Capítulo 13

RENA - reconhecedor de entidades

José João Dias de Almeida

Diana Santos e Nuno Cardoso, editores, *Reconhecimento de entidades mencionadas em português: Documentação e actas do* HAREM*, a primeira avaliação conjunta na área*, Capítulo 13, p. 157–172, 2007.

O RENA (Alves e Almeida, 2006) é um protótipo de sistema de extracção/marcação de entidades mencionadas construído por Edgar Alves sob supervisão de J.J. Almeida no âmbito do projecto IKF.

O projecto IKF (Information + Knowledge + Fusion) (Silva, 2004; Oliveira e Ribeiro, 2003; Tettamanzi, 2003) foi um projecto Eureka (E!2235) envolvendo participantes universitários e industriais de seis países, cuja finalidade básica foi o desenvolvimento de uma infraestrutura distribuída baseada em ontologias para o manuseamento inteligente de conhecimento – contemplando um ambiente documental multifonte e distribuído.

O IKF *framework* baseia-se num modelo de representação de conhecimento sofisticado (baseado em ontologias, facetas, lógica vaga (*fuzzy*), informação incompleta, e raciocínio temporal) (Silva, 2004), e é constituído por um conjunto de módulos envolvendo, entre outros:

1. Extractores básicos – extracção de conhecimento a partir de documentos heterogéneos de modo a construir um sistema de assimilação documental:
   - organização de um conjunto de ficheiros de modo a construir uma base documental
   - extracção de informação (rica) a partir desse conjunto de documentos
   - classificação facetada, *fuzzy* vaga e parcial de documentos e da informação neles contida
   - fusão da informação extraída dos vários documentos
2. Renovador de conhecimento (*Knowledge Renovator*) (Oliveira e Ribeiro, 2003) – ligada à evolução (temporal ou não) da informação e do conhecimento.
3. Enfermaria do Conhecimento – ligado a sistemas legados, e à reparação de inconsistências por razões variadas.
4. Navegadores – um conjunto de navegadores sobre a base de conhecimento e a base documental.

A título de exemplo de aplicação considere-se o caso da assimilação documental de caixas de correio electrónico: ao extrair e fundir conhecimento, pretende-se obter informação capaz de responder a perguntas como:

- quem é a pessoa F?
- qual a lista dos amigos de F? quais os parceiros de X?
- qual o conjunto de áreas de interesses de Y?
- que documentos são relevantes acerca de Z?

Tendo em vista estes objectivos, para além das tarefas principais (as tarefas estruturais ligadas ao projecto), foi realizado um conjunto de pequenas tarefas/experiências exploratórias, envolvendo recursos muito limitados e frequentemente envolvendo alunos finalistas.

É neste contexto que surge o protótipo RENA que, não fazendo directamente parte do projecto IKF, foi desenhado como um caso de estudo com a intenção de fazer extracção de conhecimento simples – extracção de uma base de entidades:

$$Rena : Ficheiro^* * BaseEnt \longrightarrow BaseDoc * BaseEnt * \ldots$$

13.1 Descrição do RENA

Na sequência do enquadramento anteriormente descrito, o protótipo RENA tem como intenção uma extracção tão rica quanto possível de informação, com vista a ser usada por sistemas de processamento e fusão de conhecimento (e em particular no projecto IKF).

À medida que a ferramenta RENA foi sendo projectada, decidiu-se que era importante que pudesse ser usada por um conjunto menos restritivo de aplicações – ou seja, que pudesse ser usada em modelos semânticos menos sofisticados (um Micro-IKF).

Dum modo resumido o RENA é um sistema REM constituído por:

- Uma biblioteca Perl:
  1. baseada num conjunto de ficheiros de **configuração** alteráveis,
  2. com funcionalidade para **extrair a lista das entidades** a partir de conjuntos de textos,
  3. ou, em alternativa, **marcar entidades** num conjunto de texto.
- Um conjunto de programas para fazer processamento de entidades.

