Eletrônico *versus* Texto Produzido Eletronicamente

Os termos *Texto Manuscrito, Texto Impresso, e Texto Digital/Eletrônico* podem fazer referência à forma em que se apresentam os textos; enquanto os termos *Texto Produzido Manualmente, Texto Produzido Mecanicamente, e Texto Produzido Eletronicamente* fazem referência especificamente à forma em que são produzidos os textos. Um texto produzido mecanicamente pode se apresentar em forma digital (por exemplo, é o caso de fac-similes digitais realizados a partir de edições impressas ou manuscritas); em contraste, um texto produzido eletronicamente pode se apresentar em forma impressa.

Neste momento trataremos especificamente dos processos envolvidos na produção dos textos.

Vamos começar listando as principais diferenças materiais entre o texto processado eletronicamente e o texto processado manualmente ou mecanicamente:

- **Suporte**:
  Os textos produzidos manualmente ou mecanicamente têm por suporte (em geral) o papel;
  O texto produzido eletronicamente tem por suporte um equipamento eletrônico
  (discos de armazenagem - discos rigidos, CDs, etc.).

- **Tecnologia**:
  Os textos produzidos manualmente ou mecanicamente se constróem como
  um conjunto de sinais gráficos desenhados com pigmentos sobre o suporte.
  O texto produzido eletronicamente se constrói como
  um conjunto de pulsos de informação digital armazenados no suporte.

- **Codificação da Informação**:
  Os textos processados manualmente ou mecanicamente dependem do conhecimento de um sistema de codificação e decodificação de informação: um *sistema de escrita*.
  O texto produzido eletronicamente, também.

  Mas neste ponto, justamente, reside uma singularidade central do texto produzido eletronicamente: a **Mediação** da Codificação de Informação:

  - **Mediação**:
    A codificação dos textos produzidos manualmente/mecanicamente é imediata:
    Os sinais gravados no suporte são codificados diretamente pelo produtor, graças ao conhecimento do sistema de escrita.

A codificação do texto produzido eletronicamente é *mediada:*
Os sinais não são codificados no suporte diretamente pelo produtor, mas sim, passam por uma etapa intermediária, o *processamento computacional.*
Este processamento codifica os sinais a serem gravados no suporte (pulsos de informação) em sinais gráficos legíveis por um usuário conhecedor do sistema de escrita.

Esta necessidade de **mediação** determina a singularidade tecnológica da produção eletrônica de textos, em relação à produção manual ou mecânica.

Pois para processar o texto digital, dependemos de tecnologias de informação, ou seja, dependemos de programas que transformem sinais digitais em sinais gráficos legíveis.

Aqui voltamos brevemente à questão dos termos referentes à produção e à apresentação dos textos.

O problema da mediação tecnológica, de fato, é relevante no processo de codificação e no processo de de-codificação das informações – ou seja: tanto para *produzir* um texto em meio eletrônico como para *receber ou ler* um texto em meio eletrônico, dependemos dos processos computacionais.

Evidentemente, os textos produzidos manualmente ou mecanicamente podem também ser recebidos no ambiente eletrônico (como já dissemos, por meio das digitalização de fac-similes). De um modo geral, entretanto, tais textos são tipicamente produzidos com a finalidade de serem recebidos na forma de objetos físicos compostos por matéria subjetiva (em geral, o papel) e matéria aparente (em geral, a tinta). Como dissemos, as informações são codificadas, nestes casos, pela aplicação da matéria aparente sobre a matéria subjetiva – de forma manual (pelo desenho da mão, com auxílio de instrumentos como a pena ou o lápis) ou mecânica (pela aplicação de desenhos, com auxílio de instrumentos como o tipo gráfico). Na outra ponta do processo – a recepção do texto – tanto no caso de produção manual como mecânica, o processo de decodificação dos sinais gráficos codificados em suportes físicos é visual.

No prosseguimento desta exposição, faremos referência a algumas questões técnicas envolvidos nas duas pontas do processamento de textos em meio eletrônico – produção e a recepção. Passamos agora, então, a usar o termos **Texto Digital** para nos referirmos àquele tipo de texto não apenas produzido em meio eletrônico, como também recebido em meio eletrônico - ou seja, o texto processado eletronicamente.

0.2.2 A codificação eletrônica da informação

Para exemplificar o tipo de processo computacional envolvido nessa mediação da codificação da informação no meio eletrônico, vamos considerar por um instante a (de)codificação dos *caracteres de escrita.*

Quando "escrevemos" um texto por sistema manual, por exemplo, o que fazemos é desenhar com a tinta no palpel caracteres de um sistema convencional de codificação da linguagem (o alfabeto). O caractere por nós desenhado é recebido imediatamente pelo leitor, sem um instrumento mediador.

desenho A > sinal gráfico A

O texto pode também ser "escrito" por sistema mecânico, ou seja, pelo auxílio de máquinas, como por exemplo a máquina de escrever. Mas a tecnologia da máquina consiste simplesmente em "*carimbar*" os desenhos que nós teríamos que traçar com as mãos. Neste sentido, não há mediação tecnológica *substantiva* , apenas uma mediação *acessória* ao traçado de letras. O mesmo se aplica para outros instrumentos mecânicos de mediação da escrita, como a prensa mecânica (o que inclui as máquinas gráficas mais modernas, que entretanto seguem o princípio do "*carimbo*" - ou seja, impressão física – de sinais gráficos a serem decodificados visualmente).

O que acontece quanto "*escrevemos*" um texto no meio eletrônico?

Parece-nos, à primeira vista, estarmos realizando uma atividade análoga à escritura com uma máquina de escrever. Mas a analogia é ilusória: o teclado do computador é apenas em aparência igual ao teclado das máquinas. O teclado da máquina, como vimos, ativa mecanicamente uma série de *carimbos* de sinais gráficos. O teclado do computador ativa *comandos matemáticos* que são processados diversas vezes por programações embutidas nos computadores, antes de aparecer nas telas como sinais gráficos.

Assim, quando digitamos um caractere "*A*", a tecla correspondente *ativa um comando* que é processado pelo computador mais ou menos assim: "*inserir o sinal gráfico identificado como [A]*":

comando x (*ativação pelo teclado:* **[A+shift]) > código y > sinal gráfico A**

Para fazer a correspondência entre os comandos e os sinais gráficos, há uma série de listas estandardizadas, como ASCII, ISO, Unicode 1 ; veja o exemplo abaixo, parte da lista de codificação Unicode:

Sintaxe dos Códigos Unicode: &#NNNN;
Lista de Códigos Decimais e equivalências:
unicode 0000 ~ 0255

	0009 **HT**	0010 **LF**			0013 **CR**		
0032 **SP**	0033 !	0034 "	0035 #	0036 $	0037 %	0038 &	0039 '
0040 (	0041)	0042 *	0043 +	0044 ,	0045 -	0046 .	0047 /
0048 0	0049 1	0050 2	0051 3	0052 4	0053 5	0054 6	0055 7
0056 8	0057 9	0058 :	0059 ;	0060 <	0061 =	0062 >	0063 ?
0064 @	0065 A	0066 B	0067 C	0068 D	0069 E	0070 F	0071 G
0072 H	0073 I	0074 J	0075 K	0076 L	0077 M	0078 N	0079 O
0080 P	0081 Q	0082 R	0083 S	0084 T	0085 U	0086 V	0087 W
0088 X	0089 Y	0090 Z	0091 [	0092 \	0093]	0094 ^	0095 _
0096 `	0097 a	0098 b	0099 c	0100 d	0101 e	0102 f	0103 g
0104 h	0105 i	0106 j	0107 k	0108 l	0109 m	0110 n	0111 o
0112 p	0113 q	0114 r	0115 s	0116 t	0117 u	0118 v	0119 w
0120 x	0121 y	0122 z	0123 {	0124 |	0125 }	0126 ~	

(fonte: http://en.wikipedia.org/wiki/List_of_HTML_decimal_character_references)

A estandardização dos códigos para caracteres garante que em qualquer computador do mundo, o comando "*A*" resulte no sinal gráfico "*A*", através da mediação de um programa. Se o programa inclui a codificação unicode, por exemplo, estas são as traduções para a "escrita" dos sinais **A** e **a**:

comando x (*ativação pelo teclado:* **A+shift) > código &#0065 > sinal gráfico A**

comando x (*ativação pelo teclado:* **A) > código &#0097 > sinal gráfico a**

Note-se que os códigos acima correspondem aos grafemas e aos módulos dos grafemas (maiúsculas ou minúsculas), não aos **tipos ou fontes** [2.]

Podemos observar melhor os efeitos da transposição dos comandos em sinais gráficos nas ocasiões em que ela *não funciona* corretamente.

Por exemplo, quando usamos um computador que não está programado para interpretar diacríticos e acentos, ou quando não sabemos se em determinado teclado devemos digitar **[shift-~]** ou **[alt-~]**, e por fim não conseguimos obter sinais gráficos como **ç, ã, é, ü** ... Isso funciona bem apenas em computadores programados para lidar com tais sinais gráficos.

Outra ocasião em que podemos ver a mediação em processo é na leitura de páginas-web. Em alguns casos, os sinais gráficos para acentos, etc., aparecem "*corrompidos*" nas páginas (por exemplo, "**corrup**[illegible]**o**" por "**corrupção**"). Isso acontece por alguma incompatibilidade entre o programa usado para mediar a transposição naquele texto e o programa presente no computador de acesso, ou por um erro de codificação. Pode-se experimentar, também em qualquer navegador, mudar a opção padrão de decodificação de caracteres, para testar seus efeitos. Numa página codificada em sistema "UTF-8", por exemplo, se mudarmos a opção de codificação para "Chinês Simplificado GB2312", teremos o seguinte efeito:

Seja bem-vindo(a) à Wikipédia, uma enciclopédia livre e gratuita, feita por pessoas como você em mais de 200 idiomas! Seja bem-vindo(a) 腠 Wikip 茅 dia, uma enciclop 茅 dia livre e gratuita, feita por pessoas como voc 锚 em mais de 200 idiomas!

Assim, o ciclo da mediação da codificação da informação, nas duas pontas do processo, funciona perfeitamente quando tanto o produtor como o receptor do texto tiverem acesso à mesma programação de codificação.

A partir do que foi exposto até este ponto, podemos definir de modo intuitivo o que é um processador de textos: é uma ferramenta que põe em funcionamento um programa que realiza a mediação entre comandos digitais e sinais humanamente legíveis.

Naturalmente, essa mediação não envolve apenas o *inventário de caracteres* (sinais gráficos). Envolve, também, a **organização espacial e informacional do texto**.

Nos processadores modernos, podemos arrumar os sinais gráficos em espaços delimitados (página), dividir o texto em blocos (parágrafos) , destacar trechos visualmente (formatos), e até ordenar as informações (estrutura de tópicos). A exemplo do que dissemos para os caracteres, cada uma dessas operações envolve um processo de codificação e decodificação complexo em diferentes etapas 3.

No entanto, ao fazermos uso dos programas de processamento de texto, não temos consciência destas operações. A tecnologia desenvolvida nestes programas é concebida, justamente, de modo a simular as ações que realizaríamos no papel, de modo a que a produção do texto seja confortável e facilitada, e que os programas possam ser operados intuitivamente pelos usuários.

Um bom exemplo dessa adaptação ao conforto humano é a **página**.

A maioria dos processadores nos apresenta, na tela, o texto organizado tal qual apareceria numa página de papel. Em princípio, o conceito de "página" não faz sentido no ambiente da tela – trata-se de fato de uma unidade espacial própria do texto em papel. Esse conceito é herdado dos textos impressos ou manuscritos, e os processadores o reproduzem para nosso conforto. Ou seja, a ferramenta cria um espaço "página" para nele dispormos o texto; esse espaço é uma interface visual mediada por programações. Evidentemente, a "página" que aparece na tela não tem consistência material direta como a página de papel; é uma representação visual mediada por códigos.

As tecnologias de interface intuitiva embutidas nos processadores de texto representam um avanço técnico considerável, e que tornou a produção de textos uma tarefa muito mais fácil (ao menos, deste ponto de vista físico).

Entretanto, nos programas atualmente disponíveis, o conforto de utilização é inversamente proporcional ao controle do usuário sobre as tecnologias de processamento. Quanto mais sofisticadas forem as simulações intuitivas das manipulações computacionais embutidas nos programas, maior o grau de mediação entre o usuário e a codificação da informação.

Quando trabalhamos com o texto como objeto de pesquisa, torna-se desejável trabalhar com um menor nível de conforto, compensado por um maior nível de *controle sobre a mediação tecnológica, como se trata na seção seguinte (2. Controle do Processamento Eletrônico). Vamos ver, então, como a maioria dos formatos para o processamento digital de textos podem ser classificados, quanto ao grau de mediação da codificação de informação.*

0.1.3 Breve tipologia dos textos eletrônicos

Atualmente, há diferentes formatos de texto eletrônico (ou tipos de arquivos), cada um com diferentes finalidades e qualidades.

Antes de tudo, quanto à finalidade dos formatos, é interessante separá-los em dois tipos:

- Documentos **locais** são aqueles tipicamente produzidos, acessados e armazenados localmente, em computadores pessoais;
- Documentos **remotos** são aqueles tipicamente produzidos para serem acessados e armazenados *em rede* - principalmente, via internet.

Procuraremos agora analisar estes dois grupos quanto a algumas qualidades que interessam centralmente ao processamento com finalidade de estudo:

- **Complexidade**: É possível codificar informações complexas neste tipo de documento?
- **Controle:** Qual o grau de controle do produtor quanto às informações que podem ser codificadas neste tipo de documento?
- **Portabilidade:** As informações codificadas neste tipo de documento serão igualmente legíveis em diferentes pontos de acesso?
- **Confiabilidade**: As informações codificadas neste tipo de documento poderão ser acessadas futuramente?

Vamos iniciar com os arquivos do grupo local, que inclui os documentos produzidos pela maioria dos processadores comuns:

txt (arquivos independentes):

- Complexidade: Não permitem a codificação de informações complexas, e nem mesmo formatações.

- Controle: Oferecem amplo grau de controle do processamento; entretanto, isto se torna pouco relevante, pelo baixo grau de complexidade das codificações possíveis.
- Portabilidade: São amplamente transportáveis, e independentes tanto em termos de programas como em termos de sistemas operacionais (Windows, Linux, Mac).
- Confiabilidade: São amplamente confiáveis; o txt é o tipo de arquivo mais antigo e mais simples atualmente disponível.

DOC, arquivos do software *Word*, da *Microsoft*:

- Complexidade: Permitem codificar uma ampla gama de informações complexas, como formatações, estruturas de documento, etc.
- Controle: Oferecem baixíssimo grau de controle do processamento, pois a programação de mediação usada neste programa é *fechada*, ou seja, seu código é desconhecido e só pode ser operado pelo fabricante.
- Portabilidade: Os arquivos .doc são feitos para serem lidos seja por computadores que possuam o mesmo programa Word, seja impressos. Recentemente, outros programas (como o Open Office) passaram a oferecer a possibilidade de leitura de arquivos .doc.
- Confiabilidade: Estão na escala mais extrema de dependência a um programa. Primeiro, pelo fato da programação ser fechada; segundo, porque o programa é *proprietário*. A única garantia de que este tipo de documento poderá ser lido/acessado no futuro é o compromisso do fabricante junto a seus clientes.

ODT ("*open document text*"), arquivos do software *Open Office,* da *Sun Microsystems*:

- Complexidade: Permitem codificar uma ampla gama de informações complexas, como formatações, estruturas de documento, etc.
- Controle: Oferecem maior grau de controle do processamento em relação aos documentos word, sendo sua programação **aberta**. Ou seja: seu código é fornecido pelo fabricante, e pode ser alterado/desenvolvido por desenvolvedores independentes.
- Portabilidade: Assim como os arquivos **.doc**, os arquivos **.odt** só podem ser produzidos e acessados no programa apropriado. Entretanto, o programa *Open Office* é **livre**.
- *Confiabilidade: Pelo fato de o programa der livre e de código aberto,* há uma garantia maior de que este tipo de documento possa ser lido/acessado no futuro; pois, ainda que o fabricante pare de produzir o programa, outros desenvolvedores poderão produzir ferramentas que permitam trabalhar com os arquivos.

De outro lado, estão os documentos de natureza fundamentalmente remota. Este tipo de documento, por natureza, tem uma vocação de informação partilhada: são textos feitos em uma máquina para serem acessados em outra máquina (e portanto, transportáveis por natureza). Há fundamentalmente dois tipos de documentos mais utilizados para leitura em máquinas:

PDF ("*Portable Document File"),* arquivos do software *Adobe Acrobat*:

- Complexidade: Permitem codificar uma ampla gama de informações complexas, como formatações, estruturas de documento, etc.
- Controle: Os arquivos PDF são produzidos tipicamente pelo software *Adobe Writer* (*proprietário* e **fechado**) e lidos pelo software *Adobe Reader* (**livre** e *fechado*). No PDF, o texto que vemos na tela é, praticamente, uma figura, uma uma fotografia do texto. Isso é o que garante a transposição das informações codificadas nesses arquivos. Assim, para o processamento de informações lingüísticas, este formato não é adequado - a não ser que se tenha acesso ao programa de produção, o *Adobe Acrobat Writer*, que como dito, é proprietário.
- Portabilidade: Como indica o próprio nome, o PDF é um formato programado para ser portátil. Ou seja, é codificado de tal maneira que as informações gráficas se transportam com absoluta integridade de um local de acesso para o outro (por exemplo, as páginas quebram nos pontos exatos; as fontes se mostram

sempre iguais, etc. Trata-se portanto de um formato altamente confiável nesta perspectiva. Entretanto, para acessar os arquivos, é preciso ter acesso ao programa *Reader, que é livre.*

- Confiabilidade: Apresentam alto grau de dependência a um programa; a programação é fechada e o programa é *proprietário*. A única garantia de que este tipo de documento poderá ser lido/acessado no futuro é o compromisso do fabricante junto a seus clientes.

