Dicionário Histórico do Português do Brasil (séculos XVI, XVII e XVIII)

Ferramentas computacionais para possibilitar o uso universal da Base Informatizada sobre o Brasil dos séculos XVI, XVII e XVIII

Sandra Maria Aluísio

NILC

USP
UNIVERSIDADE DE SÃO PAULO

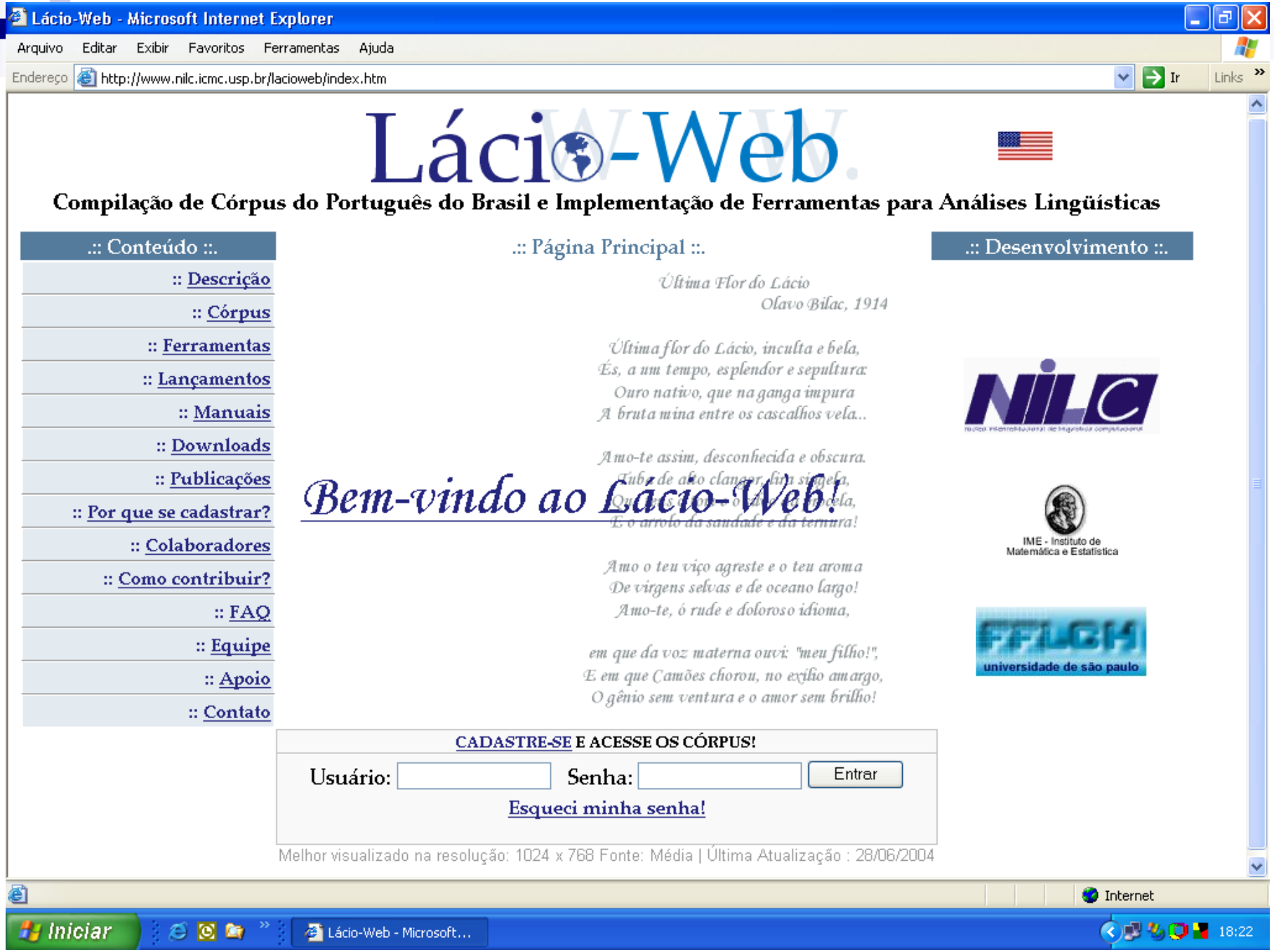
Reunião de Trabalho – 19/12/2005

Roteiro

- Como fizemos o Portal do Projeto Lácio-Web (LW)
  - Definição dos cabeçalho dos textos
    - Tipologia Textual quadripartida: gênero, tipos de texto, domínio e meio de distribuição
  - Editor de cabeçalhos e a marcação em XML
  - Base de dados que dá suporte às buscas para montagem de subcórpus
- O que aprendemos: o que faríamos diferente
- Proposta para o Projeto do Dic_Hist
  - Proposta para o Portal e mão-de-obra necessária
  - Proposta para o gerenciamento do projeto: uso de uma Web colaborativa – coteia do ICMC

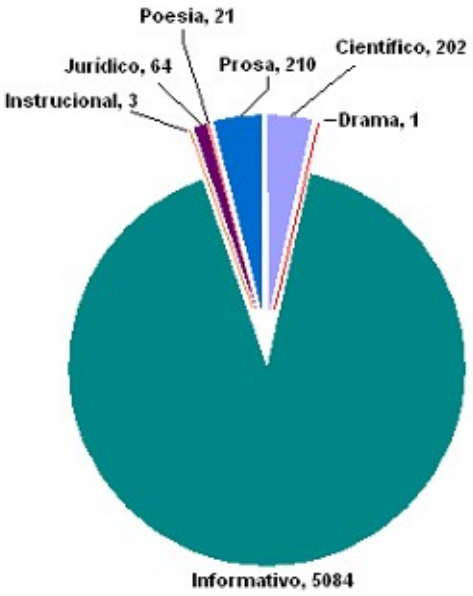
Projeto Lácio-Web

- 30 meses de projeto
- 4 tipos diferentes de córpus: Lácio-Ref, Par-C, Comp-C, Mac-Morpho (embora 6 fossem previstos)
- Obtida a Autorização de Uso – via assinatura de um termo – para todos os textos dos córpus
- 16 bolsistas entre DTI e ITI
- 3 pesquisadores – 2 computação, 1 lingüista de córpus