Muita da capacidade de extracção depende de um conjunto de ficheiros e de regras – elementos de configuração – que descrevem conhecimento geral e regras de contexto usados na extracção.

Pretendeu-se desde o início que esses elementos de configuração fossem *externos* ao RENA, de modo a que o utilizador os pudesse adaptar à sua visão do mundo e à sua intenção concreta de utilização. Assim, foi requisito dos elementos de configuração que fossem legíveis, expressivos e compactos.

13.1.1 Estrutura interna do RENA

Do ponto de vista algorítmico, o RENA:

1. começa por procurar entidades e construir uma sequência de textos simples e entidades: $(texto \times entidade)^*$
2. seguidamente, esse objecto é processado por uma série de filtros com assinatura

$$f : (texto \times entidade)^* \rightarrow (texto \times entidade)^*$$

que vão processar os pares texto-entidades, enriquecendo a informação, alterando fronteiras e unindo zonas, com base nos recursos de configuração atrás referidos e utilizando ferramentas internas ou externas (como por exemplo o analisador morfológico jspell (Simões e Almeida, 2002; Almeida e Pinto, 1995)).

3. no final, de acordo com a saída pretendida, é criado:
   - um texto com as entidades anotadas
   - um resumo das entidades presentes

   O formato final pretendido pode ser:
   - *XML*, contendo uma versão do texto original onde são anotadas todas as referências a entidades encontradas.
   - *YAML* (Ben-Kiki et al., 2005, 2006), descrevendo todas as entidades com alguma referência no texto, bem como todas as classificações atribuídas.

Os filtros que gerem texto nos formato acimas referidos, que, aliás, podem ser desactivados, fazem tarefas como:

- tratamento de entidades com elementos de uma única letra,
- tratamento de aspas ligado às entidades
- remoção de entidades entre aspas (este filtro só deverá ser usado se se pretender ignorar este tipo de ocorrências).
- tratamento de entidades com traços interiores (por exemplo, *Benfica-Sporting*)
- tratamento de entidades em início de frase
- enriquecimento por análise de regras de contexto
- enriquecimento por análise do almanaque de nomes
- enriquecimento por análise do almanaque de cultura geral
- tratamento de acrónimos
- reconhecimento e unificação de entidades iguais (ou abreviadas) e criação de atributos de ligação entre as várias ocorrências da mesma entidade.

13.1.2 Ficheiros de configuração

A configuração de base do RENA é constituída por um conjunto de recursos:

1. Ontologia de classes – que estabelece relações (hierárquicas) entre os tipos de entidades existentes;
2. Tabela de contextos – com regras para deduzir qual o tipo das entidades com base no contexto esquerdo;
3. Almanaque de cultura geral – onde se registam termos/conceitos geográficos, culturais, patrimoniais, cultura geral;
4. Sistema de tratamento de nomes – em que se guardam alguns dos nomes/apelidos mais comuns e regras para determinar se um nome próprio se refere a pessoas;
5. Tabela de conversão/adaptação de nomes;
6. Tabela de contextos atributivos (em fase de construção).

Vários destes recursos são definidos usando linguagens de domínio específico (DSL) construídas com a intenção de conseguir uma descrição eficaz dessa informação.

Seguidamente vamos detalhar alguns destes recursos e apresentar alguns exemplos.

Ontologia de classes

A ontologia de classes armazena os tipos de entidades e respectivas relações. A definição dos tipos de entidades e dos seus relacionamentos é uma actividade delicada, sensível: corresponde a uma descrição do nosso modo de ver o mundo. Há zonas desta ontologia que são facilmente reutilizáveis, outras que são dependentes do projecto concreto.

Normalmente é importante ter controlo total sobre esta ontologia pelo que ela deve ser construída manualmente. No entanto, alguma zonas podem ser obtidas por aprendizagem automática.

No nossos exemplos vimos que pode haver utilidade em usar (pequenos) extractos de ontologias como o CDU, o tesauro da Unesco, o tesauro da Biblioteca de Alexandria, ou outros sistemas classificativos.