Hipertexto - arquivos .html, .xhtml, .xml, etc. - linguagem livre:

- Complexidade: Permitem codificar uma ampla gama de informações complexas. Além das informações codificadas nos demais arquivos (como formatações, estruturas de documento, etc.), permitem a interação com tecnologias complexas como bases de dados, buscas, etc.
- Controle: Oferecem grau muito amplo de controle do processamento, em especial no formato XM; mais adiante, falaremos dos diferentes formatos de hipertexto (cf. Seção 2).
- Portabilidade: O Hipertexto é o texto eletrônico por excelência: trata-se de um formato concebido para acesso em máquinas, inteiramente independente do suporte tradicional (papel). Por isso, apresenta maior potencialidade para leitura na tela, explorando as possibilidades de relação complexa entre diferentes partes do texto (as hiper-ligações ou hiperlinks). A origem em rede significa que o hiptertexto surgiu como formato compartilhado por naturez; *para acessá-los, basta o acesso a um navegador de hipertexto (Internet Explorer, Mozilla, Firefox, Safari, etc.; quase todos livres, alguns deles abertos)* .
- Confiabilidade: Por ser concebido para a leitura em diferentes máquinas, o texto é processado de acordo com normas estandardizadas internacionalmente. O processo de normatização da codificação de hipertextos é **livre** e *aberto*, e tem sido conduzido por um consórcio de pesquisadores associados sem fins lucrativos. Àparte o .txt, o hipertexto está na escala mais extrema de independência com relação a programas, não sendo necessário um programa específico para produzir os textos (embora seja possível utilizar um dos muitos programas proprietários ou livres disponíveis hoje). .

0.3 Controle do Processamento Eletrônico

0.3.1 Motivações para o processamento controlado

A motivação central para o processamento controlado de textos é a **preservação da integridade das informações codificadas nos textos**.

Isso pode ser fundamental no trabalho de natureza filológica - no qual o texto é o objeto central do estudo. Ultimamente, a transcrição e edição de textos antigos tem sido realizada em meio eletrônico - ao menos, no plano do arquivamento da documentação. Neste trabalho de transcrição e edição, codificam-se as informações relevantes sobre os textos. De um modo geral, costumamos adaptar as codificações anteriormente utilizadas para o meio impresso.

Por exemplo, para indicar adição de informações ao texto original (caso dos desdobramentos de abreviaturas), é comum utilizar-se o código de *itálico*:

Texto original: **VM** > Texto editado: **V***ossa* **M***ercê*

Ou seja, utilizamos processos de *formatação* dos textos eletrônicos para codificar informações importantes.

Entretanto, tendo em vista o que vimos acima sobre a mediação tecnológica, precisamos lembrar que a formatação é uma codificação intermediária; e não há garantias, em princípio, de que a formatação aplicada a um texto em determinado processador seja legível por outros processadores, nem que seguirá sendo legível no futuro.

De fato: não há garantias de que *qualquer informação* incluída em um texto eletrônico seja legível no futuro.

Pois, ao contrário do texto impresso, o texto digital depende de programas para ser lido. Se os programas mudarem, *as informações se perdem*.

No trabalho filológico, a formatação codifica informações cruciais, que não desejaríamos ver perdidas. Assim, podemos listar ao menos duas razões pontuais pelas quais o controle do processamento de textos pode ser desejável neste tipo de trabalho:

- **Codificação de caracteres**:
  O processamento controlado pode garantir a transposição correta de caracteres independente do programa de processamento utilizado.

- **Codificação de informações complexas:**
  O processamento controlado pode garantir a tradução correta de informações complexas (como organização gráfica e formatação) independente do programa de processamento utilizado.

0.3.2 O Hipertexto como forma de processamento controlado

A breve tipologia esboçada na seção 1.3 listou alguns dos principais formatos de arquivos de texto eletrônico hoje disponíveis, de acordo com o grau de complexidade de informações que podem ser codificadas, o controle que se pode exercer sobre esta codificação, a portabilidade do texto produzido, e o grau de confiança que se pode ter em cada formato.

Dali se depreende que o Hipertexto é a principal opção a ser explorada, quando se pensa em conjugar complexidade, controle, portabilidade e confiabilidade.

O conteúdo da rede mundial de computadores se constrói, tipicamente, neste formato. A apresentação do texto, tal como visualizada na tela, esconde um código, que pode ser acessado pela barra de ferramentas do navegador (normalmente, no item *exibir>codificação* (ou: *código-fonte*).

O Hipertexto é tipicamente codificado na linguagem **html** ou *HiperText Markup Language* . Vamos ver como é um mesmo trecho de texto codificado em **html**, na versão para tela, e em em código:

Trecho visível – apresentação na tela:

As sentenças deste trecho estão inseridas em parágrafos, que são estruturas codificadas como "**p**". As propriedades de formatação, como **negrito**, *itálico* e sublinhado, são indicadas por diferentes códigos (respectivamente, "**b**", "**i**", e "**u**"). Os códigos correspondentes às estruturas são indicados por parênteses angulares, ou < >. Observe que os parênteses angulares são caracteres especiais, e no código-fonte não aparece o sinal gráfico, mas o equivalente em código **iso-8859-1**. Isso vale também para todos os acentos e diacríticos usados neste texto. Por exemplo, " **ó**" equivale a "**ó**".

Outras informações, como o recuo do texto, são também codificadas (neste caso, na estrutura "**blockquote**"). Alguns atributos, como o alinhamento do texto, são indicados na respectiva estrutura. Por exemplo, **p align="justify"** gera um parágrafo justificado, como este.

As estruturas são organizadas, no código, em árvores. Assim, neste trecho, as estruturas de formatação de caracteres estão contidas nos parágrafos, e os parágrafos estão contidos na estrutura de formatação "**blockquote**".

Trecho em código:

```
<p align="justify">As senten&ccedil;as deste trecho est&atilde;o inseridas em
par&aacute;grafos, que s&atilde;o estruturas codificadas como &quot;<b>p</b>&quot;. As
propriedades de formata&ccedil;&atilde;o, como <b>negrito</b>, <i>it&aacute;lico</i> e
<u>sublinhado</u>, s&atilde;o indicadas por diferentes c&oacute;digos (respectivamente,
&quot;<b>strong</b>&quot;, &quot;<b>i</b>&quot;, e &quot;<b>u</b>&quot;).
</p>

<p align="justify">Os c&oacute;digos correspondentes &agrave;s estruturas s&atilde;o
indicados por par&ecirc;nteses angulares, ou &lt; &gt;. Observe que os par&ecirc;nteses
angulares s&atilde;o caracteres especiais, e no c&oacute;digo-fonte aparece n&atilde;o o sinal
gr&aacute;fico, mas o equivalente em c&oacute;digo iso-8859-1. Isso vale tamb&eacute;m
para todos os acentos e diacr&iacute;ticos usados neste texto. Por exemplo,
&quot;<b>&amp;oacute;</b>&quot; equivale a &quot;<b>&oacute;</b>&quot;. </p>

<blockquote>

<p align="justify">Outras informa&ccedil;&otilde;es, como o recuo do texto, s&atilde;o
tamb&eacute;m codificadas (neste caso, na estrutura &quot;<b>blockquote</b>&quot;).
Alguns atributos, como o alinhamento do texto, s&atilde;o indicados na respectiva estrutura
(por exemplo,<b>p align=&quot;justify&quot;</b>gera um par&aacute;grafo justificado)
</p>
<p align="justify">> As estruturas s&atilde;o organizadas, no c&oacute;digo, em
&aacute;rvores. Assim, neste trecho, as estruturas de formata&ccedil;&atildeo de caracteres
est&atilde;o contidas nos par&aacute;grafos, e os par&aacute;grafos est&atilde;o contidos na
estrutura de formata&ccedil;&atilde;o &quot;<b>blockquote</b>&quot;</p>

</blockquote>
```

Todo texto em formato **html** que circula pela internet tem mais ou menos esta estrutura de código. Embora isto possa parecer um tanto complexo, é fundamental ressaltar que esta característica de "codificação embutida" não é exclusiva do formato hipertexto – é de fato própria a todo tipo de arquivo de texto eletrônico.

De fato: se pudéssemos olhar a codificação dos documentos que produzimos nos nossos processadores locais, veríamos codificações ainda mais complexas (incluindo quebras de página; formato da página; tamanho e tipo das letras; estrutura de tópicos, etc.).

Ou seja: o hipertexto codificado em **html** não é mais complexo; apenas permite o acesso ao código.

Os códigos **html** padrão são normatizados, e constituem um conjunto fechado de etiquetas (ou rótulos, como **<b>**, **<i>,** ou **<p>**, exemplificados acima), que precisam ser respeitadas, para poderem ser processadas pelos navegadores (saiba mais em http://www.w3.org/MarkUp/).

Assim, o **html** é uma linguagem acessível, livre e bastante controlada, ideal para o processamento e formatação do hipertexto de natureza mais geral; os códigos **html** disponíveis são próprios para codificar estruturas de *organização* e *formatação* dos textos.

Entretanto, o **html** não é necessariamente a melhor opção para processamentos mais específicos, como é o caso das edições eletrônicas de natureza filológica. Pois nestes casos, além de informações gerais sobre organização e formatação, há outras informações específicas que precisamos codificar. Podemos exemplificar a gama de informações comumente codificadas em edições filológicas com este pequeno quadro retirado de Marquilhas (1996):

Códigos de Edição

{.} sinal ilegível por cancelamento
{...} trecho ou palavra ilegível por cancelamento
[.] sinal ilegível por deterioração do material
[...] sinal ilegível por deterioração do material
[?] sinal de difícil leitura
[???] vocábulo de difícil leitura
(.) branco superior ao espaco entre palavras e inferior à largura da linha
(*) linha em branco
(&) linha escondida pela encadernação
[&] linhas escondidas pela encadernação
($) seção em branco
(-) linha cancelada
(~) trecho manchado
{*texto*} trecho ou palavra cancelados
|**texto**| trecho coberto por segunda camada gráfica
/**texto**/ texto escrito na entrelinha
//texto// texto escrito na margem
(**texto**) texto de outra caligrafia

Para codificar informações deste tipo, precisamos de um controle ainda maior sobre o processamento dos textos.

De fato: precisamos poder *criar nossas próprias categorias de informação estruturada*.

Isto é possível com a linguagem **XML**, ou **eXtended Markup Language**, como se apresenta em seguida.

0.3.3 A linguagem de anotação extendida, ou XML

A linguagem **XML** apresenta todas as vantagens da linguagem **html**; mas a isso se acrescenta sua natureza inteiramente flexível.

Na anotação dos textos do *Corpus Histórico Tycho Brahe*, por exemplo, além de anotarmos a categoria **parágrafo<p>** (como no trecho **html** acima), anotamos a categoria **sentença** (**<s>**); além disso, numeramos cada sentença com um código identificador, o que facilita a pesquisa posterior (exemplo: **<s id="g_008_611">** é a sentença número *611* do texto *g_008*) .

As categorias usadas por cada anotador são inteiramente abertas, desde que se respeitem parâmetros estandardizados de estruturação - ou seja, há limitações sintáticas, mas não semânticas.

Além disso, a linguagem **XML** permite uma grande flexibilidade na apresentação dos documentos. Enquanto cada documento **html** é apresentado na tela diretamente pela interpretação do código por parte do navegador, o documento **XML** pode servir de base para se gerarem inúmeras versões de cada documento (por exemplo, documentos **html** para serem acessados via navegador, ou documentos txt para serem processados por programações computacionais especializadas).

O mais importante é que nas diferentes versões, as informações podem ser re-ordenadas, selecionadas, re-formatadas, etc., sem prejuízo da anotação de base. Isso confere à linguagem **XML** um nível muito elevado de controlabilidade.

Um manual completo da linguagem XML e suas correlatas (XSL, X-Q) pode ser consultado em:
http://www.w3schools.com/xml/default.asp

No sistema de edições eletrônicas que aqui apresentamos, a linguagem XML (e sua linguagem de programação associada, o XLST) é utilizada em todas as etapas do processamento de textos:

- **Transcrição** dos documentos
  (na *aplicação de códigos de estruturas como quebras de linhas e parágrafos*);

- **Edição** dos documentos
  (na inserção de comentários e modificações do editor, sob forma codificada e controlada);

- **Apresentação** dos documentos
  (na *geração de diferentes versões a partir da anotação estrutural e edição de base*).

A seção 4 a seguir detalha um pouco de cada etapa, procurando explicar sua relação com as etapas tradicionais de edição especializada.

0.4 Etapas do Trabalho de Edição Eletrônica

0.3.1 Procedimentos de Edição – Visão Geral

Os procedimentos básicos a serem conduzidos na edição de textos podem ser assim resumidos:

Rescensão: Estudo, reunião e seleção do material a ser editado

Cópia: Transcrição, ou reprodução do texto em novo suporte material;
inclui eventuais ajustes na forma e organização espacial do texto e transliteração grafemática.

Estabelecimento do Texto: Intervenção do editor no sentido de facilitar a leitura; inclui a proposição de conjecturas sobre lições difíceis e a uniformização da forma gráfica. Inclui também a seleção de variantes, nos casos de textos com mais de um testemunho ou versão original.

Apresentação: Organização da edição na forma final a ser lida publicamente.

As etapas de cópia e de preparação envolvem procedimentos de **interpretação** do texto. Nas edições especializadas, esta interpretação deve ser tornada explícita para o leitor final. A interpretação e sua explicitação podem ser realizadas de acordo com diferentes critérios, que variam, sobretudo, quanto ao grau de interferência realizado pelo editor em relação ao texto. Tradicionalmente, as edições são agrupadas de acordo com este critério de maior ou menor proximidade em relação ao texto original. A combinação dos procedimentos acima pode resultar em três tipos básicos de edição especializada:

1. **Edições Conservadoras, ou *Diplomáticas*:**

   A transcrição é conservadora (mantém as variantes e arcaísmos grafemáticos, a separação vocabular, a paragrafação, as lacunas, etc.); não há procedimentos de intervenção do editor.

2. **Edições Semi-interpretativas, ou *Semi-Diplomáticas*:**

   A transcrição é menos conservadora (uniformiza variantes e arcaísmos grafemáticos e separação intravocabular, intervém na paragrafação); os procedimentos de intervenção do editor incluem a proposição de conjecturas sobre lições difíceis e o desenvolvimento de abreviaturas.

3. Edições Interpretativas, ou *Modernizadas*:

Além dos procedimentos conduzidos nas edições semi-interpretativas, neste caso a intervenção do editor pode incluir a uniformização das grafias do original transcrito.

A diferença central entre o trabalho de edição tradicional e o trabalho de edição eletrônica é que com o processamento eletrônico, os "*tipos de edição*" podem ser re-definidos como "*tipos de apresentação*". No sistema aqui desenvolvido, a transcrição é sempre conservadora; os procedimentos interpretativos podem ser mais ou menos intensos a depender do texto e da finalidade da edição – mas cada texto pode ser apresentado em uma versão *Diplomática*, *Semi-Diplomática* ou *Interpretativa-Modernizada*, separando-se na etapa da apresentação as diferentes camadas de intervenção executadas no texto, como se mostra a seguir.

0.3.2 Procedimentos da Edição Eletrônica

O processo de produção das edições eletrônicas neste sistema envolve as seguintes etapas (cf. também diagrama 1 a seguir):

A. Rescensão – Pesquisa e Catalogação

O processo se inicia pela pesquisa acerca da obra a ser editada, o que inclui um levantamento de sua história editorial e a reunião das edições disponíveis, selecionando-se, entre elas, a que se julgar mais adequada para os objetivos da edição.

As informações sobre cada texto selecionado são sistematizadas em uma ficha catalográfica onde se incluem informações sobre o texto original (autor, data, gênero, etc.); sobre a edição utilizada como fonte, caso esta não coincida com o original (editor, data, processo de estabelecimento do texto); e sobre a edição eletrônica (editor, processo de estabelecimento do texto, etapas de edição).

Estas informações são anotadas nos cabeçalhos dos documentos de base onde será realizada a edição, e podem ser acessadas posteriormente para fins de classificação e busca dos textos, por meio de uma programação gravada no servidor onde se armazenam os documentos (programação X-Query).

B. Cópia – Transcrição ou Transposição eletrônica

A etapa de cópia equivale a uma transposição sistematizada do suporte original para o suporte eletrônico. Ela envolve procedimentos distintos a depender da natureza do material que está sendo editado eletronicamente: textos originais, ou preparações de outros editores. Nos dois casos, chamaremos o texto a ser editado de *texto-fonte:*

- Quando o texto-fonte é um texto original (impresso ou manuscrito; ascessado diretamente ou via fac-simile), esta etapa se resume na transcrição fidedigna do texto. A transcrição inclui, neste sistema, a transposição dos caracteres de escrita e a reprodução da organização do texto (via anotação XML).

- Quando o texto-fonte é uma preparação de outro editor, esta etapa envolve a adaptação do material produzido pelo primeiro editor. Temos acesso a este material sob diferentes formas: livros impressos e documentos eletrônicos. Os livros impressos são escanerizados, e o resultado da escanerização é revisto palavra por palavra (corrigindo-se os erros do sistema de escanerização, em geral o OCR). Os documentos eletrônicos nos chegam, em geral, no formato .doc ou .rtf. Estes arquivos são transformados em documentos XML com a correta codificação de caracteres. As indicações do primeiro editor sobre a organização do texto são também adaptadas para a anotação XML.

No caso de se utilizar mais de um testemunho do texto para a edição – ou seja, na realização de edições

críticas – o processo de transposição aqui desenvolvido permite que as características do texto em cada uma das edições sejam trancritas, e as interferências de cada editor sejam codificadas individualmente, de forma a que todo este material seja incluido em um único documento. Este documento crítico permite o acesso posterior a todas as camadas aplicadas subsequentemente ao texto.

C. Preparação e Estabelecimento do Texto – Edição Interpretativa

A profundidade das intervenções editoriais em um texto depende da finalidade da edição. Este sistema de edições eletrônicas foi desenvolvido fundamentalmente para preparar os textos para a análise lingüística automática. Assim se estabelecem dois objetivos:

- o resultado da edição deve ser fidedigno ao original nos planos lexical, morfológico, sintático e textual;

- o resultado da edição precisa ser uniformizado no aspecto ortográfico.