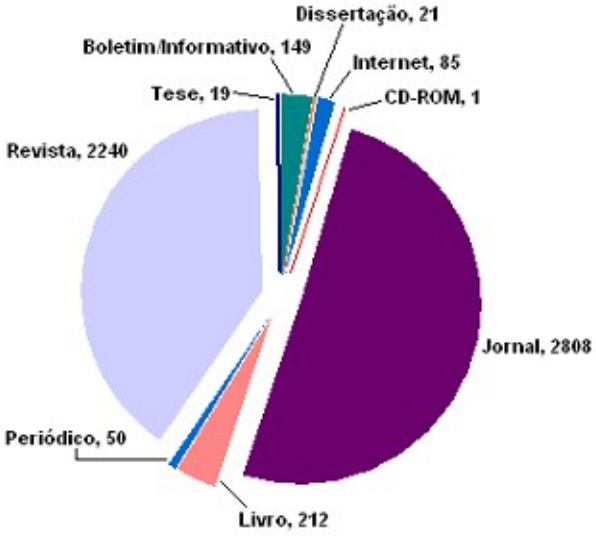


http://www.nilc.icmc.usp.br/lacioweb/

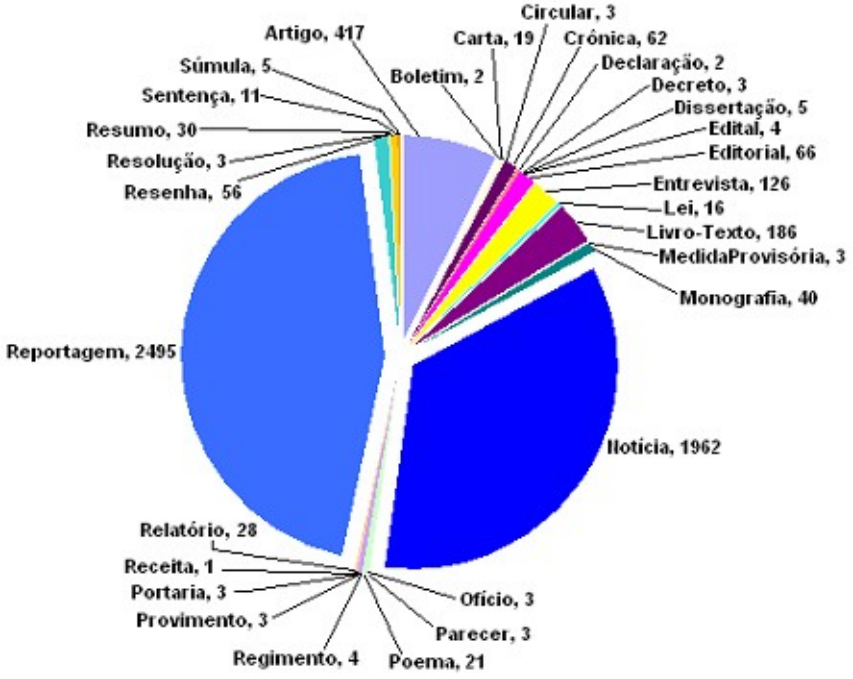
CLASSIFICAÇÃO DE GÊNEROS



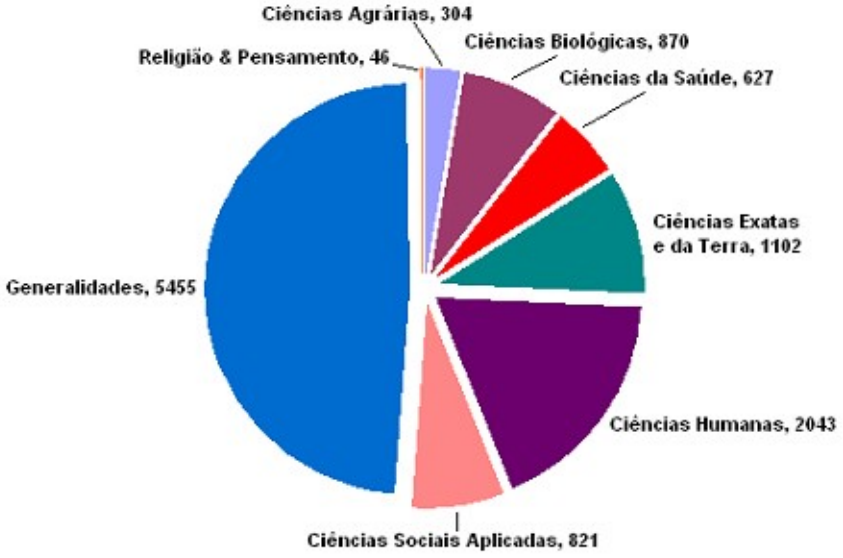
CLASSIFICAÇÃO DE MEIOS DE DISTRIBUIÇÃO



CLASSIFICAÇÃO DE TIPOS DE TEXTOS



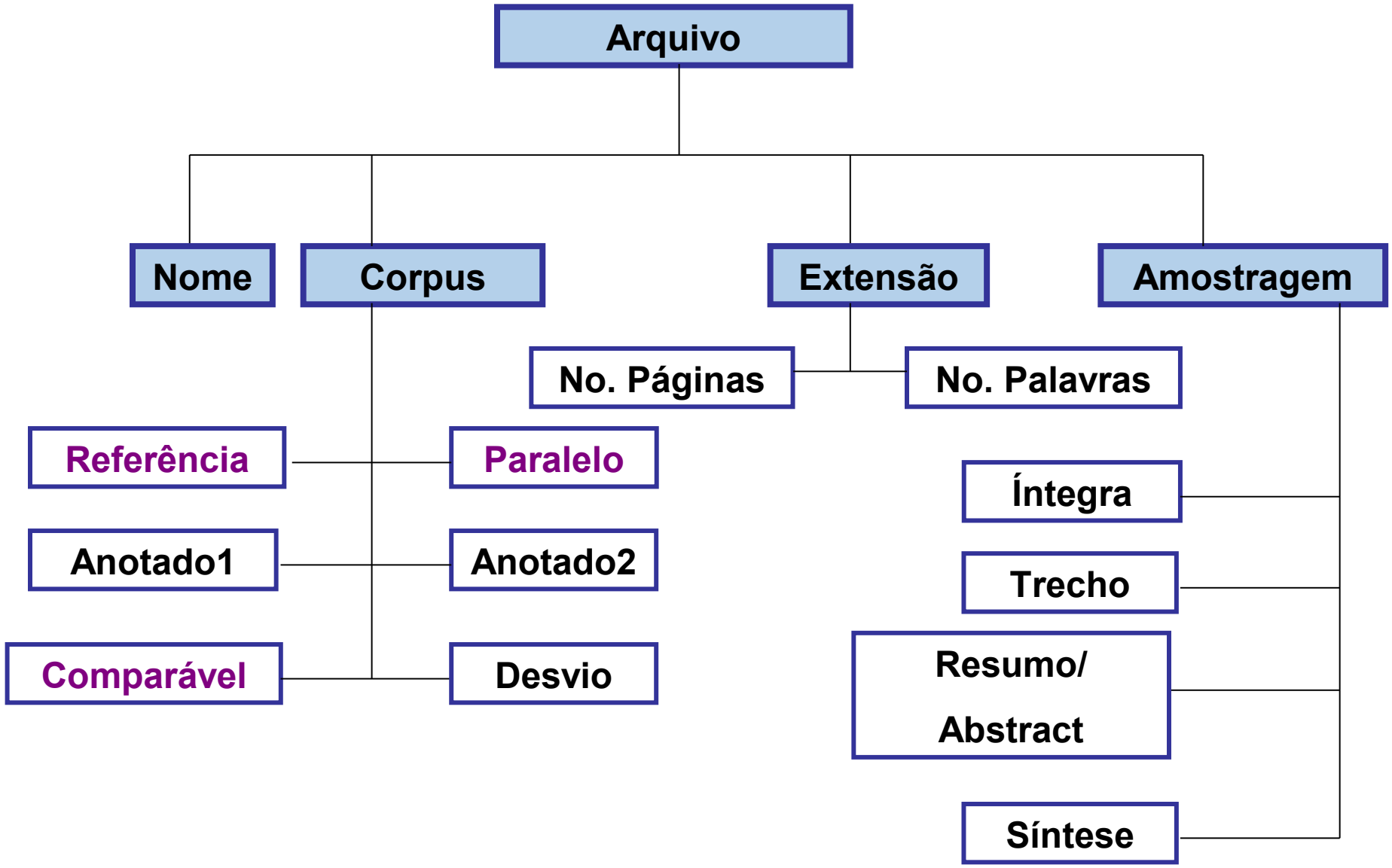
CLASSIFICAÇÃO DE DOMÍNIOS

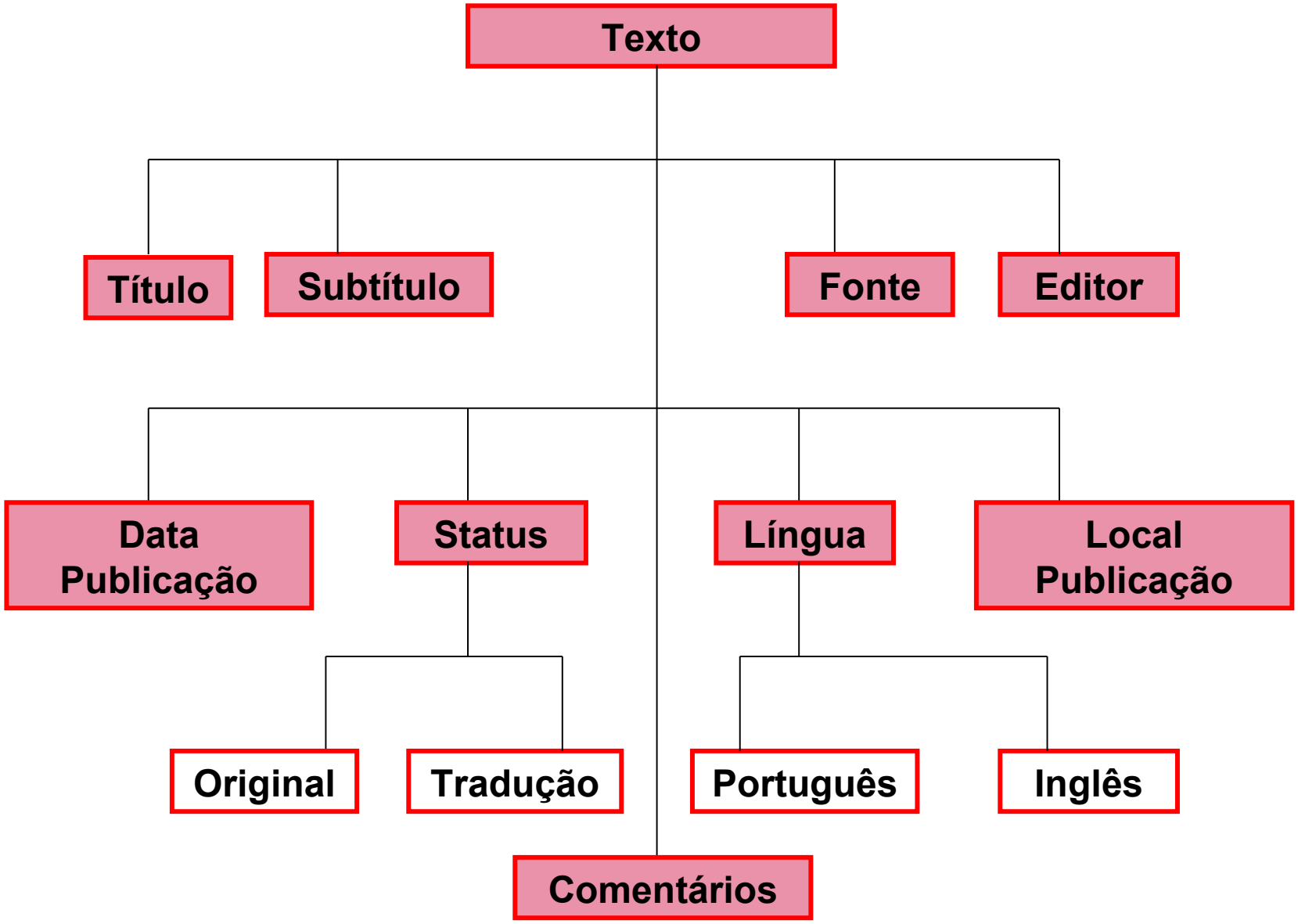


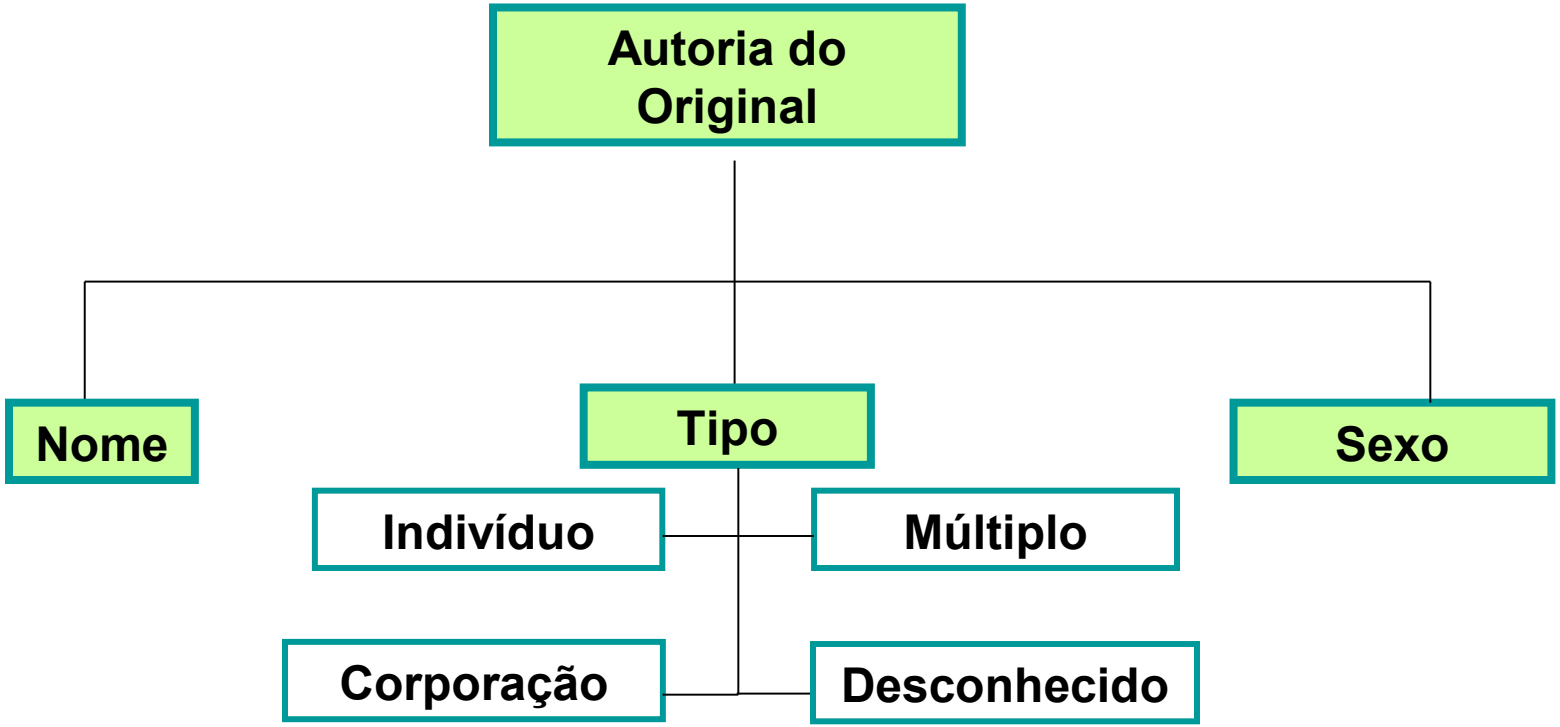
Como fizemos o Portal do LW

- Decisão sobre as informações do cabeçalho
  - dados bibliográficos comuns,
  - dados de catalogação como tamanho do arquivo, tipo da autoria, resumo do texto (se houver), e a tipologia textual
- Nomeação e formatação dos arquivos
- Edição dos cabeçalhos com ajuda de um EDITOR
  - Obtém dados dos nomes (preenche automaticamente alguns campos), ajuda na edição e gera um texto em XML
- Subida dos dados acima na base de dados que dá suporte às buscas

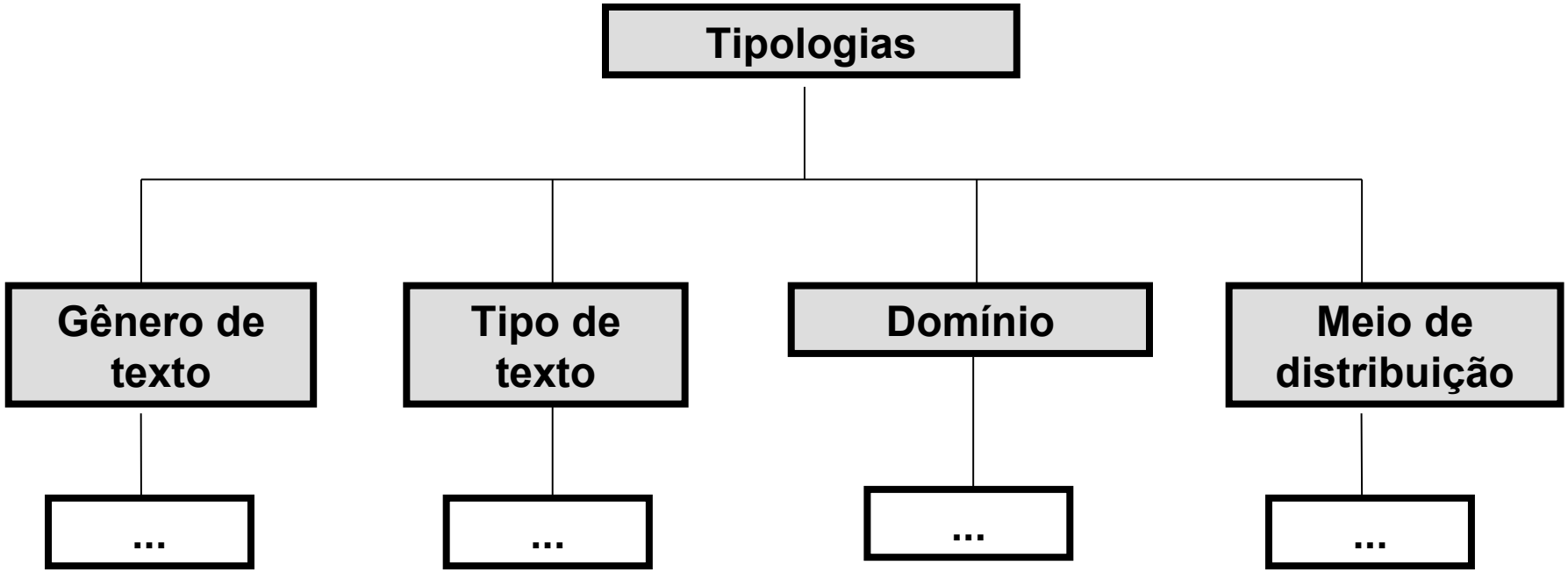
Informações do cabeçalho


Arquivo
Nome
Corpus
Extensão
Amostragem
No. Páginas
No. Palavras
Referência
Paralelo
Anotado1
Anotado2
Comparável
Desvio
Íntegra
Trecho
Resumo/
Abstract
Síntese