A existência deste recurso é crucial para se conseguir:

- fazer inferência parcial de tipos de entidades,
- facilitar a fusão de análises complementares,
- obter uma maior adaptabilidade da informação extraída.

```
- pessoa:
   - advogado
   - arquitecto
   - atleta:
      - futebolista
      - nadador
   - escritor:
      - poeta
   - jornalista
   - militar:
      - general
      - almirante
      - brigadeiro
      - sargento
      - tenente
      - capitão
   - músico:
      - compositor
      - pianista
      - trompetista
   - político:
      - presidente da república
      - deputado
```

Figura 13.1: Extracto da ontologia de classes.

Sempre que possível pretende-se que esta ontologia tenha um grão fino de modo a poder registar toda a informação extraída, mas ao mesmo tempo deseja-se que permita uma posterior abstracção/síntese.

A dimensão e conteúdo da ontologia de classes deverá ter em conta a pragmática e o conteúdo e dimensão do conjunto documental em análise. No caso concreto, utilizamos uma ontologia exemplo com cerca de 120 classes. Na Figura 13.1 representa-se um extracto da ontologia de classes (visto como uma taxonomia para mais fácil visualização).

Saliente-se mais uma vez que a ontologia para descrever as classes difere conforme a intenção e o conjunto de documentos em análise. Por exemplo, embora haja muitas coisas comuns, há uma enorme diferença entre o conjunto das classes referentes a um arquivo de biologia, a um arquivo de etnomusicologia, ou a um arquivo de *software* de PLN.

Tabela de contextos

A tabela de contextos permite que de um modo compacto se possa definir uma associação entre uma **expressão de contexto** esquerdo e uma classe (ver Figura 13.2).

```
  cidade (de do da)     => cidade !lctx
  freguesia (de do da)  => freguesia
  distrito (de do da)   => distrito
  concelho (de do da)   => concelho/90
  estado (de do da)     => estado
  capital               => cidade !lctx

  (Rio Oceano Lago Mar Serra Cordilheira) => $_
  Cabo (do de da)       => cabo
  Golfo (do de da)      => golfo

  (Lugar Largo Lg. Praça Rua R. Avenida) (de da do das dos)?  => lugar
  (Travessa Beco Quinta Viela Rotunda) (de da do das dos)?    => lugar

# Monumentos $

  (Convento Mosteiro Igreja Ig. Palácio Museu Sé) (de da)? => monumento
```

Figura 13.2: Extracto da tabela de contextos.

Note-se que:

- as regras podem ter valores de confiança, de modo a permitir distinguir entre indícios mais fortes e indícios mais fracos,
- a grafia maiúscula é usada para indicar se o termo de contexto esquerdo deverá ou não ser incluído na entidade,
- os padrões das regras podem incluir variantes alternativas, elementos opcionais, comentários, etc.

Embora esta tabela possa ser construída, consolidada e revista manualmente, uma boa base de início pode ser obtida através da extracção dos bigramas de palavras do contexto direito e do início de entidade (das entidades antes ou depois de classificadas) – podendo ser usadas técnicas de *bootstraping* habituais em situações idênticas[1].

Muitas regras são gerais; no entanto, no caso geral, esta tabela depende do problema concreto.

Almanaque de cultura geral

Conforme atrás se referiu, o almanaque de cultura geral pretende guardar alguma informação de cultura geral de índole diversa.

[1] No estado actual do RENA, há apenas um esqueleto de ferramentas de ajuda à construção dessa tabela segundo o método referido.

```
Rio Douro =
 rio Douro
        IOF =>  rio
        AFLUENTES =>    rio Mau,
                rio Sousa,
                rio Varosa,
                rio Tâmega,
                rio Pinhão,
                ....
                rio Torto,
                rio Távora,
                rio Esla,
                rio Tua
        COMPRIMENTO =>  927
        FOZ =>  Porto
        IN =>   Portugal,
                Espanha
        NASCE =>        serra do Urbião
        PASSA_EM =>     barragem do Pocinho,
                barragem de Miranda,
                barragem de Crestuma,
                Miranda do Douro,
                barragem do Carrapatelo,
                Régua,
                barragem da Bemposta
```

Figura 13.3: Extracto da informação existente no almanaque de cultura geral.