A satisfação desses dois objetivos molda as diretrizes da preparação dos textos neste sistema de edição. Assim, depois da etapa de transcrição acima descrita, os textos passam por procedimentos de unformização ortográfica; entretanto, são tomados todos os cuidados no sentido de preservar a forma do texto original nos planos lexical, morfológico e sintático.

Nesta etapa de preparação, a diferença primordial entre o sistema eletrônico e o sistema tradicional é que nas edições eletrônicas, é possível preservar as formas originais no mesmo plano em que se acrescentam formas interpretadas pelo editor (uniformizações, conjecturas, etc.).

D. Apresentação

Nas edições eletrônicas, o sistema de apresentação da edição – ou organização da leitura pública final dos textos – difere também bastante do sistema tradicional. No sistema tradicional, a apresentação final é uma etapa consequente da transcrição e do estabelecimento do texto , ou seja, o próprio documento preparado é organizado para a leitura final.

Nas edições eletrônicas, a apresentação final é uma etapa paralela à transcrição e ao estabelecimento do texto – ou seja: as apresentações para leitura final resultam da formação de documentos paralelos ao documento-base onde se codificou a preparação. É o que chamamos de geração de versões.

Nos documentos-base, a transcrição do texto e as intervenções do editor estão codificadas ou anotadas em linguagem XML; essas anotações podem ser lidas por programações computacionais que selecionam e organizam as estruturas dos documentos-base e formam, a partir delas, os documentos de apresentação[4].

Como visto acima, a dupla articulação presente nos objetivos destas edições (fidedgnidade lingüística e uniformização de grafia) resulta na edição em camadas dos documentos-base. Na etapa da apresentação, estas camadas de edição podem ser separadas para o acesso do leitor final. Assim, um mesmo documento pode ser lido nas versões Edição Conservadora, Edição Semi-Interpretativa, Edição Interpretativa, etc. A geração das versões, por ser paralela, não traz nenhuma conseqüência para os documentos-base, que permanecem inalterados.

Diagrama 1

Edições Eletrônicas: Fluxo dos Procedimentos

<table>
<tr><td rowspan="2">I. Pesquisa
e
Catalogação</td><td colspan="2">• Pesquisa sobre a história editorial
• Reunião e seleção de edições disponíveis
• Composição das fichas catalográficas
• Descrição e classificação segundo a história editorial
• Inclusão no catálogo eletrônico via cabeçalhos de metadata (xml)</td></tr>
<tr><td>Textos Originais

• manuscritos
• impressos

• em papel
• digitalizados</td><td>Outras Edições

• semi-diplomáticas
• modernizadas</td></tr>
<tr><td>II. Transcrição</td><td>• Transcrição (xml)

• Anotação das estruturas de texto (xml)</td><td>• Adaptação do sistema original (revisão de OCR, adaptação de documentos Word) (xml)

• Adaptação das indicações de estruturas de texto (xml)</td></tr>
<tr><td>III. Edição Interpretativa</td><td>• Uniformização Grafemática (xml)

• Uniformização Ortográfica (xml)</td><td>• Adaptação das indicações editoriais anteriores (xml)

• Uniformização Ortográfica (xml)</td></tr>
<tr><td>IV. Apresentação</td><td colspan="2">Geração de Versões (xslt):

• Edição Diplomática
• Edição Semi-Interpretativa
• Edição Interpretativa
• Edição Interpretativa para ferramentas automáticas
• Glossário de Interferências Editoriais</td></tr>
<tr><td colspan="3">Acesso externo via Catálogo Eletrônico</td></tr>
</table>

Notas da Seção

1Para saber mais sobre codificação de caracteres, consulte:
http://www.unicode.org/ (home page do sistema Unicode)
http://www.w3.org/TR/html4/charset.html (padronização de codificações para html)
http://www.cs.tut.fi/~jkorpela/chars.html (tutorial)
http://en.wikipedia.org/wiki/Character_encoding (resumo enciclopédico)
http://www.unicode.org/charts/PDF/U0000.pdf (lista completa unicode)

2Ou seja, o desenho ou aparência do sinal gráfico pode ser distinto a depender dos conjuntos de tipos/fontes disponíveis em cada computador. No exemplo com o código **&#0065** acima, o grafema [A] apresenta-se no tipo "Gothic"; poderia se apresentar no tipo "Arial" (como em **A**) ou "Times New Roman" (como em A) – e ainda, em diferentes tamanhos. Em todos os casos, o código de caracter é o mesmo.

3Se, ainda, desejarmos imprimir o texto eletrônico produzido, precisamos de uma outra etapa de mediação: o programa que transforma a codificação em imagens impressas (ou seja, que organiza o "texto" como pontos de tinta a serem gravados no papel).

4 Isso é realizado por um comando que ativa uma transformação via XLST, a linguagem de programação que atua sobre os documentos anotados com XML. Os arquivos XLST com as programações podem ser armazenados em um servidor, e o usuário final pode ativá-los através de comandos remotos, sem precisar manipular a computação.

I. Catalogação

I.1 Diretrizes

O sistema de catalogação e ordenação, buscas, e geração de documentos no Corpus é chamado de Catálogo Dinâmico.

A catalogação dos textos é realizada com base nas anotações de informações externas sobre o texto em seu cabeçalho: nome do autor, título, gênero, editores, etc. As categorias assim formadas servem a dois funcionamentos:

1. Formação de Fichas Catalográficas ou "Portais"

As metainformações listadas nos arquivos xml são a base das fichas catalográficas de cada documento, onde se informam os dados básicos sobre o texto original, sobre a edição utilizada no processamento para o corpus, e sobre o processamento realizado para o corpus.

No atual sistema, a apresentação das fichas catalográficas funciona como um portal para cada texto. Cada"portal" é na realidade o próprio arquivo XML do texto. A este arquivo foi adicionada uma ligação a uma programação que obriga o navegador a apresentar o documento XML sob esta forma de "portal", mostrando as versões disponíveis para o texto em questão. Cada uma das opções oferecidas no portal leva à ativação da geração de diferentes **versões** ou **formas de apresentação** do texto.

2. Alimentação de um sistema de buscas por metainformações.

As metainformações servem de base para buscas com o objetivo de selecionar documentos de determinado perfil (quanto à periodização, tipo de texto, tipo de processamento, etc.). Essas buscas podem ser programadas na linguagem X-Query, que, da mesma forma como a linguagem XSLT, baseia-se na estrutura arbórea da anotação XML dos documentos. A busca em X-Query lê os cabeçalhos dos documentos listados num arquivo central (catalog.xml) cujos elementos são os endereços de cada documento XML armazenados no servidor. Com base neste arquivo, uma programação X-Query (**catalog.xq**) pode ler as estruturas incluidas nos cabeçalhos dos documentos previstos, e pode formar listas re-ordenadas por diferentes critérios.

Os procedimentos para esses funcionamentos estão expostos na seção I.2 abaixo:

- A seção I.2.1 abaixo mostra a estrutura do arquivo de catálogo.
- Na seção I.2.2 abaixo estão resumidos os procedimentos de estruturação dos cabeçalhos em cada documento.
- Na seção I.2.3 resumem-se os procedimentos necessários para a ativação do catálogo (de modo local ou remoto).

A seção I.3, em seguida, apresenta a organização da documentação interna necessária para administrar o Corpus.

I.2 Procedimentos

I.2.1 Arquivo Básico de Catalogação

O arquivo **catalog.xml** representa o catálogo básico do Corpus; é nele que as modificações no inventário geral devem ser registradas (adição de textos, retirada de textos, modificações nos nomes dos textos, etc). O arquivo está gravado em <http://www.ime.usp.br/~tycho/cgi-bin/catalog.xml>; seu estado atual é reproduzido abaixo:

```
<catalog>
<corpusfile filename="a_001.xml"/>
<corpusfile filename="a_002.xml"/>
<corpusfile filename="a_003.xml"/>
<corpusfile filename="a_004.xml"/>
<corpusfile filename="a_005.xml"/>
<corpusfile filename="b_001.xml"/>
<corpusfile filename="b_002.xml"/>
<corpusfile filename="b_003.xml"/>
<corpusfile filename="b_005.xml"/>
<corpusfile filename="b_006.xml"/>
<corpusfile filename="b_007.xml"/>
<corpusfile filename="b_008.xml"/>
<corpusfile filename="c_001.xml"/>
<corpusfile filename="c_002.xml"/>
<corpusfile filename="c_003.xml"/>
<corpusfile filename="c_004.xml"/>
<corpusfile filename="c_005.xml"/>
<corpusfile filename="c_006.xml"/>
<corpusfile filename="c_007.xml"/>
<corpusfile filename="d_001.xml"/>
<corpusfile filename="e_001.xml"/>
<corpusfile filename="f_001.xml"/>
<corpusfile filename="g_001.xml"/>
<corpusfile filename="g_002.xml"/>
<corpusfile filename="g_003.xml"/>
<corpusfile filename="g_004.xml"/>
<corpusfile filename="g_005.xml"/>
<corpusfile filename="g_006.xml"/>
<corpusfile filename="g_008.xml"/>
<corpusfile filename="g_009.xml"/>
<corpusfile filename="h_001.xml"/>
<corpusfile filename="l_001.xml"/>
<corpusfile filename="l_002.xml"/>
<corpusfile filename="m_001.xml"/>
<corpusfile filename="m_002.xml"/>
<corpusfile filename="m_003.xml"/>
<corpusfile filename="m_004.xml"/>
<corpusfile filename="o_001.xml"/>
<corpusfile filename="p_001.xml"/>
<corpusfile filename="p_002.xml"/>
<corpusfile filename="s_001.xml"/>
<corpusfile filename="v_001.xml"/>
<corpusfile filename="v_002.xml"/>
<corpusfile filename="v_003.xml"/>
<corpusfile filename="v_004.xml"/>
</catalog>
```

I.2.2 Anotação dos metadados

Cada documento .xml é composto de duas partes principais: <head> e <body>; <body> codifica o "corpo" do documento, incluindo o texto e todos os seus elementos; <head> codifica as informações sobre o documento, ou seja, as meta-informações, codificadas como <metadata>.

O sistema para anotação das estruturas de <metadata> em uso no Corpus está estruturado em **camadas de informações** (atributo "generation"), para dar conta das sucessivas etapas envolvidas nas diversas edições dos textos.

Em cada geração, há elementos <meta>, compostos de pares de nomes e valores <name>, <value>. Há uma lista fechada para os nomes; os valores podem ser abertos ou fechados.

- Em I.2.2.1 a seguir há um quadro-resumo do esquema de anotação de cabeçalhos, seguido de uma lista dos pares <name>, <value>.

- Em I.2.2.2 a seguir pode-se ver um exemplo de cabeçalho completo, para o texto "Reflexão sobre a Vaidade dos Homens" (Matias Aires), a_001.xml; e em seguida, c omo as informações do cabeçalho do texto a_001.xml formam um portal e uma ficha catalográfica.

I.2.2.1 Esquema da anotação de cabeçalhos

A. Estrutura Geral

```
<head>
```

Metadata, Geração "Corpus Processing"

Para anotar as informações relativas ao processamento para inclusão no Corpus, e outros indicadores técnicos internamente relevantes.

```
<metadata generation="corpus_processing">

<meta> <name>Corpus Processing Type</name>          <val>...</val> </meta>
 <meta> <name>Corpus Processing Edition Level</name> <val>...</val> </meta>
 <meta> <name>Document Name</name>                   <val>...</val> </meta>
 <meta> <name>Document Title</name>                  <val>...</val> </meta>
 <meta> <name>Document Author ID</name>              <val>...</val> </meta>
 <meta> <name>Period by Birthdate</name>             <val>...</val> </meta>
 <meta> <name>Word Count</name>                      <val>...</val> </meta>
 <meta> <name>First Processing Information</name>    <val>...</val> </meta>
 <meta> <name>Last Revised by</name>                 <val>...</val> </meta>
 <meta> <name>XML Markup by</name>                   <val>...</val> </meta>
 <meta> <name>Last Saved Date</name>                 <val>...</val> </meta>
 <meta> <name>Inserted in Catalog by</name>          <val>...</val> </meta>
 <meta> <name>POS</name>                             <val>...</val> </meta>
 <meta> <name>PSD</name>                             <val>...</val> </meta>
 <meta> <name>POS Tagged Version</name>              <val>...</val> </meta>
 <meta> <name>Syntactic Annotation Version</name>    <val>...</val> </meta>

</metadata>
```

Metadata, Geração "Original Text"

Para anotar as informações sobre o texto original, ordenadas em pares de nomes e valores.

```
<metadata generation="original_text">

<meta> <name>Author Name</name>             <val>...</val> </meta>
 <meta> <name>Author Year of Birth</name>    <val>...</val> </meta>
 <meta> <name>Original Text Title</name>     <val>...</val> </meta>
 <meta> <name>Original Text Editor</name>    <val>...</val> </meta>
 <meta> <name>Original Text Date</name>      <val>...</val> </meta>
 <meta> <name>Original Text Reference</name> <val>...</val> </meta>
 <meta> <name>Genre</name>                   <val>...</val> </meta>

</metadata>
```

Metadata, Geração "Immediate Source"

Para anotar as informações sobre a edição-fonte utilizada na transcrição

Essa camada só precisa ser utilizada quando a Fonte Imediata não é a edição original do texto.

```
<metadata generation="immediate_source">

 <meta> <name>Immediate Source Type</name>          <val>...</val> </meta>
  <meta> <name>Immediate Source Edition Level</name> <val>...</val> </meta>
  <meta> <name>Immediate Source Title</name>         <val>...</val> </meta>
  <meta> <name>Immediate Source Rights</name>        <val>...</val> </meta>
  <meta> <name>Immediate Source Editor</name>        <val>...</val> </meta>
  <meta> <name>Immediate Source Date</name>          <val>...</val> </meta>
  <meta> <name>Immediate Source Reference</name>     <val>...</val> </meta>

</metadata>
```

Metadata, Geração "Intermediate Source"

Para anotar as informações sobre a edição intermediária.

Essa camada só precisa ser utilizada quando houver uma edição intermediária entre a Fonte Imediata e a edição original do texto.

```
<metadata generation="intermediate_source">

<meta> <name>Intermediate Source Type</name>          <val>...</val> </meta>
 <meta> <name>Intermediate Source Edition Level</name> <val>...</val> </meta>
 <meta> <name>Intermediate Source Title</name>         <val>...</val> </meta>
 <meta> <name>Intermediate Source Rights</name>        <val>...</val> </meta>
 <meta> <name>Intermediate Source Editor</name>        <val>...</val> </meta>
 <meta> <name>Intermediate Source Date</name>          <val>...</val> </meta>
 <meta> <name>Intermediate Source Reference</name>     <val>...</val> </meta>

</metadata>
```

```
<head>
```

B. Guia das Nomenclaturas

Notas: (1) Todas as categorias <name> são fechadas
(2) As categorias <val> podem ser abertas ou fechadas:
as categorias <val> fechadas são precedidas por (*) na tabela abaixo;
as categorias <val> abertas podem ter formatos específicos (indicados na tabela abaixo entre parênteses)
(3) As buscas baseadas em metadados são Case-sensitive
(portanto a digitação deve manter o uso de maiúsculas e minúsculas aqui indicado).

(i) Pares de `<name><val>` em `<meta>`, para `<metadata generation="corpus_processing">`

<name>	**<val>**	*Uso*
Document Name	(cf. Tabela de Nomeação abaixo)	Código do documento
Last Saved Date	(formato DD.MM.AAAA)	Data da última modificação
Document Title	(*sem especificação*)	Título do Texto
Document Author ID	(formato: Último Sobrenome, Nome - ex. Aires, Matias)	Nome do Autor para Catálogo
Word Count	(sem especificação)	número de palavras do texto
Digitalized by	(formato: Inicial. Sobrenome, separados por vírgulas)	Nomes dos Digitadores
XML Markup by	(formato: Inicial. Sobrenome, separados por vírgulas)	Nomes dos Anotadores
Inserted in Catalog by	(formato: Inicial. Sobrenome, separados por vírgulas)	Nomes de quem enviou o arquivo para o servidor e atualizou o arquivo corpus.xml
POS Tagged Version	(*sem especificação*)	URL do arquivo etiquertado, onde se aplica
Syntactic Annotation Version	(*sem especificação*)	URL do arquivo parseado, onde se aplica
First Processing Information	(formato: Inicial. Sobrenome, separados por vírgulas)	Nomes dos preparadores da Fase I, onde se aplica
Corpus Processing Status	(*) Final Release | Revision in Progress | Modernization in Progress	Estado de Preparação dos textos: Final Release - prontos; Revision in Progress - em revisão ; Modernization in Progress - em edição
Corpus Processing Type	(*) generation 1 | generation 2 | generation 3 | generation 4	Tipo de processamento no corpus quanto ao número de edições anteriores:1 - digitalização de manuscritos originais; 2 - digitalização de impressos originais; 3 - digitalização de edição intermediária ; 4 -digitalização de edição de uma edição intermediária
Corpus Processing Edition Level	(*) Complete | Technical	Tipo de processamento quanto à profundidade da edição para o Corpus: Complete - modernização integral; Technical - modernização parcial (textos da Fase I).
Period by Birthdate	(*) 1300-1349 | 1350-1399 | 1400-1449 | 1450-1499 | 1500-1549 | 1550-1599 | 1600-1649 | 1650-1699 | 1700-1749 | 1750-1799 | 1800-1849 | 1850-1899	Período de nascimento do autor (50 anos)
POS	(*) yes | no	Texto etiquetado morfologicamente
PSD	(*) yes | no	Texto analisado sintaticamente

(ii) Pares de <name><val> em <meta>, para <metadata generation="original_text">

<name>	**<val>**	***Uso***
Author Name	(formato: Nome Sobrenome)	Nome do autor
Author Year of Birth	(*sem especificação*)	Ano de nascimento do autor
Original Text Title	(*sem especificação*)	Título original da obra
Original Text Editor	(*sem especificação*)	Editor original da obra
Original Text Date	(*sem especificação*)	Data da edição original
Original Text Reference	(ABNT)	Referência Completa da edição original
Genre	(*) Letters | Narrative | Dissertation | Theatre | diverse	Gênero do texto