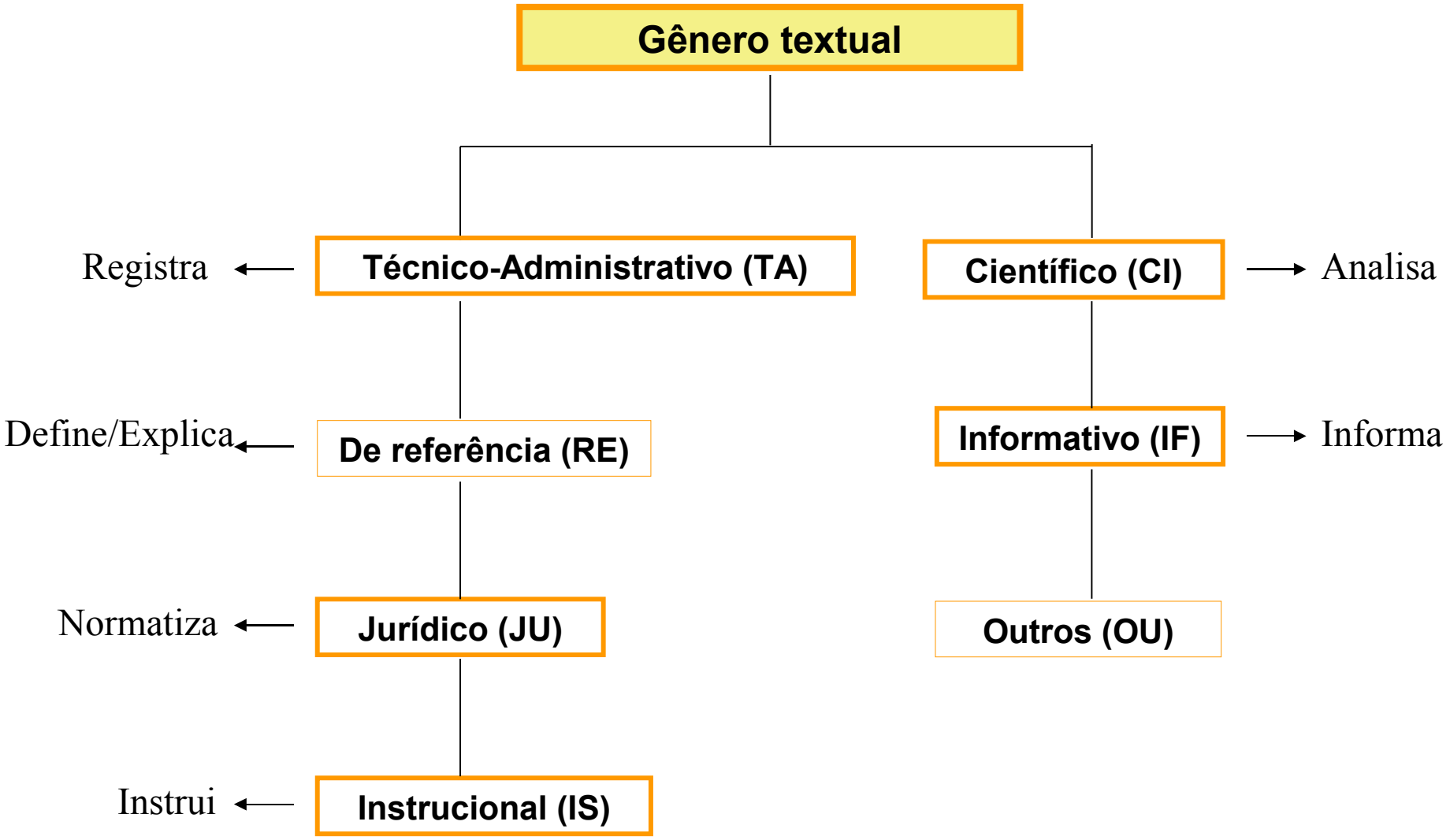

Texto
Título
Subtítulo
Fonte
Editor
Data Publicação
Status
Língua
Local Publicação
Original
Tradução
Português
Inglês
Comentários



OBS: Este esquema deve ser produzido para duas outras categorias de autoria: 1) autor da síntese 2) autor da resenha e 3) orientador. Neste último caso, a categoria “corporação” não deve constar dentre as possibilidades.

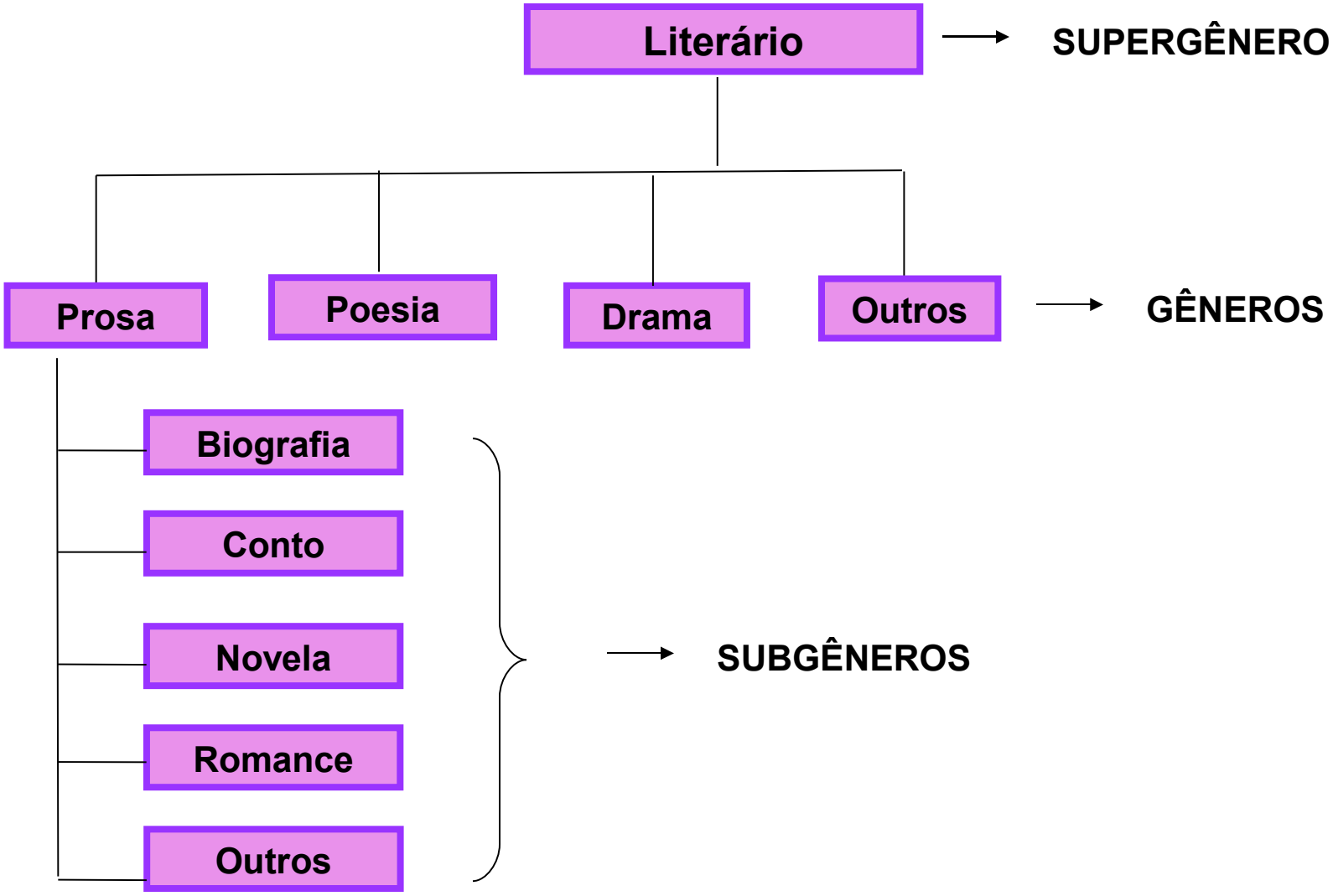

Tipologias
Gênero de texto
Tipo de texto
Domínio
Meio de distribuição
...
...
...
...

Gêneros de texto



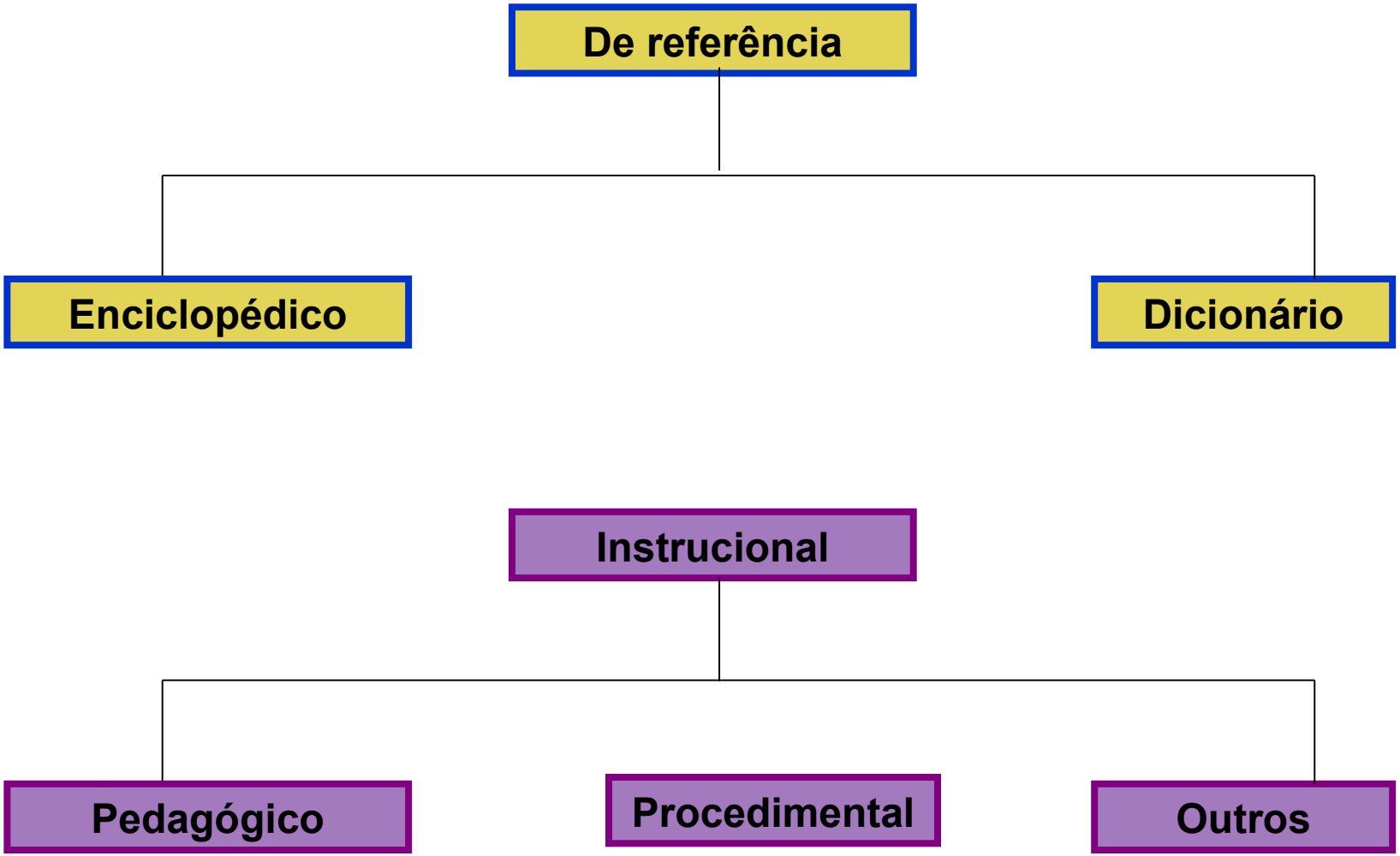
Critérios distintivos: formação do gênero (Bakhtin) / comunidade discursiva (Swales) / Intenção comunicativa