Presentemente este almanaque tem por base informação criada no âmbito do projecto T_2O (Almeida e Simões, 2006a,b), e a informação associada a cada entidade é por vezes rica (ainda que heterogénea): além duma classe de base, pretende-se armazenar um conjunto de atributos e ligações tão rico quanto possível.

Simplificadamente este almanaque corresponde a uma vista sobre a projecção de uma ontologia T_2O, seleccionando-se os termos por exemplo referentes a geografia, personagens famosas, ou eventos.

Na Figura 13.3 mostra-se um extracto da informação existente no almanaque associada a **Rio Douro**, demonstrando a intenção de dispor de um conjunto de dados de base rico e estruturado que permita processamento posterior (interactivo ou não).

Sistema de tratamento de nomes

A intenção subjacente ao **sistema de tratatamento de nomes**, demonstrado na Figura 13.4, é permitir dispor de dados para determinar se certos identificadores constituem (ou não)

```
26.62287        Maria   nome
13.70273        Ana     nome
6.85846         José    nome
5.16030         Silva   apelido
4.90977         António nome
3.95357         Carla   nome
3.51606         Manuel  nome
3.50263         João    nome
...
0.02148         Dinis   misto
```

Figura 13.4: Extracto do sistema de tratamento de nomes.

prováveis nomes de pessoas (quando não houver fortes indícios noutro sentido).

De um modo simplificado, guarda-se uma tabela que indica a taxa de ocorrência (por milhão de palavras) de determinada palavra, indicando ainda se o seu uso é preferencialmente nome, apelido ou misto. Esta tabela tem por base uma lista de 150.000 nomes completos, de várias proveniências.

Tabela de conversão/adaptação de nomes

Dado que há necessidade de poder usar ontologias de classes e tabelas de contextos adaptadas a cada projecto concreto, temos necessidade de criar mecanismos para conversão de classes.

Esta tabela pretende criar um grau de indirecção de modo a permitir uma mais fácil alteração da estrutura da ontologia de classes, criando alguma independência entre a ontologia de classes, o almanaque e a tabela de contextos.

Tabela de contextos atributivos

O objectivo da tabela de contextos atributivos é, para além de eventualmente inferir classes, ajudar a inferir mais atributos, factos e informações acerca das entidades – numa palavra, informação mais rica.

Considere-se o seguinte extracto exemplo:

```
a atleta portuguesa A :: atleta(A), nacionalidade(A,portuguesa)
X , no norte de Y     :: geo(X), geo(Y), norte(X,Y)
o francês Z           :: pessoa(Z), nacionalidade(Z,francês)
```

Quando for encontrada uma ocorrência do tipo **...a atleta portuguesa Rosa Mota ...** é feita a inferência de que Rosa Mota é uma atleta (e portanto uma pessoa, etc), e que o atributo nacionalidade da entidade em causa é preenchido com o valor **portuguesa**.

Esta tabela é crucial para aumentar a riqueza da informação extraída. Até ao momento, ela tem sido construída manualmente, no entanto há planos para a construção de ferramentas que proponham regras e extraem pistas a partir de textos anotados.

13.2 Participação no HAREM

A participação no HAREM foi muito importante e produtiva para nós já que:

- envolveu discutir e trocar impressões com os outros participantes e com a organização,
- envolveu lidar com um problema para o qual o RENA não tinha sido pensado,
- levantou uma série de questões que nunca nos tinham ocorrido referentes à necessidade de criação de camadas de adaptação de notações e de adaptação de estruturas classificativas.

Há, no entanto, alguma diferença entre o tipo de avaliação que pretendíamos (mais ligada a um uso de extracção de informação enciclopédica) e a avaliação feita no HAREM.