(iii) Pares de <name><val> em <meta>, para <metadata generation="immediate_source">

<name>	**<val>**	***Uso***
Immediate Source Type	(*) Original manuscript | Original manuscript, fac-simile | Original print | Original print, fac-simile | Transcript of original print | Transcript of original manuscript	Tipo de texto utilizado como Fonte Imediata, quanto ao estágio de versão
Immediate Source Edition Level	(*) Preserved ortography | Edited ortography	Profundidade da Edição utilizada como Fonte Imediata
Immediate Source Title	(*sem especificação*)	Título Completo da Fonte Imediata
Immediate Source Rights	(*sem especificação*)	Detentor dos Direitos Legais da Edição utilizada como Fonte Imediata
Immediate Source Editor	(*sem especificação*)	Nome do Editor da Fonte Imediata
Immediate Source Date	(*sem especificação*)	Data da Fonte Imediata
Immediate Source Reference	(ABNT)	Referência Completa da Fonte Imediata
Immediate Source URL	(*sem especificação*)	Endereço web da Fonte Imediata, se houver

(iv) Pares de <name><val> em <meta>, para <metadata generation="intermediate_source">

<name>	**<val>**	***Uso***
Intermediate Source Type	(*) Transcript of transcript of original print | Transcript of transcript of original manuscript | Transcript of original print | Transcript of original manuscript	Tipo de texto utilizado como Fonte Intermediária, quanto ao estágio de versão
Intermediate Source Edition Level	(*) Preserved ortography | Edited ortography	Profundidade da Edição utilizada como Fonte Intermediária
Intermediate Source Title	(*sem especificação*)	Título Completo da Fonte Intermediária
Intermediate Source Rights	(*sem especificação*)	Detentor dos Direitos Legais da Edição utilizada como Fonte Intermediária
Intermediate Source Editor	(*sem especificação*)	Nome do Editor da Fonte Intermediária
Intermediate Source Date	(*sem especificação*)	Data da Fonte Intermediária
Intermediate Source Reference	(ABNT)	Referência Completa da Fonte Intermediária

I.2.2.2 Amostra: Cabeçalho e Ficha catalográfica de um texto do Corpus

cf. <http://www.ime.usp.br/~tycho/corpus/texts/xml/a_001.xml>:

A. Cabeçalho

```
<head id="a_001" >
 <metadata generation="corpus_processing">
   <meta> <name>Corpus Processing Type</name>          <val>generation 3</val> </meta>
   <meta> <name>Corpus Processing Edition Level</name> <val>Technical edition
only</val> </meta>
   <meta> <name>Document Name</name>                   <val>a_001</val> </meta>
   <meta> <name>Document Title</name>                  <val>Reflexões sobre a Vaidade dos
Homens</val> </meta>
   <meta> <name>Document Author ID</name>              <val>Aires, Matias</val> </meta>
   <meta> <name>Period by Birthdate</name>             <val>1700-1749</val> </meta>
   <meta> <name>Word Count</name>                      <val>56.479</val> </meta>
   <meta> <name>First Processing Information </name>   <val>Digitalized by P. Abdo, T.
Menegatti, C. Namiuti; First Revised by P. Abdo, T. Menegatti, C. Namiuti; Last Revised
by P. Abdo, T. Menegatti, C. Namiuti; First Uploaded by P. Abdo, V. Vinha, L. Chociay;
First Uploaded Date 21.01.2002;</val> </meta>
   <meta> <name>XML Markup by</name>                   <val>M.C. Paixão de Sousa, D.
Störbeck, T. Trippel</val> </meta>
   <meta> <name>Last Saved Date</name>                 <val>24.02.2005</val> </meta>
   <meta> <name>Inserted in Catalog by</name>          <val>M.C. Paixão de Sousa</val>
 </meta>
   <meta> <name>POS</name>                             <val>yes</val>  </meta>
   <meta> <name>PSD</name>                             <val>yes</val>  </meta>
   <meta> <name>POS Tagged Version</name>
<val>http://www.ime.usp.br/~tycho/corpus/texts/pos/txt/aires_00_tcc.txt</val>    </meta>
   <meta> <name>Syntactic Annotation Version</name>
<val>http://www.ime.usp.br/~tycho/corpus/texts/psd/txt/aires_02_psd.txt</val>    </meta>
 </metadata>
 <metadata generation="original_text">
   <meta> <name>Author Name</name>          <val>Matias Aires</val> </meta>
   <meta> <name>Author Year of Birth</name> <val>1705</val> </meta>
   <meta> <name>Original Title</name>       <val>Reflexão sobre a Vaidade dos
Homens</val> </meta>
   <meta> <name>Genre</name>                <val>Dissertation</val> </meta>
 </metadata>
 <metadata generation="immediate_source">
   <meta> <name>Immediate Source Type</name>           <val>generation 4 </val>   </meta>
   <meta> <name>Immediate Source Edition Level</name>  <val>Edited ortography</val>
 </meta>
   <meta> <name>Immediate Source Title</name>          <val>Reflexões sobre a Vaidade dos
Homens e Carta sobre a Fortuna</val>  </meta>
   <meta> <name>Immediate Source Rights</name>         <val>Imprensa Nacional - Casa da
Moeda</val>  </meta>
   <meta> <name>Immediate Source Editor</name>         <val>Jacinto do Prado Coelho</val>
 </meta>
   <meta> <name>Immediate Source Date</name>           <val>1980</val>   </meta>
   <meta> <name>Immediate Source Reference</name>      <val>AIRES, Matias. Reflexões
sobre a Vaidade dos Homens e Carta sobre a Fortuna   (seleção, prefácio e notas por
Jacinto do Prado Coelho e Violeta Crespo Figueiredo). Lisboa, Imprensa Nacional - Casa da
Moeda. 1980.</val>   </meta>
 </metadata>
</head>
```

B. Ficha catalográfica

A Apresentação em forma de ficha catalográfica, cf. abaixo, é obtida pela aplicação da folha **catalog_files.xsl** sobre os arquivos xml do corpus. A folha está armazenada no mesmo diretório dos

arquivos de textos:

<http://www.ime.usp.br/~tycho/corpus/texts/xml/catalog_files.xsl>

No servidor, todos os arquivos xml estão gravados com a instrução de serem apresentados automaticamente desta forma. Esta instrução se obtém adicionando-se a seguinte linha em seguida à declaração de xml do documento:

```
<?xml-stylesheet href="catalog_files_2006.xsl" type="text/xsl"?>
```

Aires, Matias (1705): *Reflexões sobre a Vaidade dos Homens*

Versões para Leitura: Transcrição do texto-fonte [gerar]

Versões para Trabalho:
- Edição Técnica, Texto Simples (.txt)
- Anotação Morfológica
- Anotação Sintática

XML Para ver o xml de base, clique em *"exibir" >"código-fonte"* na barra superior do navegador

Ficha Catalográfica

Edição Transcrita:

AIRES, Matias. Reflexões sobre a Vaidade dos Homens e Carta sobre a Fortuna (seleção, prefácio e notas por Jacinto do Prado Coelho e Violeta Crespo Figueiredo). Lisboa, Imprensa Nacional - Casa da Moeda. 1980.

Informações sobre a obra original:

Author Name	Matias Aires
Author Year of Birth	1705
Original Title	Reflexão sobre a Vaidade dos Homens
Genre	Dissertation

Informações sobre a edição utilizada:

Immediate Source Type	Transcript of original print
Immediate Source Edition Level	Edited ortography
Immediate Source Title	Reflexões sobre a Vaidade dos Homens e Carta sobre a Fortuna
Immediate Source Rights	Imprensa Nacional - Casa da Moeda
Immediate Source Editor	Jacinto do Prado Coelho
Immediate Source Date	1980
Immediate Source Reference	AIRES, Matias. Reflexões sobre a Vaidade dos Homens e Carta sobre a Fortuna (seleção, prefácio e notas por Jacinto do Prado Coelho e Violeta Crespo Figueiredo). Lisboa. Imprensa Nacional - Casa da Moeda. 1980.

Informações sobre o processamento no corpus:

Document Name	a_001
Last Saved Date	31.10.2006
Corpus Processing Status	Final Release
Corpus Processing Type	generation 4
Corpus Processing Edition Level	Technical
Document Title	Reflexões sobre a Vaidade dos Homens
Document Author ID	Aires, Matias
Period by Birthdate	1700-1749
Word Count	56.479
Digitalized by	P. Abdo, T. Menegatti, C. Namiuti
XML Markup by	M.C. Paixão de Sousa, D. Störbeck, T. Trippel, A.C. Garcia de Souza
Inserted in Catalog by	M.C. Paixão de Sousa
POS	yes
PSD	yes
POS Tagged Version	http://www.ime.usp.br/~tycho/corpus/texts/pos/txt/aires_00_tcc.txt
Syntactic Annotation Version	http://www.ime.usp.br/~tycho/corpus/texts/psd/txt/aires_02_psd.txt
First Processing Information	Revised by P. Abdo, T. Menegatti, C. Namiuti; Revised by P. Abdo, T. Menegatti, C. Namiuti; Verificomment by ; First Uploaded by P. Abdo, V. Vinha, L. Chociay; First Uploaded in 21.01.2002;

I.2.3 Formação do Catálogo

O Catálogo Eletrônico, ou Catálogo Dinâmico, é criado por meio de uma programação em linguagem X-Query.

Um manual completo da linguagem X-Q pode ser consultado em: http://www.w3schools.com/xquery/default.asp

A programação percorre os cabeçalhos (I.2.2) de todos os textos do Corpus (por intermédio do arquivo que indica todos os textos a serem pesquisados e seu local de armazenamento, catalog.xml, cf. I.2.1), e organiza as informações ali contidas na forma de uma página de entrada.

Assim, o Catálogo funciona, para o usuário, como portal de acesso geral ao Corpus. O usuário pode selecionar os textos diretamente por autor ou título; ou acessar listas que agrupam os textos por data, gênero, etc.

Quando se seleciona o texto que deseja visualizar, o sistema leva a um "portal" preparado para cada texto, como em B acima.

O arquivo de programação para o Catálogo é:

http://www.ime.usp.br/~tycho/cgi-bin/catalog.xq

Fundametalmente, a programação X-Query em catalog.xq "percorre" o Corpus, selecionando e organizando as estruturas de metadata conforme programado. Por exemplo, para gerar uma lista dos documentos em ordem alfabética, a programação tem o seguinte código:

trecho do arquivo catalog.xq :

```
{let $corpuslocation := "../corpus/texts/xml/"
    for $filename in
doc(concat($corpuslocation,'catalog.xml'))/catalog/corpusfile/@filename
    order by string(for $file in doc(concat($corpuslocation,$filename))
    return
               for $metadata in $file/document/head/generation/meta
               where $metadata/name="Document Author ID"
               return $metadata/val/text()) ascending
    return
              for $file in doc(concat($corpuslocation,$filename))
              return
                     for $generation in $file/document/head/generation
                     return
                           for $metadata in $generation/meta
                           where $metadata/name="Document Author ID"
                           return
                                 {
                                  for $tb_title in $metadata/../../generation/meta
                                  where $tb_title/name="Document Name"
                                  return
                                         [{$tb_title/val/text()}]
                                 }
                                 {
                                  {$metadata/val/text()}
                                 }
                                 {
                                  for $text_title in $metadata/../../generation/meta
                                  where $text_title/name="Document Title"
                                  return
                                         {$text_title/val/text()}
                                 }
        }
```

O que se traduz:

"Tomando o universo de elementos listados no arquivo ../corpus/texts/xml/catalog.xml, escolha em cada arquivo os elementos incluidos no caminho /document/head/generation/meta nos quais o elemento name tiver o texto Document Author ID"; para cada arquivo, gere uma lista ordenada contendo esse elemento Document Author ID; retorne para cada elemento assim definido e ordenado o código do documento e o seu título".

A programação de busca pode ser ativada de duas formas: local e remota, cf. abaixo.

I.2.3.1 Ativação do Catálogo – via local

- A ativação local pode ser realizada em uma máquina que contenha os seguintes softwares:
  - saxon (versão 8 ou superior; software livre, disponível em <http://www.saxon.net>)
  - java (4.2; idem)
- A ativação no terminal Unix ou Dos segue o seguintes passos:
  1. chamar `java` <+espaço>
  2. chamar modo `-cp` <+espaço>
  3. indicar o software de transformação (Saxon) – incluir caminho na estrutura de diretórios;
  4. digitar o comando para transformação: `net.sf.saxon.Query` <+espaço>
  5. indicar o nome do documento de query – no caso, sempre `catalog.xq` (incluir o caminho), e digitar parêntese angular: >
  6. indicar o nome do arquivo a ser gerado – incluir caminho onde ele será gravado, e terminação (.hml), <+enter>
  7. O arquivo gerado estará disponível no diretório que foi indicado em (f).

- Esquema
  (para uma máquina windows onde o Saxon está gravado na raiz (c:); e catalog.xq está num diretório cgi-bin local:

```
c:\>java -cp c:\saxonb8-3\saxon8.jar net.sf.saxon.Query c:\~tycho\cgi-
bin\catalog.xq > c:\~tycho\corpus\texts\xml\catalog.html
```

I.2.3.2 Ativação do Catálogo – via remota

O acesso remoto ao catálogo é feito pela ativação de scripts via hipertexto.

Na página de entrada do catálogo gerado conforme os procedimentos acima, o usuário tem acesso à seguinte tela:

Atualização

Este catálogo foi gerado em 23.05.07.
Clique [aqui] para atualizá-lo no servidor
(isso levará alguns instantes)

Ao clicar onde indicado, o usuário irá ativar o script em http://www.ime.usp.br/~tycho/cgi-bin/catalog.pl, que ativa a geração do catálogo, da mesma forma do que acontece por via local. É gerado um documento temporário, que é a interface do Catálogo Dinâmico atualizado pelo usuário.

I.3 Documentação (Catalogação Interna)

Esta seção apresenta as informações básicas sobre a arquitetura de diretórios e o sistema de identificação de arquivos do Corpus. Foram criados diretórios para os novos tipos de documentos, e os alguns diretórios existentes foram renomeados (cf. I.3.1 para a lista dos diretórios e arquivos relevantes). Além disso, para facilitar a gestão dos arquivos (em especial tendo em vista a geração de múltiplas versões), um novo sistema de nomeação foi introduzido (cf. lista em I.3.2).

Incluem-se por fim informações sobre a documentação de anotação (cf. I.3.3); estão ali reproduções da gramática de anotações (documento .dtd, cf. I.3.3.1) e do documento .xml modelo a ser usado como ponto de partida das transcrições (cf. I.3.3.2) .

I.3.1 Estruturas de Diretório

				Notas
`~tycho/`	`cgi-bin/`			Contém: - os arquivos de programação XSLT e XQ, - o documento XML que une o catálogo, - os scripts de acesso ao catálogo (.pl) e os softwares (saxon).
	`corpus/`	`obras/`		Contém os arquivos atualmente em transcrição (backup semanal)
		`style/`		Contém as folhas de estilo CSS que organizam a aparência/formatação dos arquivos e versões geradas
		`docs/`		Contém cópias das folhas de programação XSLT e XQ que organizam o sistema de apresentação e busca dos documentos xml (para melhor acesso externo)
		`texts/`	`orto/`	Contém as transcrições dos textos na versão da Fase I
			`pos/`	Contém os arquivos morfologicamente etiquetados
			`psd/`	Contém os arquivos sintaticamente anotados
			`xml/`	Contém os documentos-base .xml, a gramática (dtd) e os modelos.

Nota: Apenas o diretório `texts` e seus conteúdos exigem senha para acesso externo.