Supergênero, gêneros e subgêneros de texto

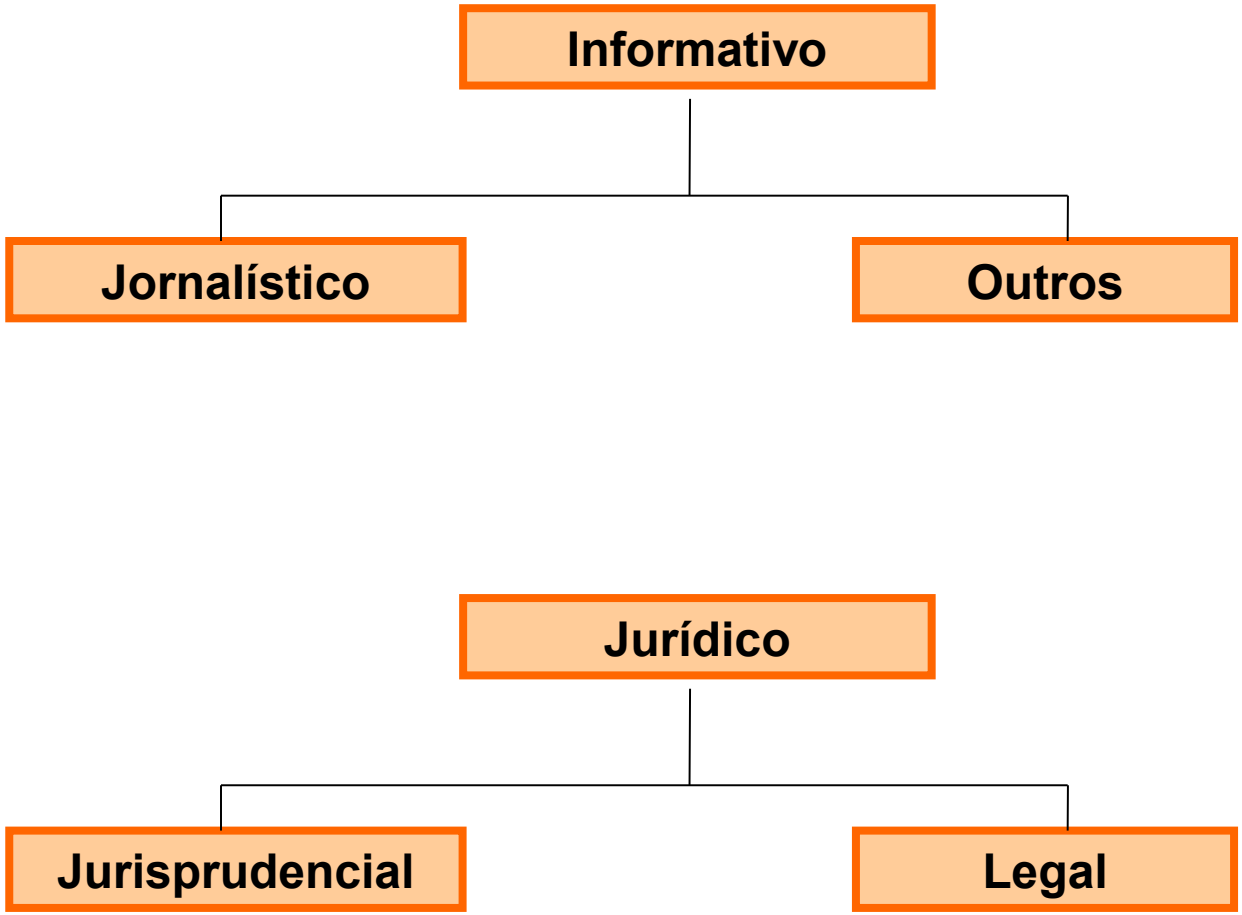


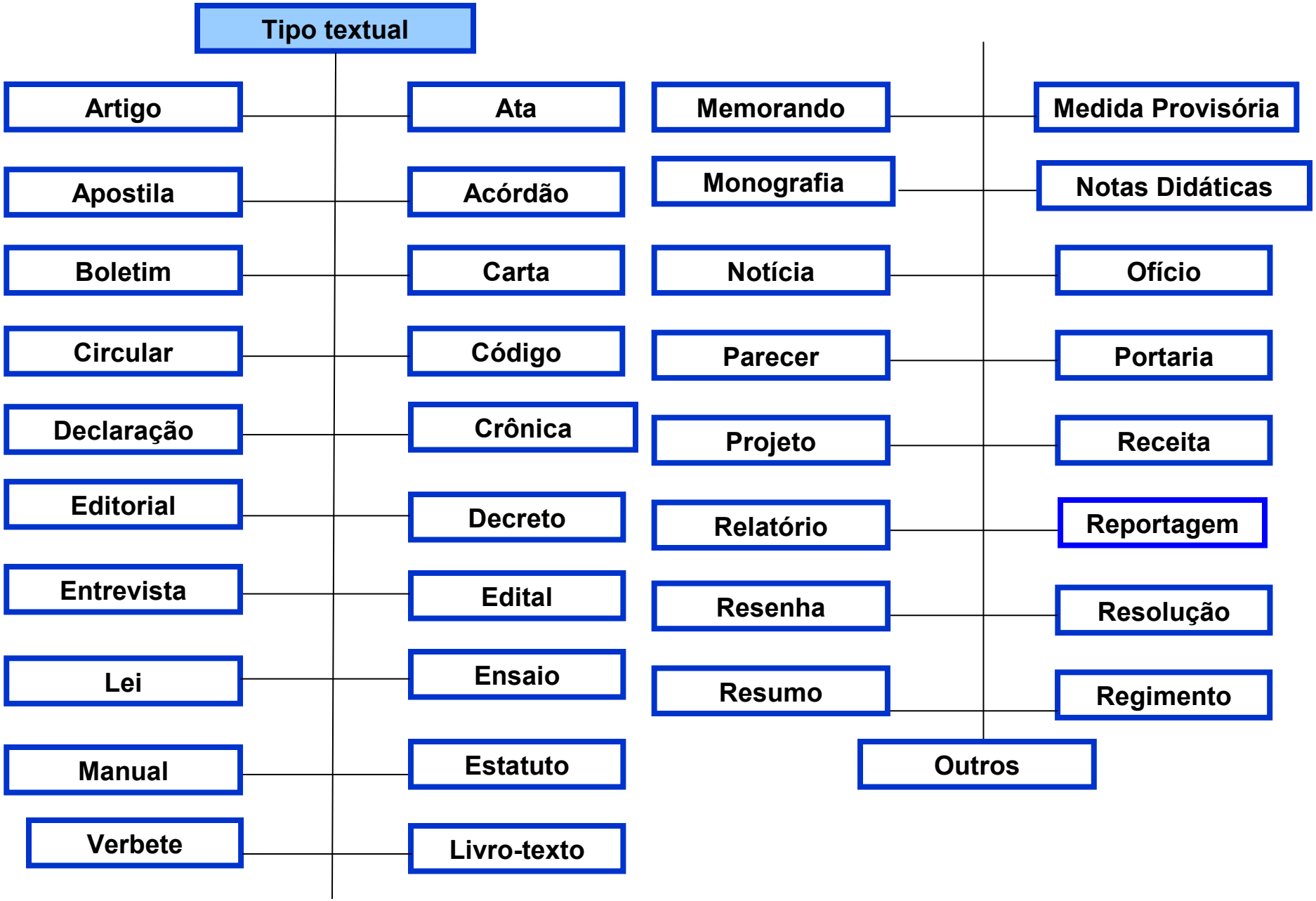
Critérios distintivos: Divisão histórica clássica, já consagrada por teóricos literários

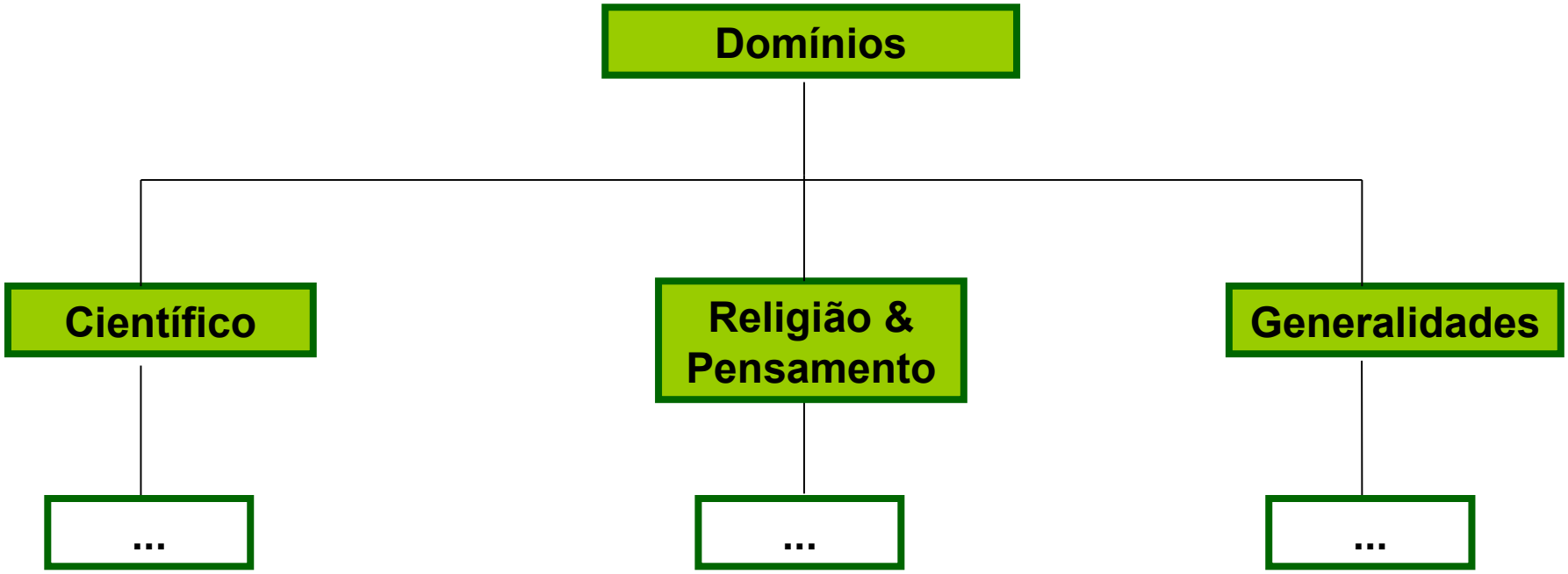
Gêneros e subgêneros de texto



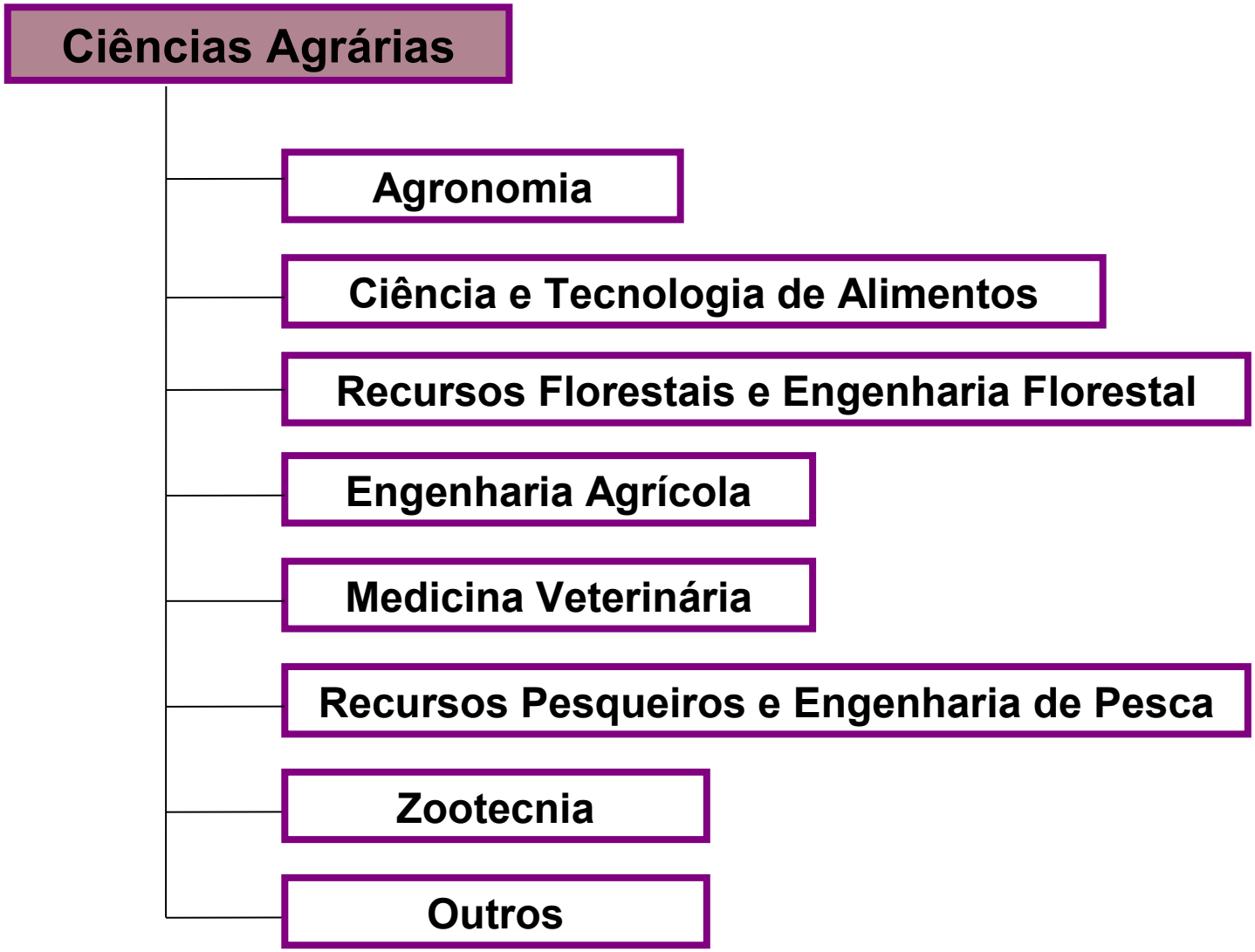
Gêneros e subgêneros de texto



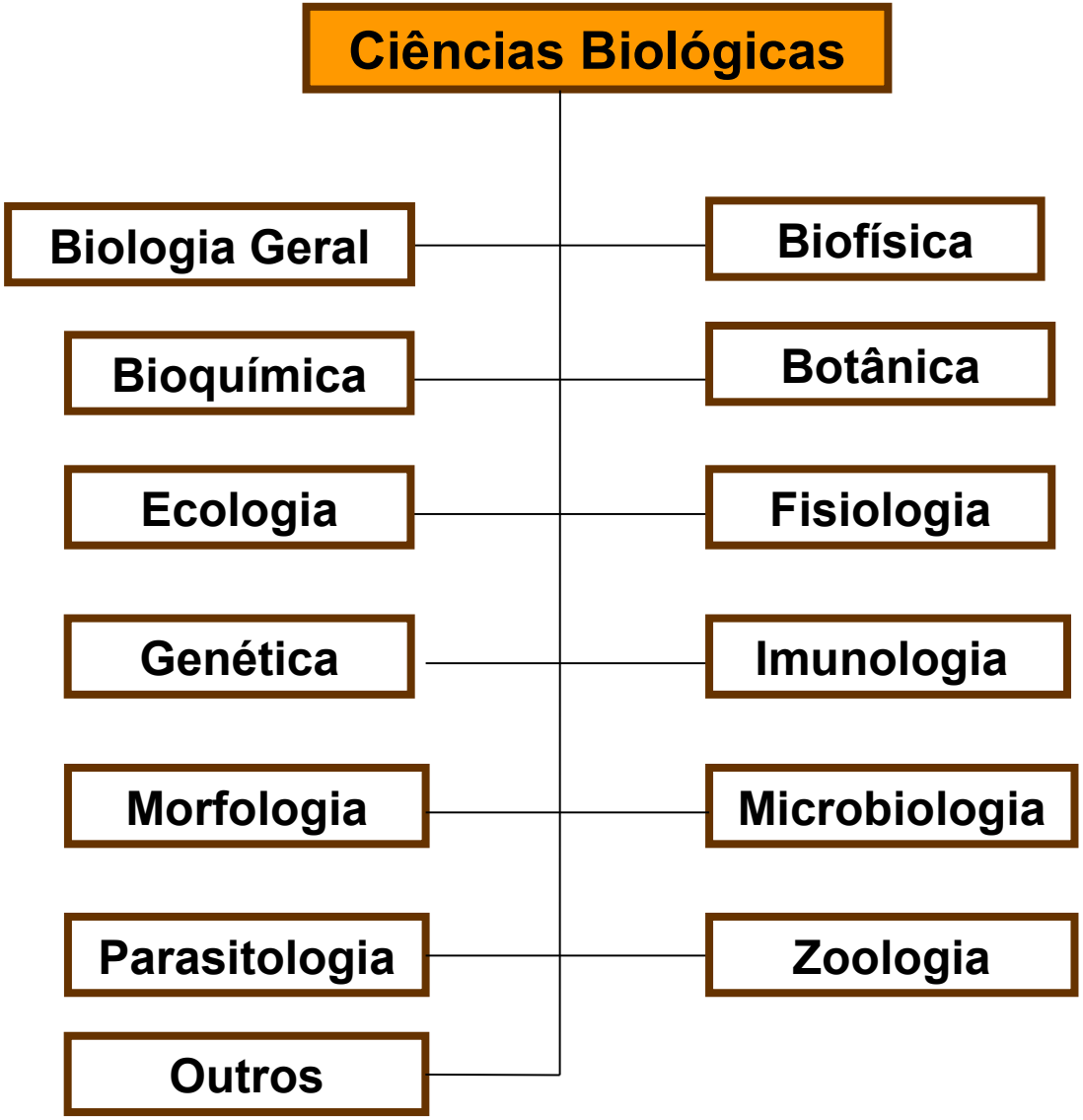

Tipo textual
Artigo
Apostila
Boletim
Circular
Declaração
Editorial
Entrevista
Lei
Manual
Verbete
Ata
Acórdão
Carta
Código
Crônica
Decreto
Edital
Ensaio
Estatuto
Livro-texto
Memorando
Monografia
Notícia
Parecer
Projeto
Relatório
Resenha
Resumo
Outros
Medida Provisória
Notas Didáticas
Ofício
Portaria
Receita
Reportagem
Resolução
Regimento


Domínios
Científico
Religião &
Pensamento
Generalidades
...
...
...

