A fronteira entre o texto e a informação estruturada

João Rafael Moraes Nicola
13 de agosto de 2013

Estrutura do texto e da informação

- As técnicas de redação de textos estruturados visam a eficiência e a eficácia do processamento **humano** do conteúdo do texto.
- No entanto, o aumento do volume de textos produzidos torna imperativo o processamento **automatizado** de textos.
- Como fazer o computador **entender** os artifícios usados na redação de textos estruturados?

O que é um documento do Word?

- **Para um ser humano**: pode ser uma lei federal, com ementa, artigos, incisos, anexos, etc.
- **Para um sistema de informação**: uma sequência de caracteres e comandos de formatação.

Extraindo informações estruturais: Parser

- Reconhecimento das partes que formam a estrutura de uma norma:
  - epígrafe, ementa, preâmbulo,
  - articulação,
  - assinatura,
  - anexos, etc.
- Reconhecimento da estrutura hierárquica da articulação
- Reconhecimento de blocos de alteração
- Reconhecimento de tabelas, fórmulas e figuras

Extraindo relações intra e inter-textuais: Linker

- Reconhecimento de remissões externas
- Reconhecimento de remissões internas (ainda não implementado)
- Acréscimo automático de “hyperlinks”
- Em uso no site LexML para documentos do Senado Federal, da Câmara dos Deputados e do STF.

Formato XML: representando a informação reconhecida

- O projeto LexML possui uma especificação de formato de arquivo baseado no padrão XML para representar todas as propriedades de documentos estruturados
- A padronização do formato permite a construção de várias ferramentas de processamento: LexEdit, LexComp, ferramentas de compilação, consolidação, etc.

A volta: Renderer

- O consumidor final desta informação é o ser humano
- O “Renderer”, ou “Renderizador” faz a conversão do documento próprio para consumo da máquina (XML), em formatos próprios para consumo do ser humano, como: formatos de processadores de texto (rtf, doc, odt), publicação e impressão (PDF, HTML), arquivo (PDFA-3).

Na prática ...

http://linker.lexml.gov.br/lexml-parser/parse/static/simulador/simulador.html