I.3.2 Tabela de nomeação dos arquivos

autor	*ID*
(1705-1763) MATIAS AIRES	a_001
(1597-1655) MANUEL PIRES DE ALMEIDA	a_002
(1802-1881) MARQUES DE F. E ALORNA	a_003
(1750-1839) MARQUESA D'ALORNA	a_004
(1676-1749) JERONIMO CONTADOR DE ARGOTE	a_005
(1675-1754) ANDRE DE BARROS	b_001
(1497-1562) JOAO DE BARROS	b_002
(1644-1710) MANUEL BERNARDES	b_003
(1825-1890) CAMILO CASTELO BRANCO	b_005
(1584-1637) ANTONIO BRANDAO	b_006
(1569-1617) BERNARDO DE BRITO	b_007
(1651-1735) JOSE DA CUNHA BROCHADO	b_008
(1702-1783) CAVALEIRO DE OLIVEIRA	c_001
(1658-1753) MARIA DO CEU	c_002
(1631-1682) ANTONIO DAS CHAGAS	c_003
(1714- ?) ANTONIO DA COSTA	c_004
(1757-1832) JOSE DANIEL RODRIGUES DA COSTA	c_005
(1601-1667) MANUEL DA COSTA	c_006
(1542-1606) DIOGO DO COUTO	c_007
(1502-1557) D. JOAO III	d_001
(1845-1900) ECA DE QUEIROZ & O. MARTINS	e_001
(1583-1655) MANUEL SEVERIM DE FARIA	f_001
(1597-1665) MANUEL DE GALHEGOS	g_001
(1724-1772) CORREIA GARCAO	g_002
(1799-1854) ALMEIDA GARRETT – cartas	g_003
(1799-1854) ALMEIDA GARRETT – teatro	g_004
(1799-1854) ALMEIDA GARRETT – Viagens	g_005
(1695-?) ALEXANDRE DE GUSMAO	g_006
(1502-1579) PERO MAGALHÃES DE GANDAVO	g_008
(1435-1517) DUARTE GALVÃO	g_009
(1517-1584) FRANCISCO DE HOLANDA	h_001
(1579-1621) FRANCISCO RODRIGUES LOBO	l_001
(1380-1460) FERNÃO LOPES	l_002
(1733-1805) DIOGO I. DE PINA MANIQUE	m_001
(1608-1666) FCO MANUEL DE MELO – Cartas Familiares	m_003
(1608-1666) FCO MANUEL DE MELO – Tácito Português	m_004
(1836-1915) RAMALHO ORTIGAO	o_001
(1510-1583) FERNAO MENDES PINTO	p_001
(1440-1522) RUI DE PINA	p_002
(1556-1632) LUIS DE SOUSA	s_001
(1672-1760) MANUEL DOS SANTOS	s_002
(1713-1792) LUIZ ANTONIO VERNEY	v_001
(1608-1697) ANTONIO VIEIRA – Cartas	v_002
(1608-1697) ANTONIO VIEIRA – Historia do Futuro	v_003
(1608-1697) ANTONIO VIEIRA – Sermões	v_004

I.3.3 Documentação de Anotação

I.3.3.1 Gramática da anotação XML no Corpus (DTD)

cf. <http://www.ime.usp.br/~tycho/corpus/texts/xml/tycho_03.dtd-->

```
<!ELEMENT    document    (head, body)>

<!ELEMENT    head        (metadata)+>
<!ELEMENT    metadata    (meta)+>
<!ELEMENT    meta        (name, val)+>
<!ELEMENT    name        (#PCDATA)>
<!ELEMENT    val         (#PCDATA)>

<!ELEMENT    body        (text)+>
<!ELEMENT    text        (sec)+>
<!ELEMENT    sec         (p | s | sec | #PCDATA | #EMPTY)+>
<!ELEMENT    p           (s | text_el | (comment*) | #EMPTY)+>
<!ELEMENT    s           (text_el | format | ed_mark | (comment*) | #PCDATA | #EMPTY)+>
<!ELEMENT    text_el     (#PCDATA)>
<!ELEMENT    format      (#PCDATA)>
<!ELEMENT    comment     (#PCDATA)>
<!ELEMENT    ed_mark     (ed | or | (ad?))>
<!ELEMENT    ed          (#PCDATA)>
<!ELEMENT    ad          (#PCDATA)>
<!ELEMENT    or          (#PCDATA)>

<!ATTLIST    metadata    generation    (original_text | immediate_source | intermediate_source |
                                       corpus_processing) #REQUIRED >
<!ATTLIST    s           t             (signature | date | heading)>
<!ATTLIST    sec         t             (ch | prologue | preface | licence | index | letter | play |
                                       act | char | sc_desc | char_desc | scene_marking |
                                       table | line | pag | col | margin |
                                       title | subtitle | language) #REQUIRED >
<!ATTLIST    text_el     t             (p_nr | line_nr | pag_nr | pag_header | pag_foot | image)>
<!ATTLIST    format      t             (cap | b | i)>
<!ATTLIST    ed          t             (mod | exp | seg | punc | cor | crux)"mod" >

<!--end dtd-->
<!--key for status indicators: -->
<!--number of occurence of a content-->
<!--"?"     content has to occure 0 or one time--> <!--"*"     any number of occurence (between 0 and n)-->
<!--"+"     at least one occurence (between 1 and n)--> <!--otherwise exactly once-->
<!--connectors-->
<!--"|"     or--> <!--","     sequence-->
```

II. Transcrição

II.1 Diretrizes

II.1.1Visão Geral

A transcrição dos textos é conservadora, mantendo todas as características grafemáticas, a organização espacial, a segmentação vocabular, e outras características do original..

II.1.2 Normas para a transcrição de originais

II.1.2.1 Caracteres alfabéticos:

- Transcrevem-se como caracteres romanos
- Reproduzem-se as diferenças de módulos do original:

N E S T E pequeno ſeruiço

- Reproduzem-se os alógrafos contextuais:

v/u: louuores, deue **ſ /s**: neſſas, ſuccedido, peſſoas **i/j**: D. Ioão

- Reproduz-se o uso de til do original:

bõs, cõ, algũa, homẽs, acrecẽte

- Registra-se o uso de capitulares:

<format t=" cap">R</format> E I N A N D O aquele mui caholico

- Registram-se as diferenças de tipo (itálico e negrito)

II.1.2.2 Abreviaturas, pontuação e organização espacial do texto:

- Transcrevem-se fielmente sinais gráficos especiais (como os caldeirões, ¶).
- Transcrevem-se fielmente as abreviaturas
- Transcreve-se fielmente a pontuação
- Reproduz-se a separação inter-linear
- Reproduz-se fielmente a separação vocabular, inclusive a inter-linear

Via preſente obra de Pero de Magalhães ("***Vi a*** *presente obra ...*") perpetualos ("*perpetuá-los*") E **iſtoaſsi** pelo mereci- <nl/> mẽto ("*E* ***isto assim*** *pelo mereci | mento*")

- Indicam-se os limites das sentenças gráficas (separação por pontos finais)
- Reproduz-se fielmente a paragrafação
- Reproduz-se fielmente a organização em colunas e em seções (capítulos, etc) e sua titulação
- Indicam-se as mudanças de páginas
- Reproduz-se a numeração de páginas, os reclames de pé de página e os cabeçalhos das páginas
- Indicam-se os textos escritos nas margens

II.1.2.3 Indicações de Dificuldades de Leitura e Aspecto do suporte material

- Indicam-se vocábulos ou trechos de difícil leitura e os pontos de deterioração do suporte material

II.2 Procedimentos

II.2.1 Digitação

Os textos são digitados diretamente em formato XML (Extended Markup Language), no editor Emacs, um processador de software livre (GNU), que sustenta grandes arquivos e evita problemas de codificação de caracteres.

- O processador está disponível para download gratuito em <http://www.gnu.org/software/emacs/>
- Um Manual completo para seu uso está disponível em <http://www.gnu.org/software/emacs/manual/html_node/emacs/index.html>

Este aplicativo oferece um modo SGML, com sistematização para inserção de etiquetas de anotação XML, o que confere segurança e consistência na anotação. Ao se abrir o documento .xml, o modo SGML começa automaticamente.

II.2.1.1 Instruções Gerais para a Digitação:

1. Os caracteres são transcritos no conjunto UTF-8. Antes de abrir o documento, certifique-se de que o "language environment" está indicando UTF-8.
2. Nos casos de carateres não previstos nos teclados dos computadores atuais, digitam-se diretamente os códigos UTF correspondentes. A lista completa do sistema Unicode se encontra em <http://www.unicode.org/charts/>. Por exemplo:

Tam glorio**ſ**os progre**ſſ**os > Tam glorio**ſ**os progre**ſſ**os Ƃ - "Latin Small letter long S"

3. Comentários: Caso pareça necessário inserir algum comentário sobre o processo de anotação, a ser conferido pelo revisor, use o sistema de comentários para XML. Os comentráios devem se iniciar com um parênteses angular aberto, um ponto de exclamação e dois hífens; e terminar com dois hífens um parênteses angular fechado; ou seja:

```
<!--TEXTO DO COMENTÁRIO-->
```

4. A digitação inclui a anotação da "estrutura bruta" do texto, ou seja, de categorias como seção, títulos, parágrafos, sentenças, quebras de linha, e outros. Veja em II.2.2 a seguir as anotações para essas categorias.

II.2.1.2 Checagem no browser (*debug*)

Um documento XML está **bem-formado** quando a hierarquia das catagorias continentes está respeitada. O navegadores só conseguem ler documentos bem-formados. Ao longo da digitação, é necessário checar a formação do documento. Para checar os documentos:

1. Abra o documento .xml na janela de um navegador (Internet Explorer, Firefox) e (em outra janela), no Emacs; Caso haja algum erro na anotação, ele será apontado pelo navegador – ex.emplo (mensagem do Firefox):

XML Parsing Error: mismatched tag. Expected: </text_el>. Location: l_002.xml Line Number 4226, Column 22:<text_el t="pag_foot">to temos</p>

2. No Emacs, vá até a linha indicada (menu: "go to line...") e corrija o erro;
3. Volte à janela do browser e recarregue o documento;
4. Prossiga assim até que o browser pare de apontar erros e leia a estrutura inteira, até </document>.

II.2.2 Anotações de Estruturas de Texto (*gross structure*)

II.2.2.1 Anotação de Formatação

Ocorrência	*Anotação*
capitulares	`<format t="cap"></format>`
itálico	`<format t="i"></format>`
negrito	`<format t="b"></format>`

II.2.2.2 Anotação de Divisões do Texto

parágrafo	`<p></p>`
sentença	`<s></s>`
quebra de linha	`<sec t="line"/>`
quebra de página	`<sec t="pag"/>`
quebra de coluna	`<sec t="col"/>`
capítulo	`<sec t="ch"></sec>`
prefácio	`<sec t="preface"></sec>`
prólogo	`<sec t="prologue"></sec>`
carta	`<sec t="letter"></sec>`
índice	`<sec t="index"></sec>`
peça (teatral)	`<sec t="play"></sec>`
ato	`<sec t="act"></sec>`
descrição dos personagens	`<sec t="char_desc"></sec>`
marcações de cena	`<sec t="scene_desc"></sec>`
nome dos personagens	`<sec t="char"></sec>`
título	`<sec t="title"></sec>`
tabela	`<sec t="table"></sec>`
texto na margem	`<sec t="margin"></sec>`

II.2.2.3 Anotação de Elementos do Texto

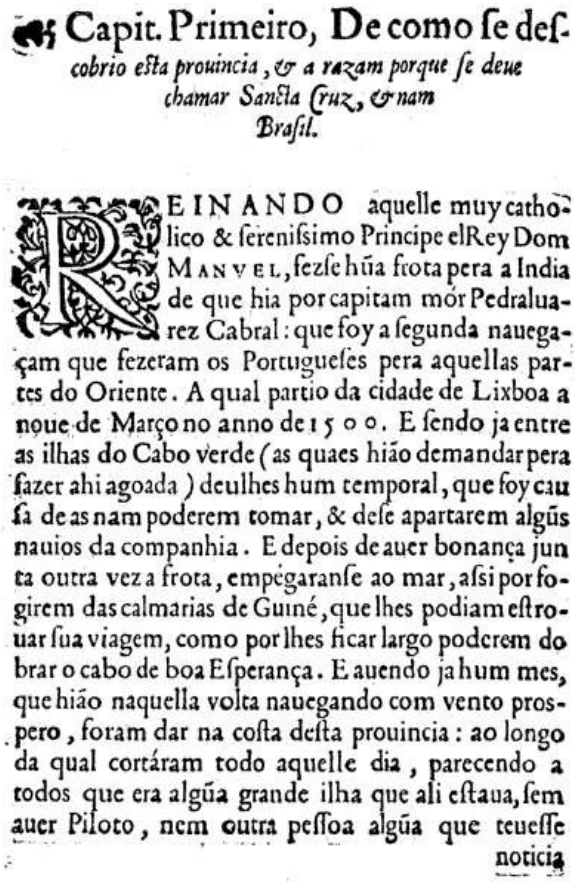
número da página	`<text_el t="pag_nr"></text_el>`
número da linha	`<text_el t="line_nr"></text_el>`
número do parágrafo	`<text_el t="par_nr"></text_el>`
cabeçalho da página	`<text_el t="pag_head"></text_el>`
reclame de pé de página	`<text_el t="pag_foot"></text_el>`
imagem	`<text_el t="image"></text_el>`

II.2.2.4 Anotação de Dificuldades de Leitura e outros Comentários

caracteres/ vocábulos / trechos de difícil leitura	`<ed_mark t="crux"></ed_mark>`
comentários em geral	`<comment></comment>`

II.2.3 Exemplo: Transcrição de um original impresso

II.2.3.1 O original impresso (fac-simile)



Capit. Primeiro, De como ſe deſ-
cobrio eſta prouincia, & a razam porque ſe deue chamar Sancta Cruz, & nam Braſil.

REINANDO aquelle muy catholico & ſereniſsimo Principe elRey Dom MANVEL, fezſe hũa frota pera a India de que hia por capitam mór Pedralua-rez Cabral: que foy a ſegunda nauegaçam que fezeram os Portugueſes pera aquellas partes do Oriente. A qual partio da cidade de Lixboa a noue de Março no anno de 1500. E ſendo ja entre as ilhas do Cabo verde (as quaes hião demandar pera fazer ahi agoada) deulhes hum temporal, que foy cauſa de as nam poderem tomar, & deſe apartarem algũs nauios da companhia. E depois de auer bonança junta outra vez a frota, empegaranſe ao mar, aſsi por fogirem das calmarias de Guiné, que lhes podiam eſtrouar ſua viagem, como por lhes ficar largo poderem dobrar o cabo de boa Eſperança. E auendo ja hum mes, que hião naquella volta nauegando com vento prospero, foram dar na coſta deſta prouincia: ao longo da qual cortáram todo aquelle dia, parecendo a todos que era algũa grande ilha que ali eſtaua, ſem auer Piloto, nem outra peſſoa algũa que teueſſe

noticia

II.2.3.2 A Transcrição

```
<sec t="ch">
<sec t="title">Capit. Primeiro, De como ſe deſ-<sec t="line"/>
<format t="i">cobrio esta prouincia , & a razam porque ſe deue <sec t="line"/>
chamar Sancta Cruz, e nam <sec t="line"/>
Braſil . </format> </sec>
<p>
<s><format t="cap">R</format>E I N A N D O aquelle muy catho- <sec t="line"/>
lico & ſereniſsimo Principe elRey Dom <sec t="line"/>
MANVEL , fezſe hũa frota pera a India <sec t="line"/>
de que hia por capitam mór Pedralua- <sec t="line"/>
rez Cabral: que foy a ſegunda nauega- <sec t="line"/>
çam que fezeram os Portugueſes pera aquellas par- <sec t="line"/>
tes do Oriente.</s> <s>A qual partio da cidade de Lixboa a <sec t="line"/>
noue de Março no anno de 1500.</s> <s>E ſendo ja entre <sec t="line"/>
as ilhas do Cabo verde (as quaes hião demandar pera <sec t="line"/>
fazer ahi agoada ) deulhes hum temporal, que foy cau <sec t="line"/>
ſa de as nam poderem tomar, & deſe apartarem algũs <sec t="line"/>
nauios da companhia.</s> <s>E depois de auer bonança jun <sec t="line"/>
ta outra vez a frota, empegaranſe ao mar, aſsi por fo- <sec t="line"/>
girem das calmarias de Guiné, que lhes podiam eſtro- <sec t="line"/>
uar sua viagem, como por lhes ficar largo poderem do <sec t="line"/>
brar o cabo de boa Eſperança.</s> <s>E auendo jahum mes, <sec t="line"/>
quehião naquella volta nauegando com vento pros-<sec t="line"/>
pero , foram dar na coſta deſta prouincia : ao longo <sec t="line"/>
da qual cortáram todo aquelle dia, parecendo a <sec t="line"/>
todos que era algũa grande ilha que ali eſtaua , ſem <sec t="line"/>
auer Piloto, nem outra peſſoa algũa que teueſſe <sec t="line"/>
<text_el t="pag_foot">noticia </text_el> <sec t="page/> ..... </s> ... </p> ...
</sec>
```

III. Edição

III.1 Diretrizes

III.1.1.Visão Geral

As interferências editoriais realizadas neste sistema têm por objetivo primordial tornar os textos legíveis para uma análise lingüística automática posterior.

Com este objetivo, é necessário normalizar a segmentação vocabular, desenvolver as abreviaturas, interpretar trechos de difícil leitura, e uniformizar a grafia dos textos, como em uma edição interpretativa tradicional.

As interferências editoriais são acrescentadas ao mesmo arquivo em que se realizou a transcrição de acordo com as diretrizes expostas na **Parte II**. São anotadas por sobre a transcrição do original, de modo a serem inteiramente recuperáveis na etapa da apresentação dos textos.

Assim, o sistema de edição aqui apresentado funciona como uma **anotação em camadas** sucessivas : ainda depois da aplicação de novas informações num texto de base, é possível distinguir as diferentes camadas do texto.

Desta forma, embora a edição seja realizada até o nível da modernização de grafia, é possível ter acesso ao texto mais conservador que permanece superposto às interferências.

Para que o sistema de camadas funcione de modo ótimo, as interferências são classificadas em categorias como *uniformização grafemática, expansão de abreviaturas, uniformização de pontuação, modernização de grafia, correção.*

Desta forma, nas etapas posteriores é possível ter acesso ao texto em diferentes versões, segundo a camada de edição escolhida. Por exemplo, é possível ter aceso ao uma versão com todas as uniformizações grafemáticas (ex.: alteração dos alógrafos contextuais u/v) , mas sem as modernizações de grafia, etc.

A partir da anotação das interferências editoriais, é possível se obter, para cada texto editado, um glossário das edições realizadas (separado, também, por categorias).