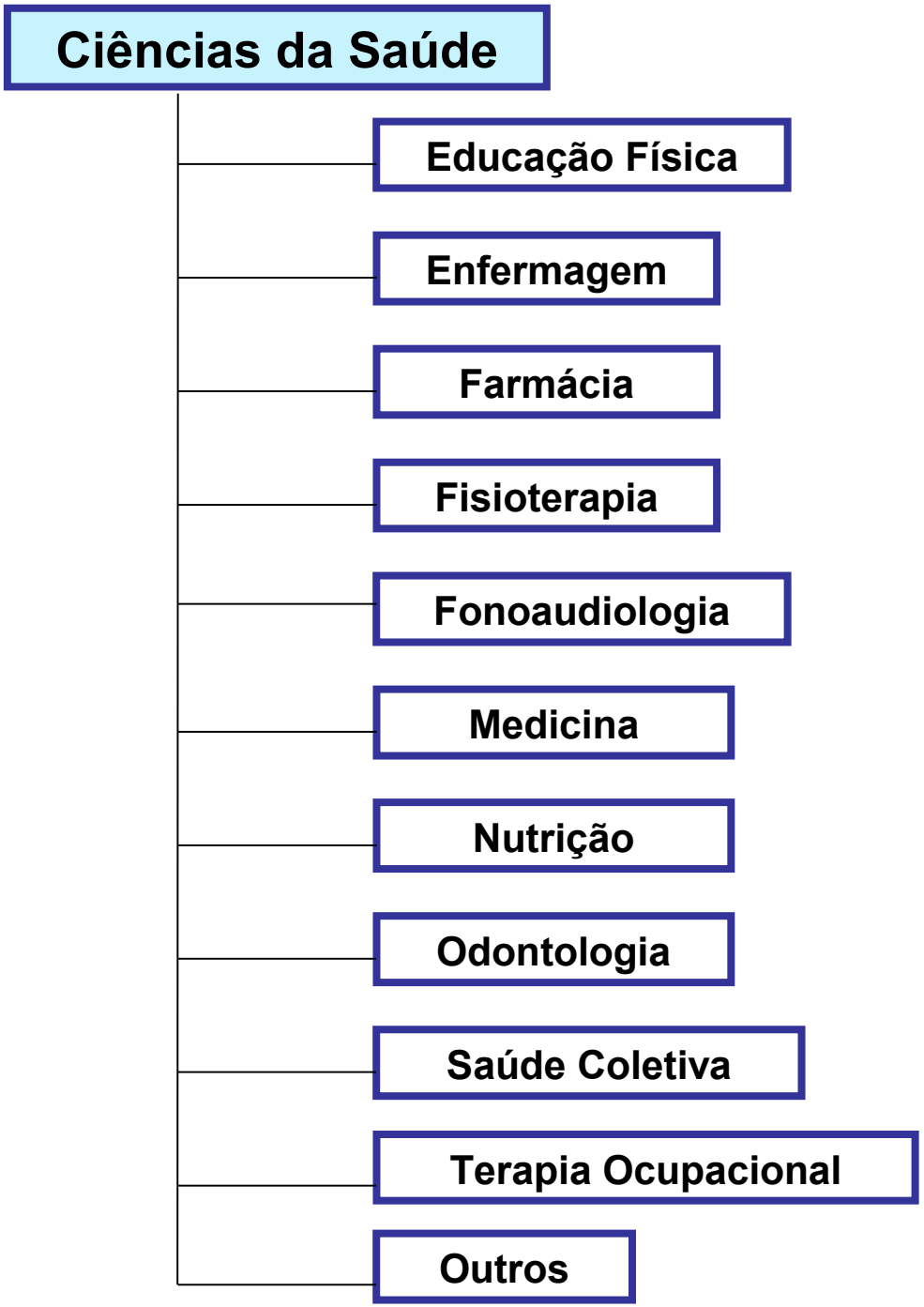
Domínio e subdomínios



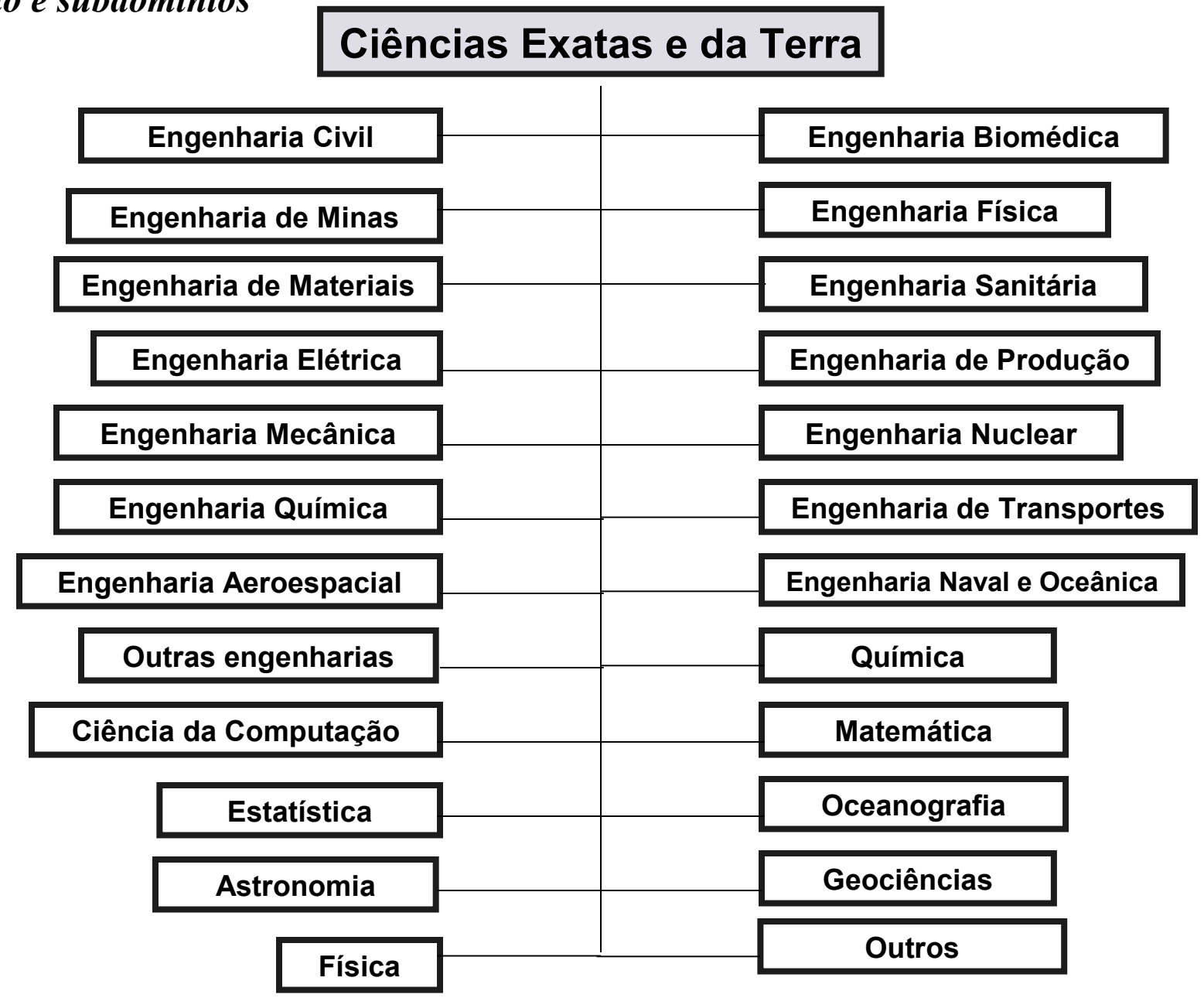
Domínio e subdomínios



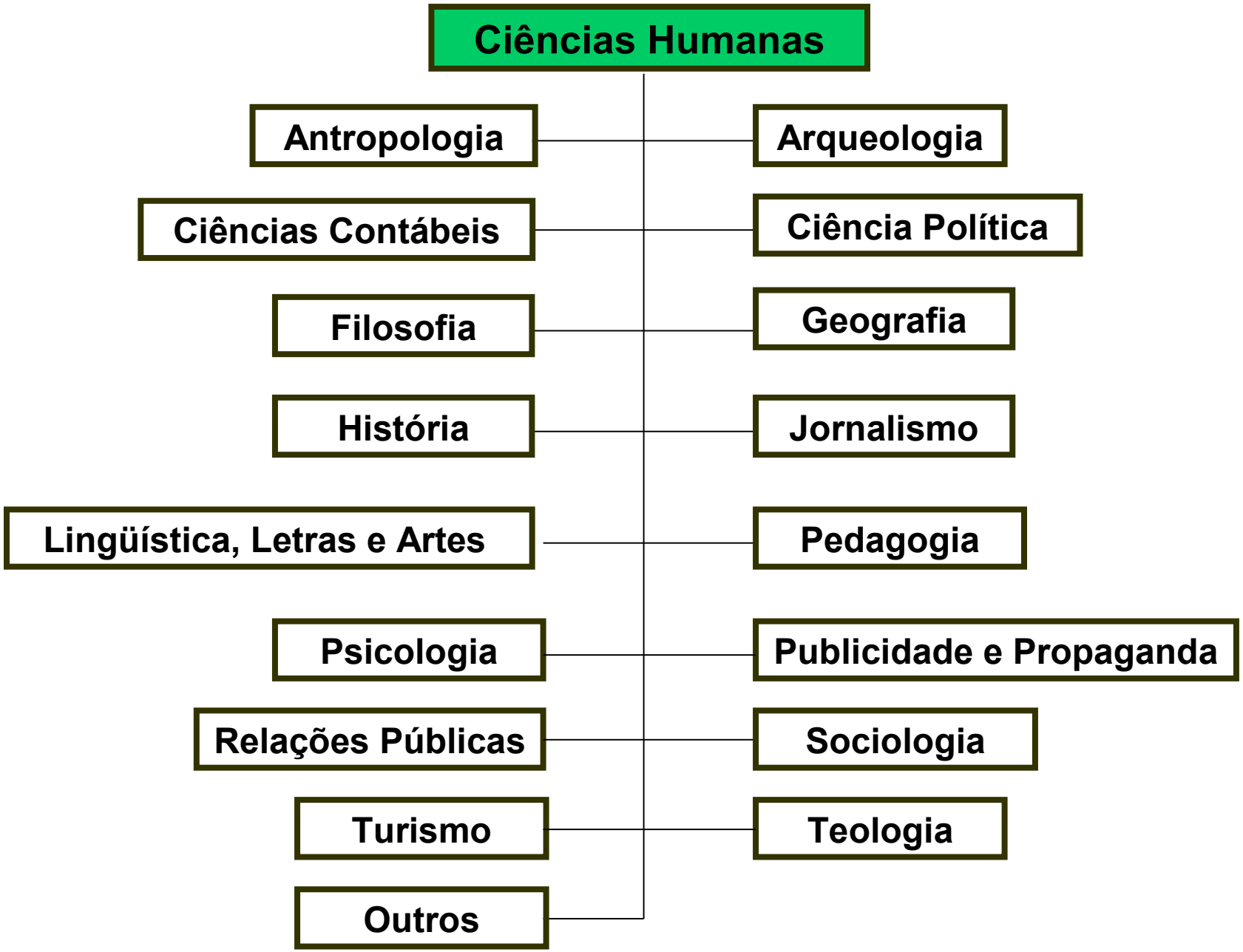
Domínio e subdomínios



Domínio e subdomínios

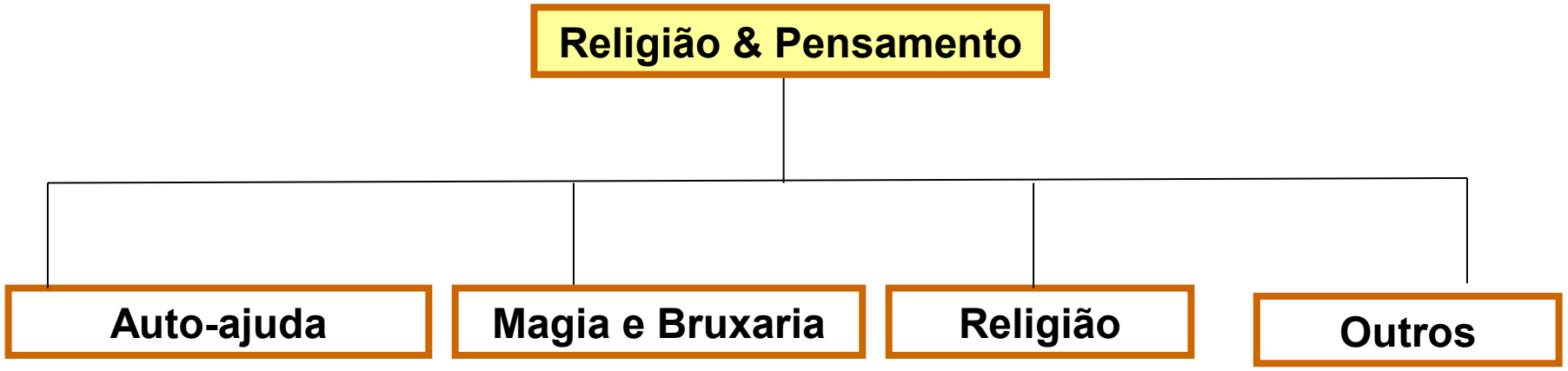


Domínio e subdomínios

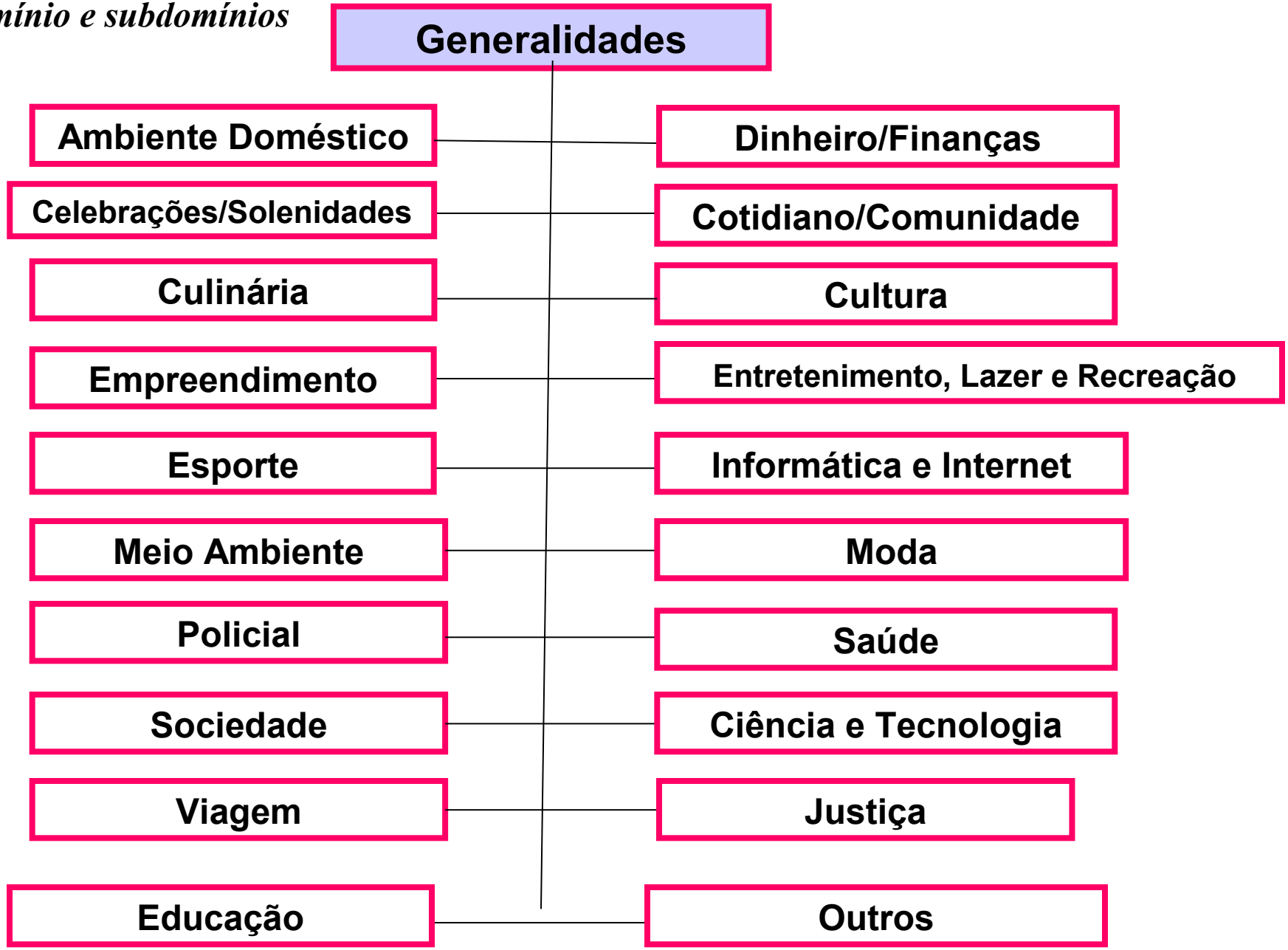


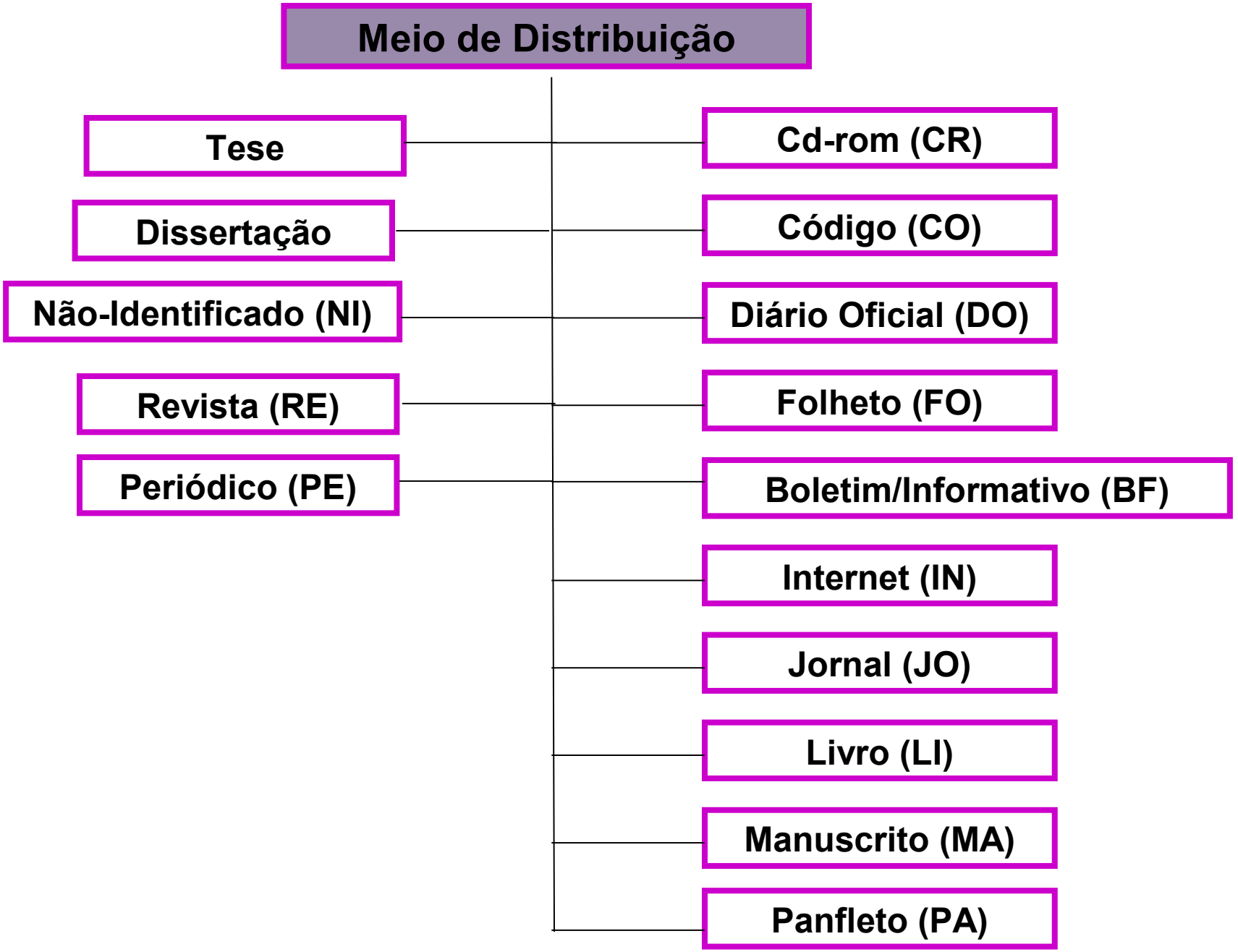
Domínio e subdomínios