Os resultados finais ficaram aquém do que seria possível por várias razões:

- um dos autores do RENA (Edgar Alves) não participou (por ter já deixado a universidade)
- houve decisões do RENA que não seguem as propostas do HAREM e das quais não quisemos prescindir,
- com o pouco tempo que nos foi possível dedicar ao RENA, optámos por melhorar alguns módulos que, não sendo os mais importantes para a avaliação no HAREM, são cruciais para o RENA.

Genericamente a identificação de entidades foi bem conseguida apesar de termos optado por não marcar valores numéricos em geral por nos parecer menos interessante para o RENA.

Os maiores problemas resultaram de uma diferente filosofia no que diz respeito às classes – diferente filosofia semântica. Enquanto que o HAREM pretende marcar a ocorrência específica em contexto específico, o RENA está menos preocupado com a ocorrência concreta mas com a entidade referida; está mais preocupado com a extracção de informação rica de cariz enciclopédico.

Considere-se o seguinte exemplo concreto:

```
(...) os diários "<OBRA TIPO="PRODUTO" >Jornal Tribuna de Macau</OBRA>" e
<OBRA TIPO="PRODUTO">Macau Hoje</OBRA> (...)
```

De acordo com a nossa intenção de extracção de informação enciclopédica, afirmar que o *Jornal Tribuna de Macau* é uma `OBRA:PRODUTO` seria completamente inaceitável: a resposta útil para o RENA (independentemente de o termos conseguido extrair) é **Jornal** ou **Jornal diário**.

Do mesmo modo demos preferência a **monumentos** em relação aos `LOCAL:ALARGADO` ou às `OBRA:ARTE`.

A participação do RENA na tarefa de classificação semântica foi feita da seguinte forma:

1. extrair a informação e usar apenas a classificação geral de acordo com a ontologia RENA,
2. traduzir (de acordo com uma tabela de tradução escrita manualmente) cada classificador RENA num par categoria:tipo do HAREM.

Esta abordagem também introduziu erros adicionais. Por exemplo, algumas classes, como monumento, acabaram por não encontrar um classificador natural na estrutura classificativa do HAREM.

Optámos por não fazer a tarefa de classificação morfológica por não nos parecer tão relevante para a nossa ferramenta específica e para não dispersar (e congratulamo-nos com a versatilidade do sistema HAREM de poder aceitar marcações parciais).

No próxima secção apresentamos mais alguns exemplos e situações em que os modelos HAREM e RENA divergiram.

13.3 Subsídio para a discussão sobre futuras edições

A organização e planeamento do HAREM foi muito boa. No entanto e tendo em conta futuras organizações vou enunciar algumas coisas que me parece ser vantajosas.

Em resumo, as propostas para futuras versões são:

1. uso de documentos seguindo (totalmente) a norma XML
2. uso claro e extensível de metadados nas colecções

$$coleccao = (MetaData \times Texto)^*$$

3. migração de taxonomia 2 níveis para uma ontologia de classes multi-nível
4. uso de etiquetagem mais versátil.

13.3.1 Uso de documentos seguindo XML

A migração para documentos XML, torna mais fácil tirar partido de um conjunto de ferramentas no sentido de:

- permitir verificar se os documentos (colecções e submissões) são bem-formados e se são válidos,
- ser claro e definido qual o sistema de encoding usado,
- poder obter mais facilmente uma variedade de vistas (*pretty-printers*), resumos, e reordenações dos documentos, de modo a se adaptar a diversas finalidades. (Usando CSS, XSL, etc.),
- ser trivial o cálculo de um conjunto de estatísticas e pesquisas (Usando XPath e afins).