As interferências editoriais são identificadas com marcas em XML (cf. III.2) **.**

III.1.2. Normas de Edição

III.1.2.1 Caracteres alfabéticos

- Uniformizam-se as diferenças de módulos do original, reservando o uso das maiúsculas para: início de período; nomes próprios; vocábulos com valor afetivo (interpretação de nomes próprios):

N E S T E pequeno ſeruiço ► Neste pequeno serviço Tropheos ► Troféus

- Uniformiza-se o sistema grafemático do texto, inclusive no caso dos alógrafos contextuais*:*

v/u: louuores, deue ► louvores, deve **ſ /s**: neſſas, peſſoas, baſtante ► nessas, pessoas, bastante **i/j**: D. Ioão ► D. João

- Uniformiza-se o uso de til do original de acordo com as normas atuais:

bõs, cõ, algũa, homẽs, acrecẽte ► bons, com, alguma, homens, acrescente naõ, saõ ► não, são attençam ► atenção

- Normaliza-se a segmentação vocabular, intra-linear e inter-linear:

Via preſente obra de Pero de Magalhães ► **Vi a** presente obra ... E auendo **jahum** mes ► E havendo **já um** mês perpetualos ► perpetuá-los E iſtoaſsi pelo mereci- mẽto ► E isto assim pelo **merecimento**

III.1.2.2 Abreviaturas, pontuação e organização espacial do texto

- Desenvolvem-se as abreviaturas da transcrição original*:*

VM ► Vossa Mercê ~q ► que

- Uniformiza-se o uso da pontuação de acordo com as normas atuais – mas apenas nos aspectos técnicos, como por exemplo o uso de aspas seguido de ponto final. Mantém-se o uso de vírgulas, pontos finais, pontos de exclamação, etc.:

"*texto entre aspas***".** ► **"***texto entre aspas***."**

III.1.2.3 Grafia

- Moderniza-se a grafia do texto, de acordo com as normas atuais (Português do Brasil):

hum ► um
succedido ► sucedido
delle ► dele
Lixboa ► Lisboa
Tropheos ► Troféus
attençam ► atenção
mes ► mês

- **Casos Específicos:**

  Nos casos em que a norma ortográfica atual está relacionada a propriedades morfo-sintáticas, a normalização ortográfica não é realizada, a não ser que a forma original indique indubitavelmente a morfo-sintaxe (ainda que de outra forma que a usada na norma atual). Assim,

- não se normaliza o uso do acento circunflexo quando indicador de flexão do verbo no plural, a não ser que o contraste de flexão esta seja indicado de alguma forma no original:

eſſes peixes **tēm** eſcamas ► esses peixes **têm** escamas *mas*
eſſes peixes **tem** eſcamas ► esses peixes **tem** escamas

- não se normaliza o uso de acento grave para indicar a crase, a não ser que esta seja indicada de alguma forma no original:

ſatisfazer **á** menor parte ► satisfazer **à** menor parte *mas*
ſatisfazer **a** menor parte ► satisfazer **a** menor parte

III.1.2.4 Léxico e Morfosintaxe

- Mantém-se rigorosamente os arcaísmos lexicais e a morfologia e a sintaxe do texto original, salvo nos casos listados a seguir.
- **Casos Específicos**: Nos casos em que se considere que a fronteira entre a variação ortográfica e lexical/morfosintática é tênue, o vocábulo é modernizado; assim,

aſsi ► assim
pera ► para
fezeſſem ► fizessem
fezeram ► fizeram
emparo ► amparo
giolhos ► joelhos

III.1.2.5 Outras Interferências

- Corrigem-se, nesta etapa, os casos em que se interpreta ter havido um **evidente** lapso de escrita ou erro na impressão:

deqois ► depois
audam ► andam
romáram ► tomaram
rrato ► trato

III.2 Procedimentos

III.2.1 Anotação das Interferências de Edição

As interferências editoriais são identificadas com marcas em XML, anotadas segundo os mesmos procedimentos que as estruturas de texto (cf. II.2.1 acima) .

A etapa da edição equivale também a uma revisão da transcrição; o editor deve estar atento a lapsos de transcrição, e para isso deve poder acessar o texto original, conferindo-o regularmente. Deve-se, inclusive, lembrar de executar a checagem no navegador, da mesma forma como se explica em II.2.1.2 acima.

Além desses dois processos referentes à estrutura da anotação, ao final da etapa de edição é necessário ainda conferir a consistência das intervenções realizadas no texto:

- Gerar uma versão de Léxico para o documento (cf. seção IV para este procedimento);
- Examinar o glossário das suas edições, conferindo se estão de acordo com as Diretrizes de Edição em III.1 acima; corrigir eventuais inconsistências.

A idéia geral da anotação de edição é identificar todos os itens acrescentados ao texto pelo editor como elementos `<ed>`, e identificar os itens originais correspondentes como `<or>`; o par de itens editado/original é encapsulado como marca de editor, categoria `<ed_mark>`. Desta forma, guarda-se a correspondência entre os dois itens, e garante-se a recuperabilidade das formas originais dos textos. As marcas de edição são mais tarde numeradas por uma programação automática.

As diferentes formas de edição (normalização de grafia, segmentação, etc) são indicadas como atributos da categoria `<ed>`, como mostra a lista a seguir.

II.2.1.1 Anotação geral

	Anotação	*Exemplo*
item original a ser editado	`<or></or>`	`<or>hum</or>`
item acrescentado pelo editor	`<ed></ed>`	`<ed>um</ed>`

conjunto dos itens original e editado	`<ed_mark></ed_mark>`	`<ed_mark>`
		`<ed>um</ed>`
		`<or>hum</or>`
		`</ed_mark>`

II.2.1.2 Atributos para os tipos de interferências

	Atributo a <ed>	*Exemplo*

uniformização grafemática e de módulo	`t="gra"`	`<ed_mark>`
		`<ed t="gra">serviço</ed><or>ſeruiço</or>`
		`</ed_mark>`

separação ou junção de vocábulos	`t="seg"`	`<ed_mark>`
		`<ed t="seg">que ſe</ed><or>queſe</or>`
		`</ed_mark>`

expansão de abreviatura	`t="exp"`	`<ed_mark>`
		`<ed t="exp">Vossa Mercê</ed><or>V.M.</or>`
		`</ed_mark>`

uniformização de pontuação	`t="punc"`	`<ed_mark>`
		`<ed t="punc">` **>>** `</ed><or> " </or>`
		`</ed_mark>`

modernização de grafia	`t="mod"`	`<ed_mark>`
		`<ed t="mod">ínclita</ed><or>inclita</or>`
		`</ed_mark>`

correções	`t="cor"`	`<ed_mark>`
		`<ed t="cor">depois</ed><or>deqois</or>`
		`</ed_mark>`

NOTA: As edições podem se sobrepor, caso mais de um dos aspectos acima listados precisem ser alterados num item. Assim, um mesmo item original pode ser editado para uniformização de grafemas e em seguida para modernização de grafia, por exemplo:

```
<ed_mark>
     <ed t="mod">sereníssimo</ed>
     <ed t="gra">serenissimo</ed>
     <or>ſereniſsimo</or>
</ed_mark>
```

III.2.2 Exemplo: edição interpretativa

III.2.2.1 Anotação das Interferências Editoriais no Texto-Base

(a partir de um trecho da transcrição em II.2.2 acima)

```
<s>
<ed_mark><ed t="gra">Reinando</ed><or><format t="cap">R</format>E I N A N D
O</or></ed_mark>
<ed_mark><ed t="mod">aquele</ed><or>aquelle</or></ed_mark>
<ed_mark><ed t="mod">mui</ed><or>muy</or></ed_mark>
<ed_mark><ed t="mod">católico</ed><ed t="seg">catolico</ed><or>catho-<sec
t="line"/>lico</or></ed_mark>
<ed_mark><ed t="gra">e</ed><or>&</or></ed_mark>
<ed_mark><ed t="mod">sereníssimo</ed><ed
t="gra">serenissimo</ed><or>ſereniſsimo</or></ed_mark>
<ed_mark><ed t="mod">Príncipe</ed><or>Principe</or></ed_mark>
<ed_mark><ed t="mod">el-Rei</ed><or>elRey</or></ed_mark> Dom <sec t="line"/>
<ed_mark><ed t="gra">Manuel</ed><or>MANVEL</or></ed_mark> ,
<ed_mark><ed t="seg">fez-se</ed><ed t="gra">fezse</ed><or>fezſe</or></ed_mark>
<ed_mark><ed t="mod">uma</ed><or>hũa</or></ed_mark> frota pera a
<ed_mark><ed t="mod">Índia</ed><or>India</or></ed_mark> <sec t="line"/> de que
<ed_mark><ed t="mod">ia</ed><or>hia</or></ed_mark> por
<ed_mark><ed t="mod">capitão</ed><or>capitam</or></ed_mark> mór
<ed_mark><ed t="seg">Pedro Álvares</ed><ed t="seg">Pedraluarez</ed><or>Pedralua-<sec
t="line"/>rez</or></ed_mark> Cabral: que
<ed_mark><ed t="mod">foi</ed><or>foy</or></ed_mark> a
<ed_mark><ed t="gra">segunda</ed><or>ſegunda</or></ed_mark>
<ed_mark><ed t="mod">navegação</ed><ed t="gra">navegaçam</ed><ed
t="seg">nauegaçam</ed><or>nauega-<sec t="line"/>çam</or></ed_mark> que
<ed_mark><ed t="mod">fizeram</ed><or>fezeram</or></ed_mark> os
<ed_mark><ed t="gra">Portugueses</ed><or>Portugueſes</or></ed_mark>
<ed_mark><ed t="mod">para</ed><or>pera</or></ed_mark>
<ed_mark><ed t="mod">aquelas</ed><or>aquellas</or></ed_mark>
<ed_mark><ed t="seg">partes</ed><or>par-<sec t="line"/>tes</or></ed_mark> do
Oriente.</s>
```

II.2.2.2 Apresentações possíveis a partir da anotação de edição

(Os mecanismos envolvidos na geração de apresentações tais como exemplificadas abaixo se explicam na seção IV. Apresentação).

REINANDO aquelle muy catho- lico & ſereniſsimo Principe elRey Dom MANVEL , fezſe hũa frota pera a India de que hia por capitam mór Pedralua- rez Cabral: que foy a ſegunda nauega- çam que fezeram os Portugueſes pera aquellas par- tes do Oriente.	Reinando aquele muito católico e sereníssimo Príncipe el-Rei Dom Manuel, fez-se uma frota para a Índia de que ia por capitão mór Pedro Álvares Cabral: que foi a segunda navegação que fizeram os Portugueses para aquelas partes do Oriente.

III.3 Diretrizes para a transposição digital de outras edições

Edições originalmente processadas por outros sistemas podem ser adaptadas ao sistema eletrônico; isto envolve alguns procedimentos de adaptação específicos, em especial quando se tratam de edições semi-interpretativas de manuscritos.

Na adaptação de outras edições, a etapa da transcrição equivale a uma transposição digital, cujo principal objetivo é adaptar as marcas do primeiro editor para o sistema eletrônico mostrado anteriormente.

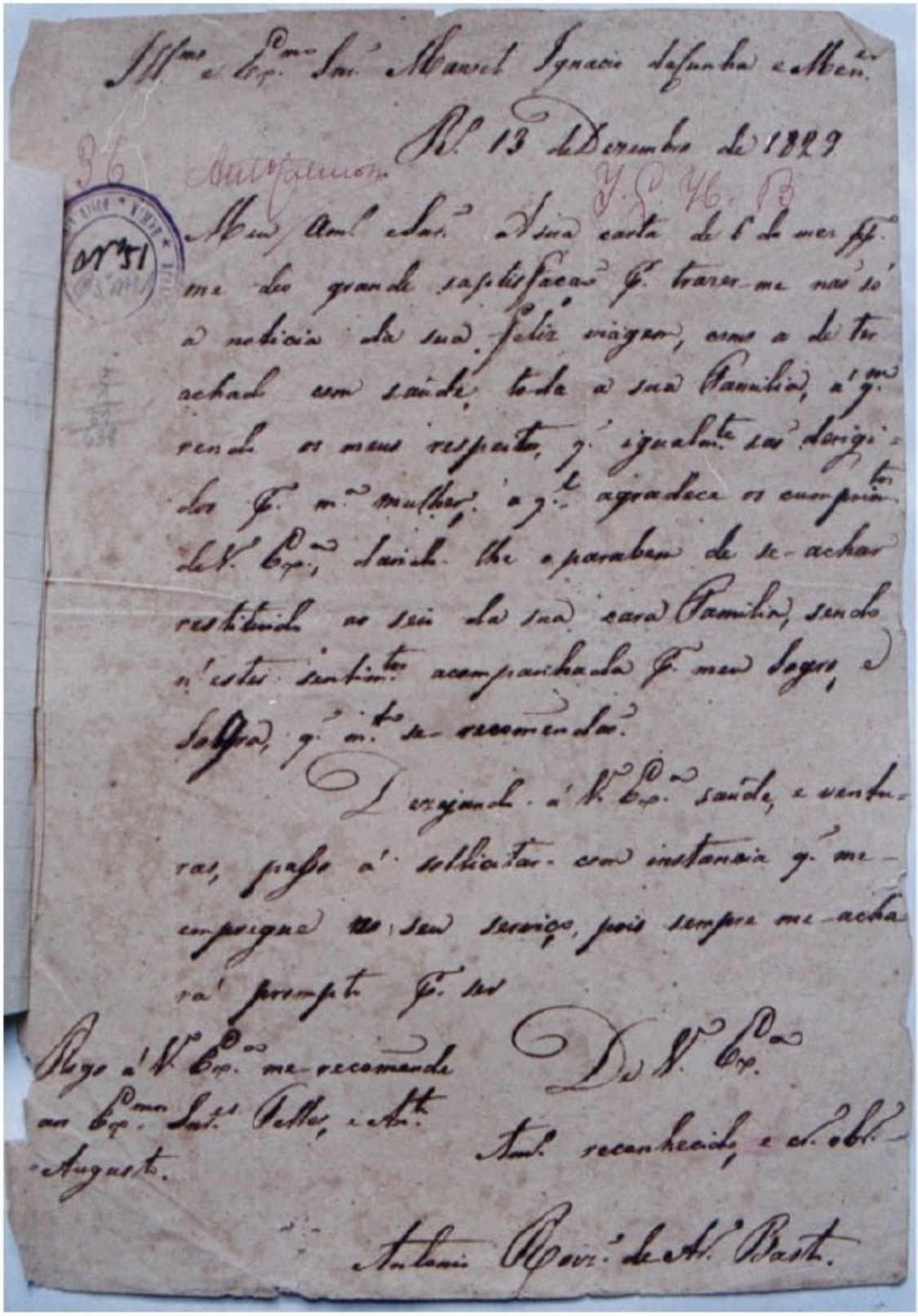
Isto se aplica tanto à transposição dos caracteres e às anotações da estrutura textual (quebras de linha e de página, etc.), mostradas na **seção II**, como para as intervenções editoriais do primeiro editor (como desenvolvimento de abreviaturas), mostradas em III.1 e III.2 acima.

A principal diferença entre os procedimentos de transcrição e preparação de textos diretamente do original tal como anteriormente apresentados e esta transposição de outras edições é que neste último caso, trata-se de identificar, nas marcas de anotação, a autoria de cada intervenção.

A seguir, mostra-se um exemplo de transposição de um conjunto de textos-fonte originalmente editados em formato Word (.doc) – o corpus "*Cartas Brasileiras*", na edição semi-diplomática de Zenaide O.N. Carneiro. O Corpus é transposto para o sistema eletrônico, adaptando-se as intervenções editoriais da preparação original, e mantendo-se um registro de autoria de cada uma delas. Posteriormente, os documentos passarão pelo processo de uniformização ortográfica

III.3.1 Exemplo: Transposição Digital de uma edição semi-diplomática

A. O Manuscrito Original (fac-simile)



Fonte: Carneiro, Zenaide de Oliveira Novais (2005). "Cartas Brasileiras (1808-1904): Um Estudo Lingüístico-Filológico". Tese de Doutoramento, Instituto de Estudos da Linguagem, Universidade Estadual de Campinas. Anexo - *Carta 1.*

B. A Edição Semi-Diplomática

Carta 1

AIGHBA. Antiga pasta 5. Documento contendo um fólio. Papel almaço sem pautas protegido por papel manteiga. Carimbo do IGHB na margem superior esquerda com a anotação, "Nº 51", além de outras a lápis, "P5m1 e, mais abaixo,"5/1/51/633". Acima do carimbo encontra-se outra anotação em vermelho: "36 Ant° Calmon I.G.H.B". No segundo fólio as informações relativas ao destinatário foram escritas na vertical.

Ill*ustrissi*mo e Ex*celentissi*mo S*enho*r Manoel Ignacio daCunha eMen*ez*es|

R*i*o 13 de Dezembro de 1829.|

Meu am*ig*o eS*enho*r. A sua carta de 6 do mes p*roximo* p*assado*| me deo grande saptisfação p*or* trazer-me não só| a noticia da sua feliz viagem, como a de ter| achado com saúde toda a sua Familia, á q*ue*m| rendo os meus respeitos, q*ue* igualm*en*te são derigi=|dos p*or* m*inh*a mulher, a q*ua*l agradece os cumprim*en*tos| deV*ossa* Ex*celênci*a, dando-lhe o paraben de se achar| restituido ao seio da sua cara Familia, sendo| n'estes sentim*en*tos acompanhada p*or* meu sogro, e| sogra[1], q*ue* m*ui*to se – recomendão.|

Dezejando á V*ossa* Ex*celenci*a saúde, e ventu=|ras passo á sollicitar com instancia q*ue* me-| empregue no seu serviço, pois sempre me – acha|rá prompto p*or* ser|

Rogo á V*ossa* Ex*celênci*a me–recomende| aos Ex*celentissi*mos S*enho*res Telles, e Ant*oni*o [?]| Augusto. |

De V*ossa* Ex*celenci*a|

Am*ig*o reconhecido, e cr*iad*o obr*igad*o|

Antonio Ro*drigu*ez de Ar*auj*o Basto.|

[1] Borrado.