Domínio e subdomínios



Domínio e subdomínios

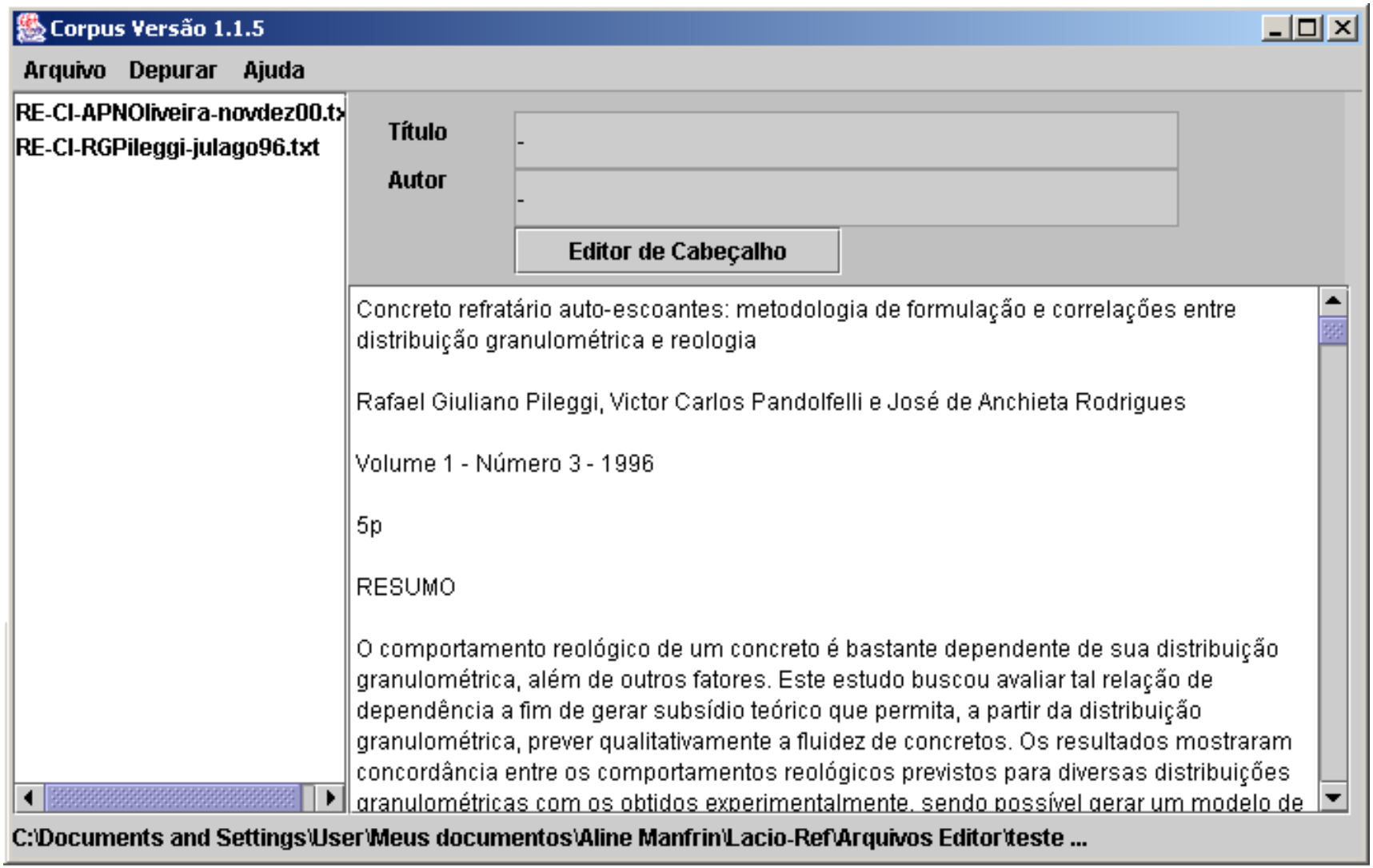


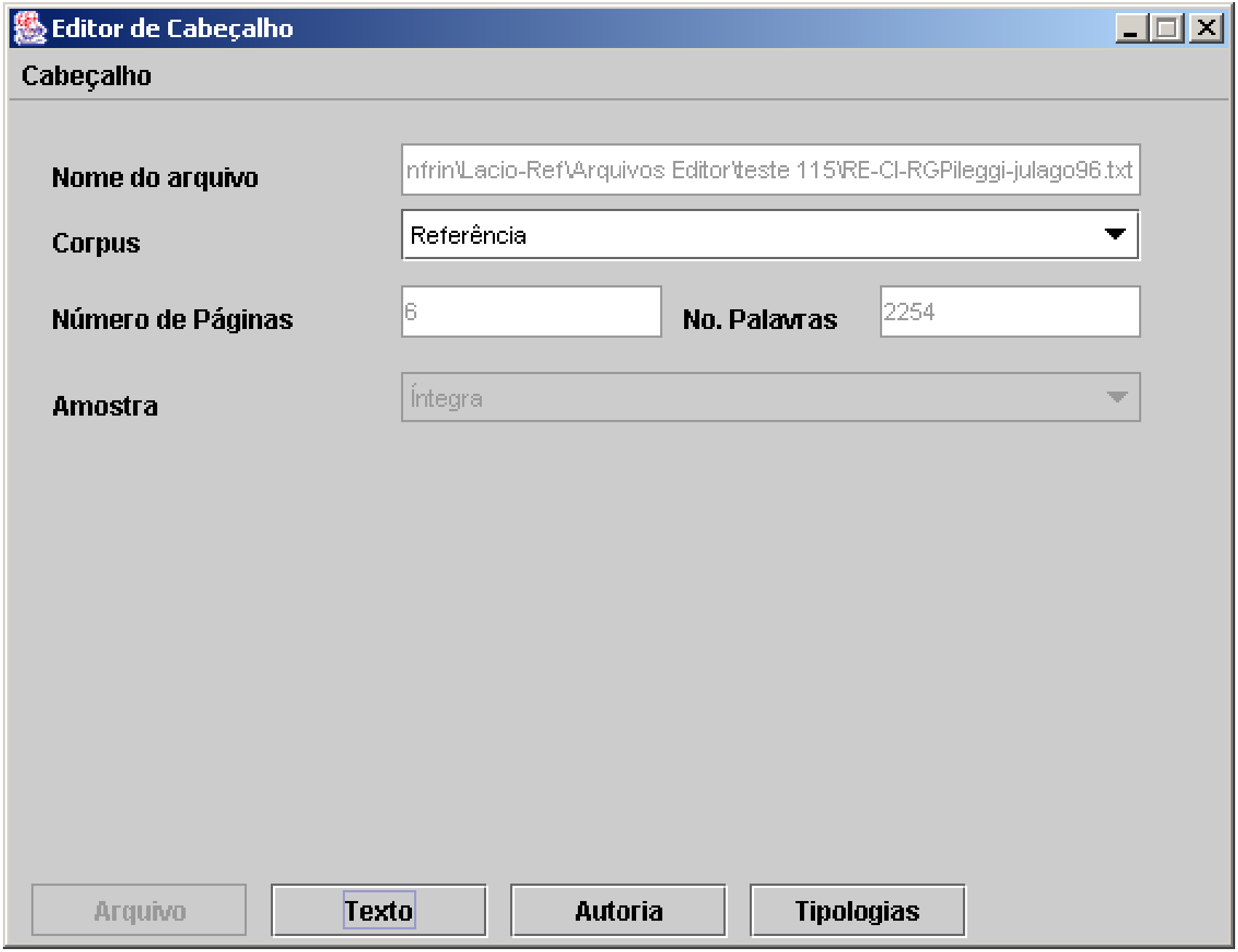

Meio de Distribuição
Tese
Dissertação
Não-Identificado (NI)
Revista (RE)
Periódico (PE)
Cd-rom (CR)
Código (CO)
Diário Oficial (DO)
Folheto (FO)
Boletim/Informativo (BF)
Internet (IN)
Jornal (JO)
Livro (LI)
Manuscrito (MA)
Panfleto (PA)

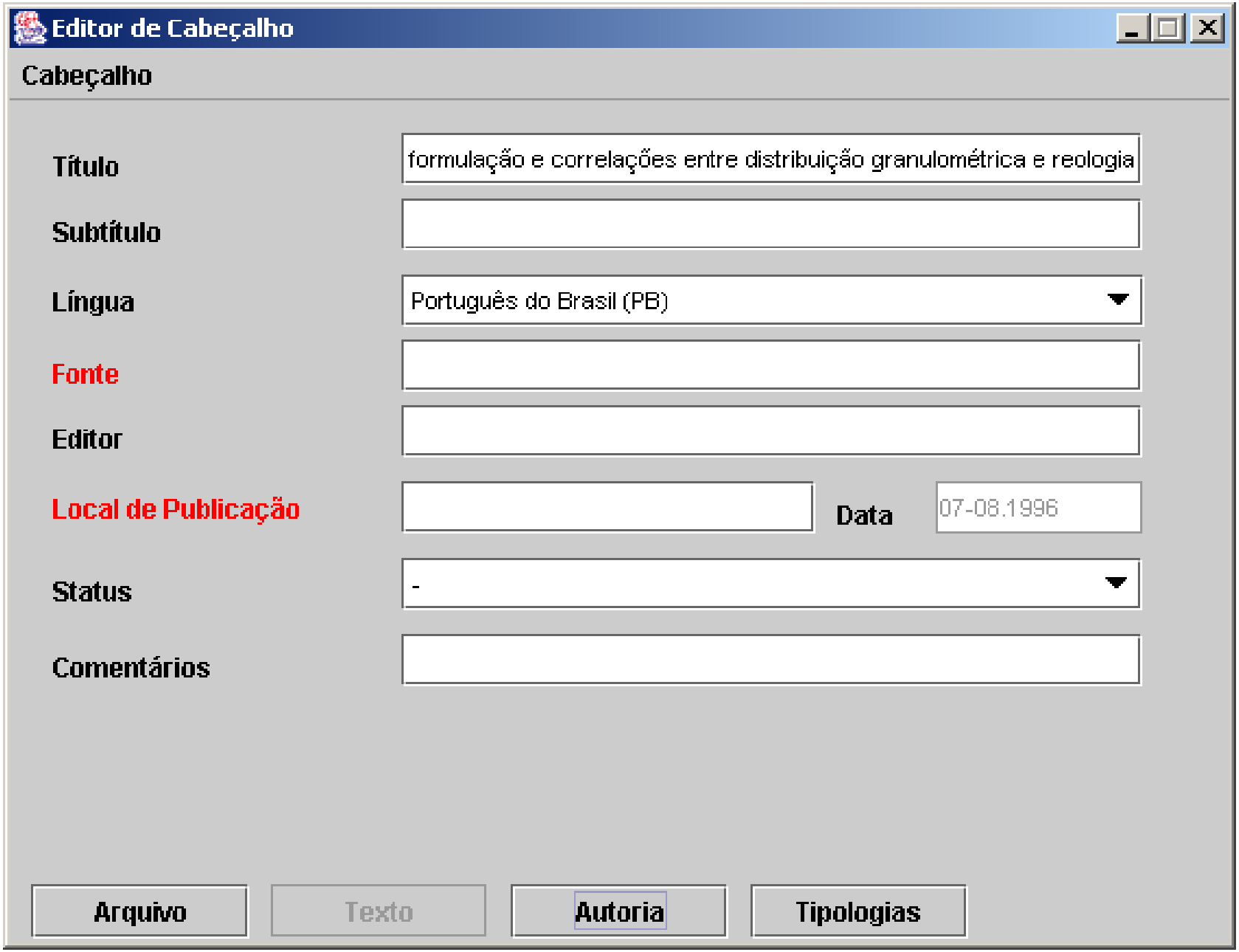
Nomeação e formatação dos arquivos

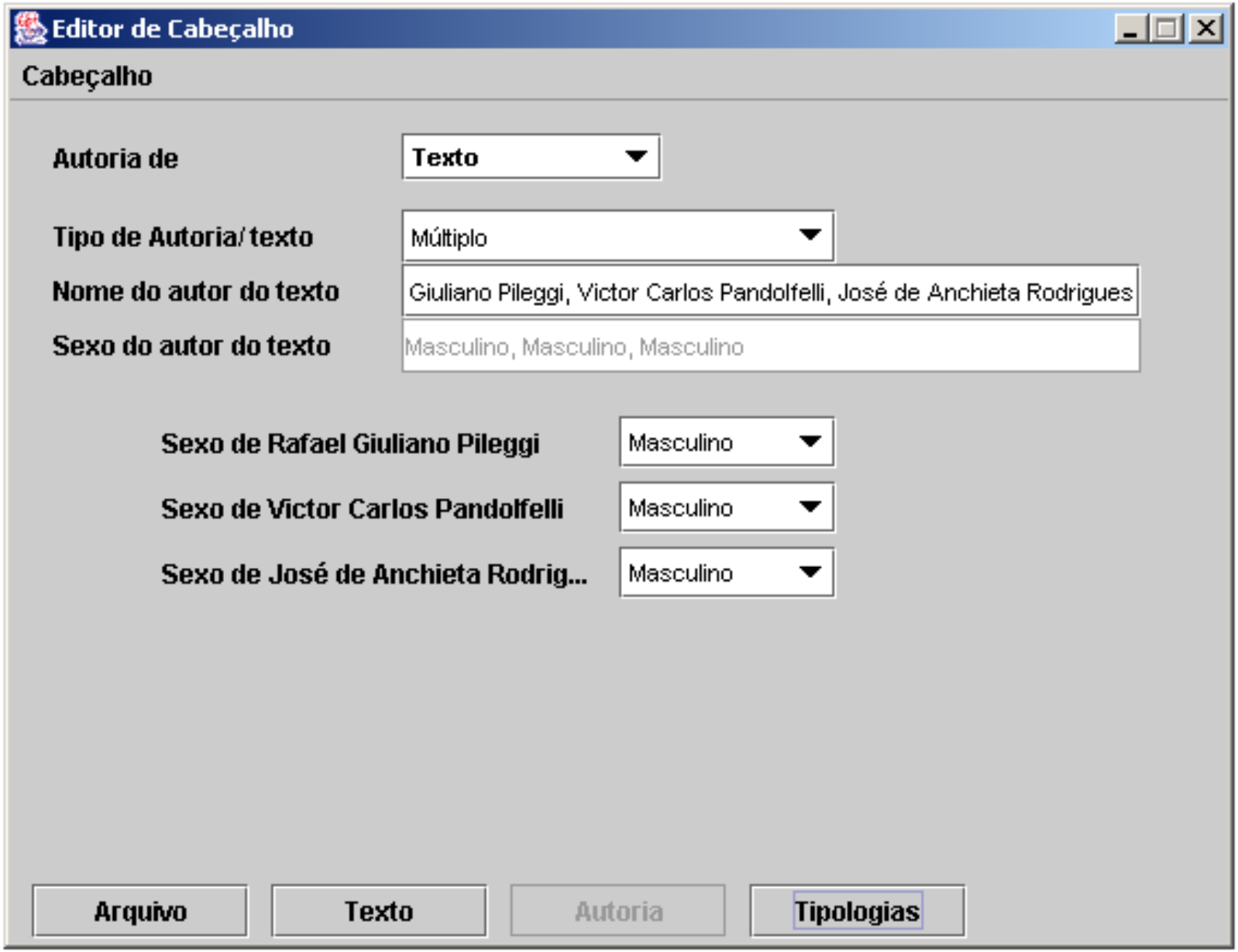
- Nomeação:
  - Meio, super-gênero e gênero e informações particulares do tipo de texto:
  - IN-JU-SE-03mar99.txt
- Formatação
  - Transformação para txt, marcação da presença de elementos gráficos

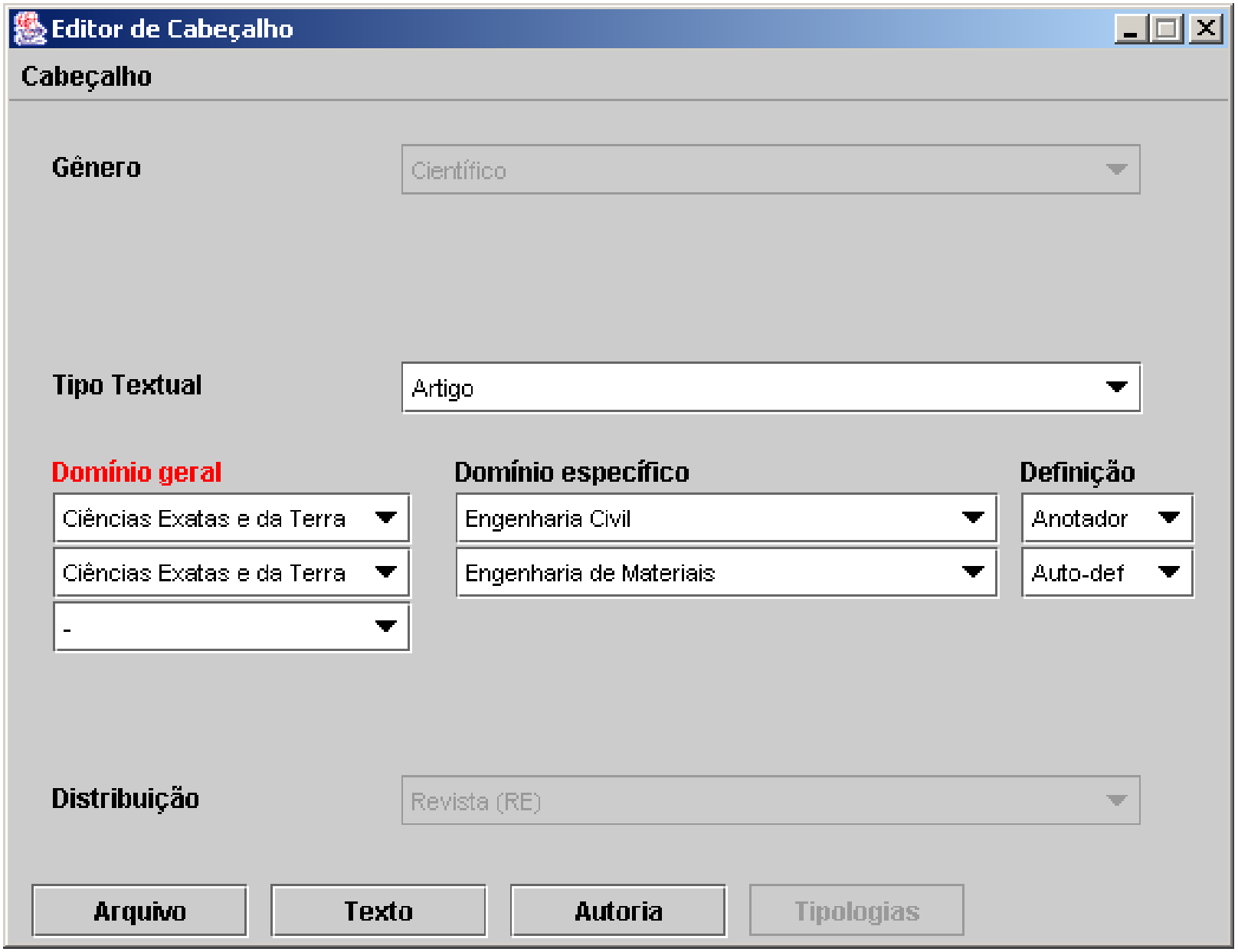
Editor de Cabeçalho – após nomeação e formatação