Fonte: Carneiro, Zenaide de Oliveira Novais (2005). "Cartas Brasileiras (1808-1904): Um Estudo Lingüístico-Filológico". Tese de Doutoramento, Instituto de Estudos da Linguagem, Universidade Estadual de Campinas. Anexo - *Carta 1.*

C. A Transposição Digital

```
<text id="cb_001">
<sec t="letter">  <sec t="title">Carta 1</sec>
  <comment t="heading" author="Z.C.">AIGHBA. Antiga pasta 5. Documento contendo um fólio. Papel almaço sem
pautas protegido por papel manteiga. Carimbo do IHGB na margem superior esquerda com a anotação,"N 51", além
de outras a lápis, "P5ml" e, mais abaixo,"5/1/51/633". Acima do carimbo encontrase outra anotação em
vermelho:"36 Ant CAlmon I.H.G.B". No segundo fólio as informações relativas ao destinatário foram escritas
na vertical.</comment>
 <s type="heading">
 <ed_mark><ed author="Z.C.">Ilustrissimo </ed> <or>Illmo</or>  </v>   e
 <ed_mark><ed author="Z.C.">Excelentissimo</ed> <or>Exmo</or>  </v>
 <ed_mark><ed author="Z.C.">Senhor </ed> <or>Snr</or>  </v> Manoel Ignacio daCunha
 <ed_mark><ed author="Z.C.">eMenezes </ed> <or>eMenzes</or></v>   <sec t="line"/> </s>
 <s t="date">Rio 13 de Dezembro de 1829.<sec t="line"/>   </s>
 <p>
 <s> Meu amigo
  <ed_mark><ed author="Z.C.">eSenhor </ed><or>eSnr</or></ed_mark>  .   </s>
 <s>  A sua carta de 6 do mes
  <ed_mark><ed author="Z.C." t="exp">proximo </ed><or>p</or></ed_mark>
  <ed_mark><ed author="Z.C." t="exp">passado </ed><or>psso</or></ed_mark>  <sec t="line"/>
  me deo grande saptisfação
  <ed_mark><ed author="Z.C." t="exp">por </ed><or>pr</or></ed_mark> trazerme não só   <sec
t="line"/>   a noticia da sua feliz viagem, como a de ter   <sec t="line"/> achado com saúde
toda a sua Familia, á
  <ed_mark><ed author="Z.C." t="exp">quem </ed><or>qm</or></ed_mark>  <sec t="line"/>
  rendo os meus respeitos,
  <ed_mark><ed author="Z.C." t="exp">que </ed><or>q</or></ed_mark> igualmente são derigi=   <sec
t="line"/>   dos
  <ed_mark><ed author="Z.C." t="exp">por </ed><or>pr</or></ed_mark>
  <ed_mark><ed author="Z.C." t="exp">minha </ed><or>ma</or></ed_mark>   mulher, a
  <ed_mark><ed author="Z.C." t="exp">qual </ed><or>ql</or></ed_mark>   agradece os
  <ed_mark><ed author="Z.C." t="exp">cumprimentos</ed> <or>cumprimtos</or>  </ed_mark>  <sec
t="line"/>
  <ed_mark><ed author="Z.C." t="exp">deVossa </ed><or>deV</or></ed_mark>
  <ed_mark><ed author="Z.Carneiro t="exp"">Excelência</ed><or>Exa</or></ed_mark> ,dandolhe o
paraben de se achar   <sec t="line"/>   restituido ao seio da sua cara Familia, sendo   <sec
t="line"/>   n'estes
  <ed_mark><ed author="Z.C." t="exp">sentimentos </ed><or>sentimtos</or></ed_mark> acompanhada
  <ed_mark><ed author="Z.C." t="exp">por </ed><or>pr</or></ed_mark> meu sogro, e   <sec
t="line"/>
  <ed_mark t="bor" author="Z.C.">sogra</ed_mark> ,
  <ed_mark><ed author="Z.C." t="exp">que </ed><or>q</or></ed_mark>
  <ed_mark><ed author="Z.C." t="exp">muito </ed><or>mto</or> </ed_mark> se recomendão.   <sec
t="line"/> </s>
</p>
 <p>
 <s>  Dezejando á
  <ed_mark><ed author="Z.C." t="exp">Vossa Excelencia</ed><or>V Exa</or>  </ed_mark> saúde, e
ventu   <sec t="line"/> ras passo á solicitar com instancia
  <ed_mark><ed author="Z.C." t="exp">que </ed><or>q</or></ed_mark> me   <sec t="line"/>
  empregue no seu serviço, pois sempre me acha   <sec t="line"/>
rá prompto
  <ed_mark><ed author="Z.C." t="exp">por </ed><or>pr</or></ed_mark> ser   <sec t="line"/>
  Rogo á
  <ed_mark><ed author="Z.C." t="exp">Vossa Excelência</ed><or>V Exa</or></ed_mark> me recomende
<sec t="line"/>   aos
  <ed_mark><ed author="Z.C." t="exp">Excelentissimos</ed><or>Exmos</or></ed_mark>
  <ed_mark><ed author="Z.C." t="exp">Senhores </ed><or>Sres</or></ed_mark> Telles, e
  <ed_mark><ed author="Z.C." t="exp">Antonio </ed><or>Anto</or></ed_mark> <sec t="line"/>
  Augusto.   <sec t="line"/> </s>
</p>
 <s>  De
  <ed_mark><ed author="Z.C." t="exp">Vossa Excelencia</ed><or>V Exa</or></ed_mark>  <sec
t="line"/>
  <ed_mark><ed author="Z.C." t="exp">Amigo </ed><or>Amo</or></ed_mark>  reconhecido, e
  <ed_mark><ed author="Z.C." t="exp">criado </ed><or>cro</or></ed_mark>
  <ed_mark><ed author="Z.C." t="exp">obrigado </ed><or>obro</or></ed_mark>  <sec t="line"/> </s>
 <s t="signature">   Antonio
  <ed_mark><ed author="Z.C." t="exp">Rodriguez </ed><or>Roez</or></ed_mark>   de
  <ed_mark><ed author="Z.C." t="exp">Araujo </ed><or>Aro</or></ed_mark>   Basto. </s> </sec>
</text>
```

III.3.2 Normas para a Transposição Digital de Outras Edições

Na transposição de outras edições, a transcrição segue as normas expostas em II.1 e II.2 acima , quanto à transcrição dos caracteres alfabéticos, à pontuação e organização espacial do texto, e às indicações de dificuldade de leitura.

A diferença primordial é que nestes casos, os comentários, as indicações de leitura e outras interferências do primeiro editor são também transcritos, **e seu autor é identificado**, como se detalha abaixo.

- Os comentários do primeiro editor são transcritos e indicados, e identificadas pelo atributo *author* na catagoria relevante*:*

```
<comment author="Z. Carneiro">AIGHBA. Antiga pasta 5. Documento
contendo um fólio. Papel almaço sem pautas protegido por papel
manteiga. Carimbo do IHGB na margem superior esquerda com a anotação,"N
51", além de outras a lápis, "P5ml" e, mais abaixo,"5/1/51/633". Acima
do carimbo encontrase outra anotação em vermelho:"36 Ant CAlmon
I.H.G.B". No segundo fólio as informações relativas ao destinatário
foram escritas na vertical.</comment>
```

- As indicações de pontos de deterioração do suporte material do original, feitas pelo primeiro editor, são indicados e identificados pelo atributo *author* na categoria relevante:

Texto a ser transposto:
[sogra] 1 (*nota de rodapé do primeiro editor*: ... 1 Borrado) Transposição:
`<ed_mark t="bor" author= "Z. Carneiro">sogra</ed_mark>`

- As interferências do primeiro editor são transcritas e indicadas, e o nome do primeiro editor é acrescentado como atributo na categoria relevante:

Texto a ser transposto:
S*en*h*o*r Manoel Ignacio daCunha eMen*e*zes Transposição:

```
<ed_mark>
     <ed author="Z.C.">Senhor</ed>
     <or>Snr</or> </ed_mark> Manoel Ignacio daCunha
<ed_mark>
     <ed author="Z.C.">eMenezes</ed>
     <or>eMenzes</or> </ed_mark>
```

III.3.1.1 Adaptação Facilitada no processador Open Office

Na maioria dos casos, os materiais já transcritos e editados foram trabalhados em formato .doc, usando formatações como *itálico*, sublinhado, etc. para indicar interferências do editor.

Se queremos transformar esses materiais em documentos xml, pode ser útil aprender algumas técnicas breves que permitem adaptar a formatação em .doc de modo a ser aproveitada pelo Emacs.

Para isso, é interessante usar o processador de textos da suite Open Office, que consegue trabalhar em formato .doc e gerar documentos de texto simples com codificação bem preservada, que então podem ser transformados em .xml no Emacs. A grande vantagem de passar os documentos no OO é que é possível fazermos substituições com expressões regulares, e assim não perdemos as informações codificadas com formatações no .doc original.

Para codificar automaticamente as informações marcadas como formatações do Word no documento de

origem na forma da nossa anotação para marcas de edição (<ed></ed>, etc.), é necessário seguir os passos abaixo:

(1) Abrir o documento .doc no Open Office

(2) Identificar a estratégia de anotação do primeiro editor (por exemplo: itálico para os trechos inseridos, como expansões de abreviatura)

(3) Realizar substituições dessas formatações por marcas de XML, com expressão regular. No caso de se desejar transformar todos os trechos em itálico do documento em trechos encapsulados entre <ed></ed>, a expressão é:

abrir *edit/find & replace* (ctrl+F)
ativar "*use regular expression*"
"*search for*" > "*format*" > " *italic*"
"*replace with*" > <ed>&</ed>

(4) Salvar o documento como texto simples (.txt)

(5) Abrir esse documento .txt no Emacs e transformá-lo em .xml, e prosseguir na anotação.

Ferramentas para esse procedimento:

Open Office (Suite de editores de texto): <http://www.openoffice.org>

IV. Apresentação

IV.1 Fundamentos

IV.1.1. Visão Geral

Como já se mencionou anteriormente, nas edições eletrônicas o sistema de **apresentação da edição** – ou organização da leitura pública final dos textos – difere bastante do sistema tradicional.

No sistema tradicional, a apresentação final é uma etapa conseqüente da transcrição e do estabelecimento do texto , ou seja, o próprio documento preparado é organizado para a leitura final. Nas edições eletrônicas, a apresentação final é uma etapa paralela à transcrição e ao estabelecimento do texto – ou seja: as apresentações para leitura final resultam da formação de documentos paralelos ao documento-base onde se codificou a preparação.

É o que chamamos de **geração de versões**. Nos documentos-base, a transcrição do texto e as intervenções do editor estão codificadas ou anotadas em linguagem XML; essas anotações podem ser lidas por programações computacionais que selecionam e organizam as estruturas dos documentos-base e formam, a partir delas, os documentos de apresentação. Como também já se mencionou, na etapa da apresentação as diversas camadas de edição aplicadas aos textos podem ser separadas para o acesso do leitor final. Assim, um mesmo documento pode ser lido nas versões Edição Conservadora, Edição Semi-Interpretativa, Edição Interpretativa, etc; e a geração das versões, por ser paralela, não traz nenhuma conseqüência para os documentos-base, que permanecem inalterados.

Na seção IV.2 , detalham-se os aspectos técnicos da preparação das programações utilizadas na geração de versões neste sistema. Aqui, apresentamos um exemplo das versões possíveis para um mesmo texto, resumindo onde relevante alguns aspectos da programação.

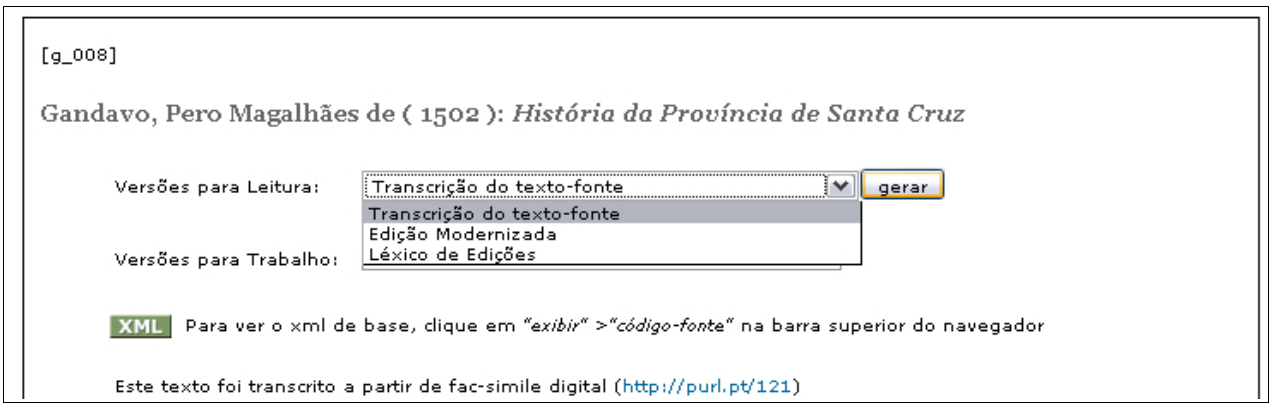
IV.1.2. Exemplo: apresentação de um texto editado eletronicamente

Tomaremos por exemplo, nesta seção, as diferentes formas de apresentação possíveis a partir da edição eletrônica do texto "*História da Província de Santa Cruz*", de Magalhães de Gandavo. A fonte original do texto é um facsímile digitalizado, trazido a público pela Biblioteca Nacional de Lisboa:

> História da prouincia Sãcta Cruz que vulgarme[n]te chamamos Brasil / feita por Pero Magalhães de Gandauo. Em Lisboa : na officina de António Gonsaluez: vendense em casa de Ioão Lopez, 1576. - 48 f. : 1 est. ; 4º (18 cm) - Assin: A-F//8. - Anselmo 709. - Faria - BN Rio de Janeiro p. 38. - B. MUseum 150 coln 204 <http://purl.pt/121>

Este texto-fonte foi transcrito de acordo com os procedimentos apresentados na seção II acima; e uma edição interpretativa foi preparada de acordo com as normas apresentadas na seção III acima.

Esta preparação, em todos os seus passos, está codificada no arquivo <http://www.ime.usp.br/~tycho/corpus/texts/xml/g_008.xml>. A partir deste documento-base, é possível produzir apresentações paralelas, sob forma de diferentes arquivos gerados automaticamente:

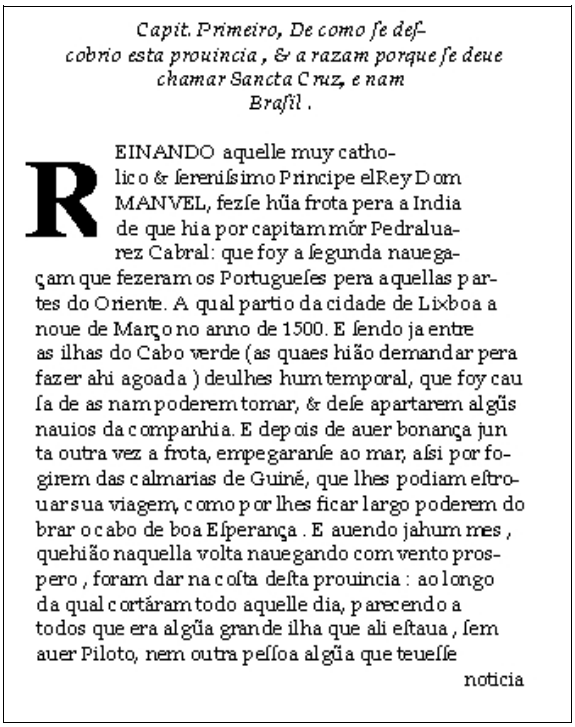


- Ao escolhar a opção "Transcrição do texto-fonte", o usuário irá ativar a geração instantânea de uma versão .html do texto, formatado e com a grafia preservada do original (cf. exemplo (A) abaixo).

- Ao escolhar a opção "Edição Modernizada", o usuário irá ativar a geração instantânea de uma versão .html do texto, formatado e com a grafia modernizada (cf. exemplo (B) abaixo).

- Ao escolhar a opção "Léxico de edições", o usuário irá ativar a geração instantânea de uma versão .html do texto, com a lista de todas as intervenções editoriais realizadas (cf. exemplo (C) abaixo).

- Ao escolhar a opção "Edição Modernizada - texto simples", o usuário irá ativar a geração instantânea de uma versão .txt do texto, com a grafia modernizada (versão adequada para o processamento posterior por ferramentas automáticas; cf. exemplo (D) abaixo).



Capit. Primeiro, De como ſe deſ-
cobrio esta prouincia, & a razam porque ſe deue
chamar Sancta Cruz, e nam
Braſil.

REINANDO aquelle muy catho-
lico & ſerenıſsimo Principe elRey Dom
MANVEL, fezſe hũa frota pera a India
de que hia por capitam mõr Pedralua-
rez Cabral: que foy a ſegunda nauega-
çam que fezeram os Portugueſes pera aquellas par-
tes do Oriente. A qual partio da cidade de Lixboa a
noue de Março no anno de 1500. E ſendo ja entre
as ilhas do Cabo verde (as quaes hião demandar pera
fazer ahi agoada) deulhes hum temporal, que foy cau
ſa de as nam poderem tomar, & deſe apartarem algũs
nauios da companhia. E depois de auer bonança jun
ta outra vez a frota, empegaranſe ao mar, aſsi por fo-
girem das calmarias de Guiné, que lhes podiam eſtro-
uar sua viagem, como por lhes ficar largo poderem do
brar o cabo de boa Eſperança. E auendo jahum mes,
quehião naquella volta nauegando com vento pros-
pero, foram dar na coſta deſta prouincia: ao longo
da qual cortáram todo aquelle dia, parecendo a
todos que era algũa grande ilha que ali eſtaua, ſem
auer Piloto, nem outra peſſoa algũa que teueſſe
noticia

Este é um trecho do texto na Versão Conservadora ou Diplomática, a partir da programação em origversion.xsl (cf. detalhes técnicos em IV.2 abaixo).

Esta apresentação do texto corresponde fundamentalmente ao texto no estado de Transcrição Conservadora, de acordo com as normas apresentadas anteriormente na seção II. Isso significa que a programação de transformação selecionou, no arquivo XML integral, todas as as estruturas de texto, titulação, etc.; mas, entre as estruturas marcadas como variantes, selecionou apenas as sub-estruturas marcadas como originais. Além disso, a programação formata o texto para melhor leitura na apresentação.

Esta forma de apresentação dirige-se à leitura do público interessado no texto em sua forma original.

```
Capítulo Primeiro, De como se descobriu
esta província ,
e a razão porque se deve
chamar Santa Cruz, e não
Brasil.

REINANDO aquele muito católico
e sereníssimo Príncipe el-Rei Dom
Manuel , fez-se uma frota para a Índia
de que ia por capitão mór Pedro Álvares
Cabral: que foi a segunda navegação
que fizeram os Portugueses para aquelas partes
do Oriente. A qual partiu da cidade de Lisboa a
nove de Março no ano de 1500. E sendo já entre
as ilhas do Cabo Verde (as quais iam demandar para
 fazer aí aguada ) deu-lhes um temporal, que foi causa
de as não poderem tomar, e de se apartarem alguns
 navios da companhia. E depois de haver bonança junta
outra vez a frota, empegaram-se ao mar, assim por fugirem
das calmarias de Guiné, que lhes podiam estorvar
sua viagem, como por lhes ficar largo poderem dobrar
o cabo de Boa Esperança. E havendo já um mês,
que iam naquela volta navegando com vento próspero,
foram dar na costa desta província : ao longo
da qual cortaram todo aquele dia, parecendo a
todos que era alguma grande ilha que ali estava , sem
haver Piloto, nem outra pessoa alguma que tivesse
                                        [ notícia ]
```

Este é um trecho do texto na Versão Edição Interpretativa. A programação usada para gerar esta versão está no arquivo edversion.xsl (cf. detalhes técnicos em IV.2 abaixo).