Editor de Cabeçalho
Cabeçalho
Nome do arquivo
nfrin\Lacio-Ref\Arquivos Editor\teste 115\RE-CI-RGPileggi-julago96.txt
Corpus
Referência
Número de Páginas
6
No. Palavras
2254
Amostra
Íntegra
Arquivo
Texto
Autoria
Tipologias


Editor de Cabeçalho
Cabeçalho
Título
formulação e correlações entre distribuição granulométrica e reologia
Subtítulo
Língua
Português do Brasil (PB)
Fonte
Editor
Local de Publicação
Data
07-08.1996
Status
-
Comentários
Arquivo
Texto
Autoria
Tipologias


Editor de Cabeçalho
Cabeçalho
Autoria de
Texto
Tipo de Autoria/texto
Múltiplo
Nome do autor do texto
Giuliano Pileggi, Victor Carlos Pandolfelli, José de Anchieta Rodrigues
Sexo do autor do texto
Masculino, Masculino, Masculino
Sexo de Rafael Giuliano Pileggi
Masculino
Sexo de Victor Carlos Pandolfelli
Masculino
Sexo de José de Anchieta Rodrig...
Masculino
Arquivo
Texto
Autoria
Tipologias


Editor de Cabeçalho
Cabeçalho
Gênero
Científico
Tipo Textual
Artigo
Domínio geral
Domínio específico
Definição
Ciências Exatas e da Terra
Engenharia Civil
Anotador
Ciências Exatas e da Terra
Engenharia de Materiais
Auto-def
-
Distribuição
Revista (RE)
Arquivo
Texto
Autoria
Tipologias

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- v:1.3.2-->
<text>
<header>
    <title>
        <fileName>IN-JU-SE-03mar99.txt</fileName>
        <corpus>Referência</corpus>
        <nPages>4</nPages>
        <nWords>1171</nWords>
        <sample>Íntegra</sample>
    </title>
    <sourceText>
        <titleOfText>ACÓRDÃO nº 23.048</titleOfText>
        <language>Português do Brasil (PB)</language>
        <source>Supremo Tribunal Regional</source>
        <pubPlace>Natal</pubPlace>
        <date>03.03.1999</date>
        <status>Original</status>
        <comments>Agravo Regimental nº 03-00722/99-0</comments>
    </sourceText>
```

Editor salva no formato XML

```
<author>
    <textAuthor>
        <name>Maria do Perpétuo Socorro Wanderley de Castro, Ezequiel
    Escolástico Bezerra, Nicodemos Fabrício Maia</name>
        <gender>Feminino, Masculino, Masculino</gender>
        <position>Juíza Presidente, Juiz, Procurador do Trabalho</position>
        <type>Múltiplo</type>
    </textAuthor>
</author>
    <textClassification>
        <textGenre>
            <genre>Jurídico</genre>
        </textGenre>
        <textType>Sentença</textType>
        <domain>
            <generalDomain defined="annotador-
    def">Científico/Ciências Sociais Aplicadas</generalDomain>
            <specificDomain>Direito</specificDomain>
        </domain>
        <distribution>Internet</distribution>
    </textClassification>
</header>
```

<body>
<omit desc="head">
ACÓRDÃO nº 23.048

Agravo Regimental nº 03-00722/99-0

Juiz Redator: Ezequiel Escolástico Bezerra

Agravante: Ministério Público do Trabalho

Agravado: Petróleo Brasileiro S/A – Petrobrás

Procedência: TRT 21ª Região/RN

Maria do Perpétuo Socorro Wanderley de Castro
Juíza Presidente

Ezequiel Escolástico Bezerra
Juiz designado para redigir o Acórdão

Nicodemos Fabrício Maia
Procurador do Trabalho
</omit>

Agravo Regimental em Mandado de Segurança. Concessão de liminar em Ação Civil Pública. Pertinência da medida concedida pelo órgão de Primeiro Grau. .....

Base de dados que dá suporte às buscas para montagem de subcórpus

- Todas as informações do cabeçalho são guardadas numa base de dados, permitindo buscas:

  - **Pesquisa Simples:** meio, super-gênero e gênero

  - **Pesquisa Avançada:** além dos campos da Pesquisa Simples, permite buscar por dados de catalogação bibliográficas, como *Nome de Autor*, *Nome do Periódico* e *Caderno*.

  - **Pesquisa Personalizada:** o usuário deve definir detalhadamente o recorte de sua investigação. Campos de seleção como: o *Tipo de Amostragem,* o *Tamanho da Amostra*, o *Tipo de Autoria,* o *Tipo Textual* e o Domínio são apresentados ao usuário, sendo que a grande maioria deles possui conteúdos dinâmicos.

O que aprendemos: o que faríamos diferente

- Nomeação dos arquivos mais simples

- Faríamos um editor de cabeçalho para a Web que após edição do cabeçalho automaticamente colocasse as informações no banco de dados
  - No LW o processo é dividido em 2 etapas pois o editor funciona no Windows e a subida à base é feita com ajuda de programas acionados manualmente

- Avaliaríamos:
  - o Plonetarium da FAPESP (http://plonetarium.incubadora.fapesp.br/portal) para fazer o mesmo Portal do LW com a infra-estrutura e servidores da incubadora de conteúdos digitais (http://incubadora.fapesp.br/)

Proposta para o Projeto do Dic_Hist

- Proposta para o Portal e mão-de-obra necessária

- Proposta para o gerenciamento do projeto: uso de uma Web colaborativa – coteia do ICMC

Portal

- 1) Implementação de um site do projeto com informações sobre o projeto, equipe, publicações, manuais de anotação, etc. em português e também em inglês para ampliar a divulgação da pesquisa.
  - Esse site será integrado o site interno; os dois terão uniformidade de *design* gráfico.
  - 4 meses de pagamento equivalente a uma bolsa ITI A do CNPq

- 2) Modelagem de um banco de dados para guardar os cabeçalhos dos textos do córpus e permitir pesquisas rápidas no córpus.
  - Posterior inclusão dos dados a partir da planilha criada pelo Projeto Resgate e criação da estrutura de diretórios que abrigará fisicamente o córpus no servidor Web.
  - 1 ano de pagamento equivalente a uma bolsa ITI A do CNPq

- 3) Adaptação das interfaces Web de pesquisa e montagem de subcórpus de estudo do projeto Lácio-Web que comporão a parte interna do site do item 1), a princípio restrita aos membros do projeto durante a sua duração (3 anos).
  - Após esse prazo a parte interna ficará disponível a outros pesquisadores. Inclusão das ferramentas básicas como concordanceadores e contadores de freqüência utilizadas no projeto Lácio-Web na parte interna do site do córpus do projeto atual.
  - 4 meses de pagamento equivalente a uma bolsa ITI A do CNPq

- 4) Inclusão das ferramentas desenvolvidas em outros projetos do NILC como lematizadores, geradores de n-gramas, extratores de termos, separadores de sentenças, interfaces com o léxico do NILC que pode ser utilizado como dicionário de exclusão na verificação de arcaísmos na parte interna do site do córpus do projeto atual.
  - Treinamento para os lingüistas do projeto quanto ao uso das ferramentas e interpretação dos dados.
  - 1 ano de pagamento equivalente a uma bolsa DTI H do CNPq

- 5) Anotação semi-automática (feita por ferramentas e corrigida por lingüistas) de uma parte do córpus de textos em português dos séculos XVI, XVII e XVIII (a princípio prevemos que esse córpus anotado tenha 500 mil palavras).
  - Essa tarefa prevê a avaliação e possível adaptação do conjunto de etiquetas utilizadas no projeto Lácio-Web para o córpus de português histórico.
  - Trabalho durante 1 ano de uma equipe contratada a partir do início do projeto que será formada por:
  - um graduado em letras com experiência de 2 anos (pagamento equivalente a uma bolsa DTI G do CNPq
  - 3 alunos da graduação em letras (pagamento equivalente a uma bolsa ITI A do CNPq

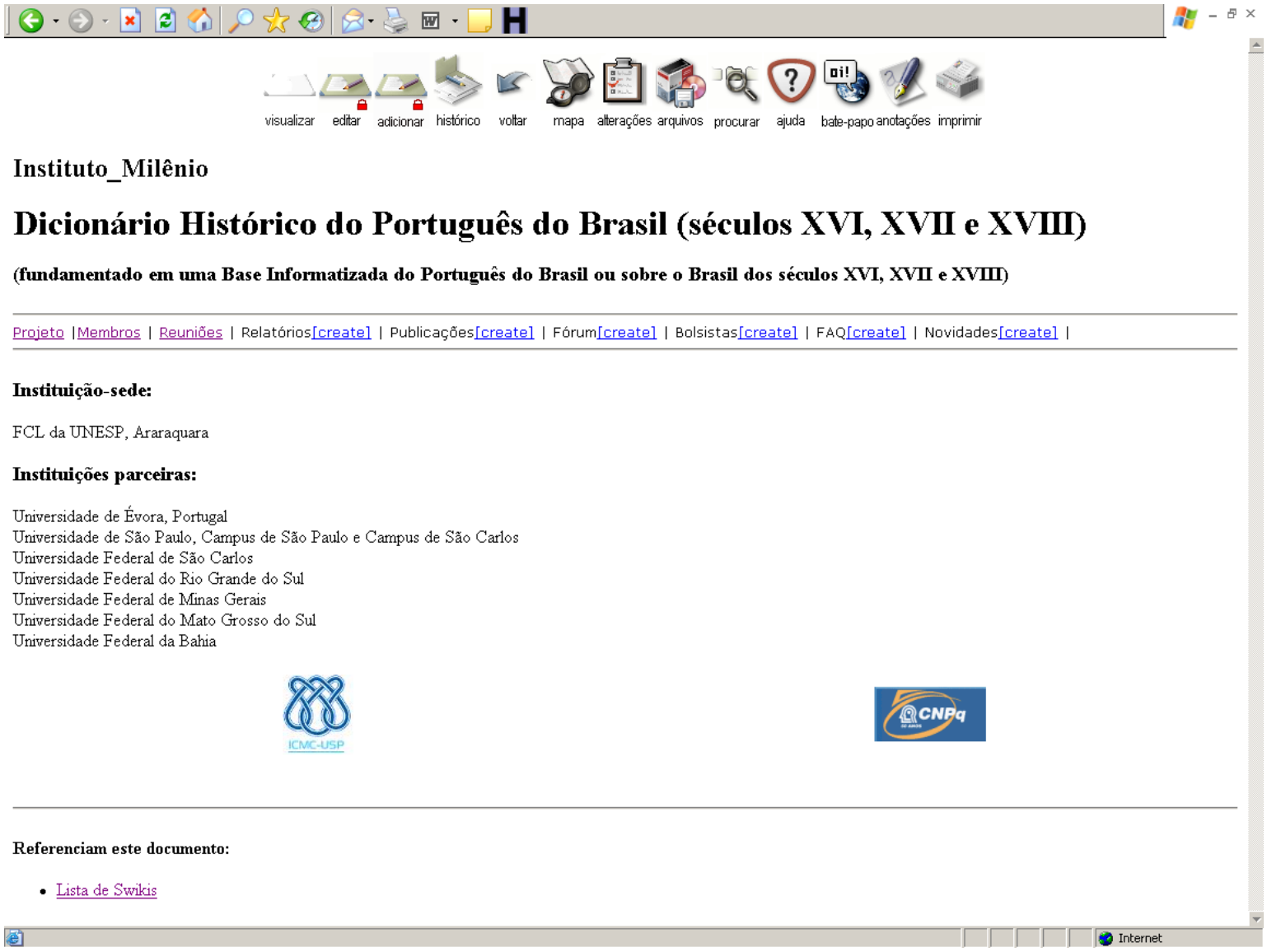
- 6) Treinamento de etiquetadores morfossintáticos com o córpus de 500 mil palavras anotado.
  - Inclusão dos etiquetadores criados na parte interna do site, além de concordanceadores que façam a pesquisa por palavra e etiqueta.
  - 4 meses de pagamento equivalente a uma bolsa ITI A do CNPq

- 7) Codificação do córpus na linguagem de marcação XML.
  - Essa codificação prevê a marcação do cabeçalho dos textos (cujas informações ficarão no banco de dados) e
  - a marcação da estrutura geral - capítulos, parágrafos, títulos e subtítulos, notas de rodapé e elementos gráficos e também a marcação da estrutura de subparágrafos - elementos que são de interesse lingüístico, tais como sentenças, citações, palavras, abreviações, nomes, referências, datas e ênfase.

  - Essa anotação do córpus seguindo padrões internacionais permitirá a sua expansão, facilidade de uso e reuso em outros projetos.
  - De interesse principal aqui estão a marcação dos parágrafos e sentenças para geração de estatísticas sobre os textos do córpus e também abreviações e transcrições ortográficas. Essa anotação explícita dos elementos dos textos permitirá a geração automática de listas de abreviações

    1 ano de pagamento equivalente a uma bolsa DTI D do CNPq

Proposta para o gerenciamento do projeto: Web colaborativa do ICMC



Endereço da Coteia

- http://coteia.icmc.usp.br/coteia/
  - Instituto_Milênio

- http://coteia.icmc.usp.br/coteia/mostra.php