Esta apresentação do texto corresponde fundamentalmente ao texto depois do proceso de uniformização da grafia, de acordo com as normas apresentadas anteriormente na seção III. Isso significa que a programação de transformação selecionou, no arquivo XML integral, todas as as estruturas de texto, titulação, etc.; mas, entre as estruturas marcadas como variantes, selecionou apenas as sub-estruturas marcadas como editadas. Além disso, a programação formata o texto para melhor leitura na apresentação.

Esta forma de apresentação dirige-se à leitura do público interessado no texto já interpretado, de mais fácil acesso.

```
START TEXT :|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|:|

 [frontispicio]
 [licenças]
 [prefácio]
 [prefácio]
 [prólogo]
 [prólogo]

[title: Capítulo Primeiro, De como se descobriu esta província ,


        [g_008_s_57] Reinando aquele muito católico e sereníssimo
        [g_008_s_58] A qual partiu da cidade de Lisboa a nove de
        [g_008_s_59] E sendo já entre as ilhas do Cabo Verde (as
        [g_008_s_60] E depois de haver bonança junta outra vez a
        [g_008_s_61] E havendo já um mês , que iam naquela volta
        [g_008_s_62] E no lugar que lhes pareceu dela mais acomod
        [g_008_s_63] Estando assim surtos nesta parte que digo, s
        [g_008_s_64] Ao outro dia seguinte , saiu Pedro Álvares e
        [g_008_s_65] E assim se punham de joelhos e batiam nos pe
        [g_008_s_66] No que mostravam claramente estarem disposto
        [g_008_s_67] Então despediu logo Pedro Álvares um navio c
        [g_008_s_68] E tornando a Pedro Álvares seu descobridor ,
        [g_008_s_69] O que não parece carecer de mistério , porqu
        [g_008_s_70] Por onde não parece razão, que lhe neguemos
        [g_008_s_71] Ao qual chamaram brasil por ser vermelho e t
        [g_008_s_72] Mas para que nesta parte magoemos ao Demônio
        [g_008_s_73] Porque na verdade mais é de estimar e melhor

 [title: Capítulo 2 Em que se descreve o sítio e qualidades desta
```

Este é um trecho do texto na Versão Edição Interpretativa para Análise Automática. A programação usada para gerar esta versão está no arquivo lain.xsl (cf. detalhes técnicos em IV.2 abaixo).

Esta apresentação do texto corresponde também ao texto depois do processo de uniformização da grafia, de acordo com as normas apresentadas anteriormente na seção III. Da mesma forma como na versão apresentada logo acima (4), a programação de transformação selecionou no arquivo XML integral, entre as estruturas marcadas como variantes, apenas as sub-estruturas marcadas como editadas. Além disso, entretanto, esta programação não inclui na apresentação todas as as estruturas de texto, mas apenas as sentenças e a titulação. Além disso, a transformação não inclui nenhuma formatação na apresentação final..

Esta forma de apresentação dirige-se à leitura de ferramentas automáticas.

História da Província de Santa Cruz

Léxico das Edições

Complete

[voltar para ficha catalográfica]

Ressegmentação | Modernização | Expansão de Abreviaturas
| Adaptação de Pontuação | Correções | Ressegmentação Técnica | Outros

Modernização

Item Editado	Item Original	
Epístola	Epiſtola	[g_008_v_3]
serviço	ſeruiço	[g_008_v_5]
ilustre	illuſtre	[g_008_v_6]
senhor	ſenhor	[g_008_v_7]
ofereço	offere-ço	[g_008_v_8]
primícias	premicias	[g_008_v_10]
desejos	deſejos	[g_008_v_13]
possibilidade	poſsibilidade	[g_008_v_14]
alguma	algũa	[g_008_v_15]
que se	que ſe queſe	[g_008_v_16]
deve	deue	[g_008_v_18]
à	á	[g_008_v_19]

Este é um trecho do texto na Versão Léxico de Edições, gerado automaticamente pela seleção de "Léxico de Edições" na figura anterior. A programação usada para gerar esta versão está no arquivo varietyex.xsl (cf. detalhes técnicos em IV.2 abaixo).

Esta apresentação do texto corresponde a um glossário das interferências editoriais anotadas no texto de base de acordo com as normas apresentadas anteriormente na seção III. A programação de transformação selecionou no arquivo XML integral apenas as estruturas marcadas como variantes (deixando de lado todo orestante do texto). As estruturas variantes são organizadas em listas, e para cada sub-estrutura marcada como editada, apresenta-se a sub-estrutura correspondente marcada como original. Inclui, além disso, o código identificador de cada estrutura. É possível, ainda, agruparem-se as interferências por categorias (uniformização grafemática, modernização de grafia, etc., tal como previsto na seção III.

A finalidade central desta forma de apresentação é fornecer um registro controlado da interferência editorial nos textos.

IV.1.3. O Hipertexto Crítico

A versão mais sofisticada deste sistema de edição eletrônica seria uma sistematização da apresentação que reuna, sob forma de um só documento de hipertexto, todas as versões atualmente apresentadas sob forma de documentos paralelos gerados separadamente.

Ou seja: um documento gerado a partir do documento de base XML que funcione como um hipertexto sinóptico, permitindo o acesso imediato às diferentes camadas de edição em cada ponto do texto (e não apenas *para cada texto*, como prevê o sistema até agora).

Este sistema servirá tanto para uma apresentação complexa de edições em diversos níveis de edição (conservador, semi-interpretativo, interpretativo) como também para uma apresentação integrada entre esses níveis de edição e as versões dos textos analisadas pelas ferramentas automáticas de análise lingüística (morfossintaxe e sintaxe).

Servirá, além disso, para uma apresentação complexa de edições realizadas a partir de mais de um testemunho de cada obra – ou seja, edições críticas e edições genéticas.

Esta modalidade de apresentação é que denominamos aqui de **Hipertexto Crítico**.

A tecnologia implementada até este momento nos estágios de preparação dos textos permite que este tipo de apresentação crítica seja realizado. O lançamento integral desta idéia depende especificamente do desenvolvimento da técnica de apresentação em si, etapa que se encontra, correntemente, em desenvolvimento.

Neste momento, há dois textos sendo preparados para a composição de Hiperedições Críticas:

Sermão das Lágrimas de São Pedro, 1679 Antonio Vieira (1608);
História da Província de Santa Cruz, 1576 Pero Magalhães de Gandavo (1502).

IV.2 Procedimentos

Detalham-se aqui os procedimentos seguidos para programar a geração de diferentes versões dos documentos .xml a partir da preparação discutida nas seções I e II, por meio de transformações XSLT (cf. Diagrama do Fluxo de Procedimentos).

A anotação XML permite que se gerem diferentes versões para um mesmo documento (com a grafia original, com a grafia editada, léxicos de edições, etc - cf. IV.1.2), programando-se transformações em linguagem XSLT.

As transformações atuam sobre as anotação XML de diferentes modos, podendo selecionar, re-ordenar ou formatar as estruturas anotadas; e geram documentos em diferentes formatos (no caso atual, .txt e .html). Para executar essas transformações, utilizamos o software **Saxon** , armazenado no servidor. Os detalhes dos procedimentos estão em IV.2.3<).

No momento, existem as seguintes programações para os textos do corpus:

Resumo das Programações XSLT

Folha .xsl	*Programação*	*Resultado*
insert_id.xsl	Reescreve o documento XML anotado, numerando as seguintes instâncias: seções (<sec>), parágrafos (<p>), sentenças (<s>), marcas de edição (<ed_mark>), itens editados (<ed>), itens originais (<or>), elementos de texto (<text_el>).	Documento com estruturas numeradas (.xml) **Nota:** *Esta programação deve ser ativada assim que a anotação do documento estiver completa, antes de todas as programações abaixo.*
origversion.xsl	Seleciona, entre os elementos <ed_mark>, apenas os sub-elementos <or>; formata o documento quanto a parágrafos, titulação, etc.	Documento com a versão grafia original (.html)
edversion.xsl	Seleciona, entre os elementos <ed_mark>, apenas os sub-elementos <ed>; formata o documento quanto a parágrafos, titulação, etc.	Documento com a versão grafia modernizada (.html)
varietylex.xsl	Seleciona do texto apenas os elementos <ed_mark>, listando os pares <ed> | <or>; formata o documento como uma tabela.	Documento com a lista das edições realizadas (.html)
plain.xsl	Seleciona, entre os elementos <ed_mark>, apenas os sub-elementos <ed>; forma uma lista de sentenças numeradas, sem formatação	Documento com as sentenças listadas (.txt)

A produção dessas versões se dá "on-line" através doCatálogo Dinâmico (cf. seção I)

As principais vantagens do sistema de geração de versões por XSLT são a flexibilidade e facilidade de armazenamento. Com esse sistema, armazenam-se no servidor apenas os documentos .xml de base; as diferentes versões são produzidas no momento do acesso do usuário ao catálogo, e não precisam portanto estar gravadas. O usuário, através do catálogo, ativa scripts em linguagem PERL, que remetem ao software Saxon, fazendo operar a programação e produzindo o documento desejado.

Além da economia de espaço de armazenamento, esse sistema de geração on-line permite grande flexibilidade, já que as folhas XSLT podem ser reprogramadas indefinidamente conforme se deseje refinar a reestruturação dos documentos XML.

IV.2.1 Procedimentos de Ativação

As programações de transformação XLST podem ser ativadas de duas formas: local e remota.

No servidor local, os membros da equipe podem ativá-las por comandos em um terminal (Linux ou Dos); por via remota, os usuários do corpus podem ativá-las mais facilmente por meio de hiperlinks que levam a scripts (gravados em ~tycho/cgi-bin), que por sua vez ativam as transformações.

IV.2.1.1 Ativação das transformações – via local

- A ativação local pode ser realizada em uma máquina que contenha os seguintes softwares:
- saxon (versão 7.7 ou superior; software livre, disponível em <http://www.saxon.net>)
- java (4.2; idem)
- A ativação no terminal Unix ou Dos segue o seguintes passos:
    - chamar java <+espaço>
    - chamar modo -cp <+espaço>
    - indicar o software de transformação (Saxon) – incluir caminho na estrutura de diretórios;
    - digitar o comando para transformação: net.sf.saxon.Transform <+espaço>
    - indicar o nome do documento-base – incluir caminho, e terminação (.xml);
    - indicar o nome da folha com a transformação desejada – incluir caminho, e terminação (.xslt), e digitar parêntese angular: >
    - indicar o nome do arquivo a ser gerado – incluir caminho onde ele será gravado, e terminação (.hml, txt, xml), <+enter>
    - O arquivo gerado estará disponível no diretório que foi indicado em (g).
- Esquema (para uma máquina windows onde o Saxon está gravado na raiz, c:

```
c:\>java -cp c:\saxon\saxon8.jar net.sf.saxon.Transform documentpahtandname
stylesheetpathandname > outputfilepathandname
```

- Exemplos:

    - para gerar um documento com grafia editada (ex. a_001_edversion.html) a partir de um documento XML (ex. a_001.xml), com a programação edversion.xslt:

```
c:\>java -cp c:\saxon7-7\saxon7.jar net.sf.saxon.Transform
c:\~tycho\corpus\texts\xml\a_001.xml c:\~tycho\cgi-bin\edversion.xslt
>c:\~tycho\corpus\obras\temp\a_001_edversion.html
```

    - para gerar um documento léxico (ex. a_001_lex.html) a partir de um documento XML (ex. a_001.xml), com a programação varietylex.xslt:

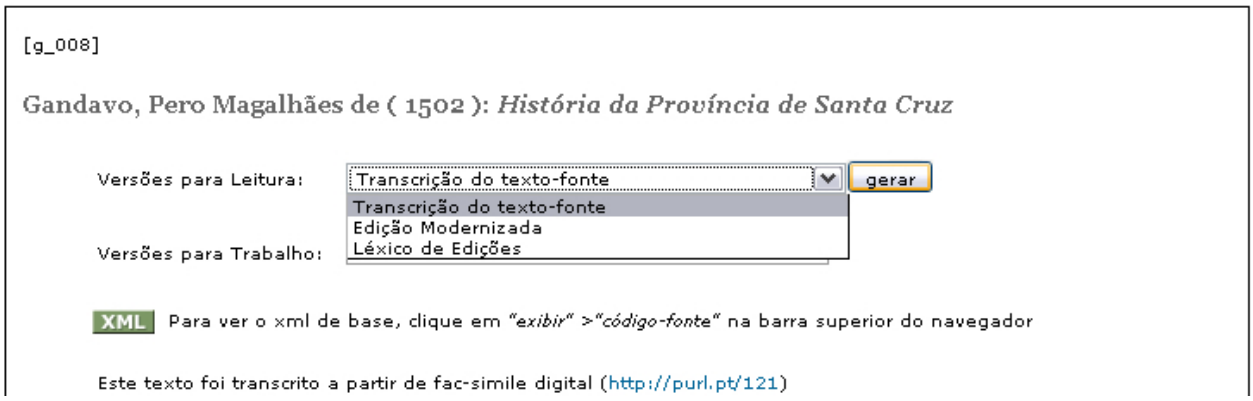
```
c:\>java -cp c:\saxon7-7\saxon7.jar net.sf.saxon.Transform
c:\~tycho\corpus\texts\xml\a_001.xml c:\~tycho\cgi-bin\varietylex.xslt
>c:\~tycho\corpus\obras\temp\a_001_lex.html
```

IV.2.1.2 Ativação das transformações – via remota

A ativação via remota não precisa proceder com comandos, pois há scripts gravados no servidor remoto que ativam automaticamente as transformações.

Ao abrir o Catálogo Dinâmico, o usuário pode escolher as versões a serem geradas, bastando clicar em um objeto de hiperlink. Este hiperlink está programado para chamar um script que ativa a programação desejada e produz um documento temporário naquela versão, da seguinte forma:

- No Catálogo Dinâmico, para cada texto forma-se a seguinte caixa de opções:



- Ao escolher uma das opções, o usuário está ativando o script getversion.pl, como mostra o código por trás da caixa acima:

```
<form name="getversion" action="getversion.pl" method="post">
    <input type="hidden" name="filename" value="{$filename}"/>
        <select name="version">
            <option value="original" selected="selected">Trasncrição do texto-fonte</option>
            <option value="modernized">Edição Modernizada</option>
            <option value="plaintext">Edição Modernizada .txt</option>
            <option value="varietylex">Léxico de Edições</option>
        </select>
    <input type="submit" value="gerar"/><input type="reset" value="Reset"/>
</form>
```

- O script getversion.pl chama a transformação desejada (origversion.xslt, edversion.xlst, varietylex.xslt);

- O documento temporário se abre no browser do usuário. Futuramente, entre as opções de versões estarão incluidos sub-corpora compactados, em arquivos .zip para download.

IV.2.1.3 Associação Direta nos Documentos

Uma terceira maneira de se obter o resultado das folhas .xsl sobre os documentos .xml é associar cada documento a uma folha, de maneira que o navegador leia o documento da maneira determinada pela programação XSLT.

No Corpus, esse método é usado para gerar a apresentação inicial dos documentos .xml sob forma de Fichas Catalográficas e Portais (cf. seção I.2.2.2, parte B)

A apresentação em forma de portal é obtida pela aplicação da folha catalog_files.xsl sobre os arquivos xml do corpus; folha está armazenada no mesmo diretório dos arquivos de textos:

```
<http://www.ime.usp.br/~tycho/corpus/texts/xml/catalog_files.xsl>
```

No servidor, todos os arquivos .xml estão gravados com a instrução de serem apresentados automaticamente desta forma. Esta instrução se obtém adicionando-se a seguinte linha em seguida à declaração de xml do documento:

```
<?xml-stylesheet href="catalog_files_2006.xsl" type="text/xsl"?>
```

Ao longo do processamento dos textos, pode ser interessante gravar nesta linha uma das outras folhas .xsl, para que se possa prever o resultado da programação XSLT sobre o documento mesmo antes de se gerar a versão desejada.

IV.2.2 Programações de Transformação

A programação XLST se dá pela interpretação do X-Path, ou caminho formado pelos nós da estrutura hierárquica de árvores na anotação XML.

Um guia completo de programação XSLT pode ser consultado em: http://www.w3schools.com/xsl/

As folhas XLST já concebidas e disponíveis no servidor trabalham sobretudo com a categoria <ed_mark>, gerando diferentes versões quanto à edição dos textos. Além desse aspecto principal, as programações organizam a estrutura das páginas quanto à paragrafação, formatação, etc. - neste ponto, contam ainda com um vínculo a uma folha de estilos (CSS) que indica o lay-out da página (margens, etc.) e detalhes da formatação (famílias de fontes, cores, etc.)

A seguir temos um pequeno trecho da programação XSLT traduzido para linguagem corrente. O ponto central é observar como cada folha trabalha com a estrutura <ed_mark>:

Trecho da folha edversion.xslt:		*Tradução:*
`<xsl:template match="ed_mark">`	(1)	ao encontrar o elemento "ed_mark",
`<xsl:apply-templates/>` `</xsl:template>`	(2)	aplique os esquemas a seguir.
`<xsl:template match="edited">`	(3)	Ao encontrar o elemento "edited",
`<xsl:apply-templates select="edited"/>` `</xsl:template>`	(4)	aplique o esquema: selecione "edited" .
`<xsl:template match="original">`	(5)	Ao encontrar o elemento "original",
`</xsl:template>`	(6)	nenhum esquema indicado (i.e.: a programação não selecionará o elemento original)

- As folhas XSLT já programadas estão gravadas no diretório ~tycho/cgi-bin e ~tycho/corpus/texts/docs (cf. I.3 acima).