13.3.2 Uso claro e expansível de metadados nas colecções

A existência de metadados nas colecções foi algo que a organização teve em conta. Existe, por exemplo, um elemento `<DOC>`, com metadados variante linguística e género textual.

```
<DOC>
  <DOCID>HAREM-871-07800</DOCID>
  <GENERO>Web</GENERO>
  <ORIGEM>PT</ORIGEM>
...
```

Por um lado, parece-me que os valores do atributo género cobrem mais que uma faceta: um documento *político* (conteúdo temático) poderá ser também uma *entrevista*, ou estar disponível (suporte) em *Web, CorreioElectrónico*. Ou seja, seria útil múltiplas ocorrências de géneros, ou separar esta informação em mais do que um campo.

Por outro lado, gostaria de ver um elemento `META` que agrupasse todos os metadados do documento de modo a permitir que possa haver mais fácil enriquecimento (por parte do HAREM ou de outro qualquer uso futuro).

13.3.3 Questões ligadas à estrutura classificativa usada

Cada entidade marcada está a ser classificada "semanticamente".

Originalmente o MUC propôs um sistema classificativo com 3 categorias e 7 tipos. O HAREM propôs subir a fasquia para uma categorização com 10 categorias e 41 tipos. A meu ver essa decisão foi necessária e acertada.[2]. Havendo uma taxonomia a dois níveis, há naturalmente a hipótese de participações parciais:

- nível 0 –> marcar apenas as entidades
- nível 1 –> apresentar apenas as classificações do primeiro nível

[2] Genericamente subir a fasquia é bom quando houver pelo menos um atleta que a transponha...

- nível 2 –> apresentar a classificação completa.
- ou ainda escolher uma subárvore da taxonomia em causa.

Por outro lado, foi construída uma função de conversão

$$harem2muc : Charem \longrightarrow Cmuc$$

que mapeia classificações HAREM em classificações MUC. – tornando possível a comparações de resultados (medidas de acerto) entre as duas competições[3]. Esta função de mapeamento entre os dois sistemas para a maioria dos casos é simples e natural, havendo no entanto zonas da estrutura HAREM que são difíceis de mapear em MUC (o que não surpreende nem impede a leitura dos valores após conversão).

Dum modo semelhante parece-me que há zonas da taxonomia HAREM que são pouco naturais e claras – vistas pelo prisma de representação de conhecimento. Constatou-se naturalmente dificuldades em arranjar consenso entre os participantes em relação ao referido sistema de classificação do HAREM, o que é natural e habitual nestas actividades, e que me parece não ter constituído obstáculo importante ao funcionamento.

Genericamente, a marcação combinada tem o seguinte aspecto:

```
<Nivel1  tipo="Nivel2">Entidade encontrada</Nivel1>
```

No que diz respeito à estrutura classificativa, os problemas com que deparamos são:

1. Apesar de existir uma etiqueta de alternativa (`<ALT></ALT>`) para descrever alternativas de que sequências de palavras compõem a entidade (vagueza na identificação textual), uma notação (`|`) para vagueza/indefinição das classes semântica e ainda uma classe especial `outra` para situações *duvidosas*, não vejo claramente como descrever ao nível da marcação:

   - **ignorância total** (ex: *o X é interessante* – não sei nada acerca de X). Um humano normalmente saberá classificar uma ocorrência mas é frequente um ferramenta não o saber; nessa situação pretendemos anotar essa ignorância.
   - dúvida (ex: *o Porto é imprevisível*: ou é uma cidade ou um clube de futebol mas não as duas ao mesmo tempo – só consegui concluir alguma informação parcial),
   - classificação múltipla (*na Biblioteca da Universidade de Coimbra encontramos o espírito barroco* – acho válidas duas ou mais classificações: Obra de arte, Local Biblioteca, ...)

[3] **Nota do editor:** A comparação entre os resultados do HAREM e os do MUC e a conversão das respectivas etiquetas não é um assunto trivial, contudo, , como é discutido nos capítulos 3 e 4.

ou seja:

```
<nivel1 tipo="não faço ideia">e1</nivel1>
<nivel1 tipo="das duas uma:A ou B mas tenho dúvidas qual">e1<nivel1>
<nivel1 tipo="tanto A como B são tipos de">e1<nivel1>
```

Estou convicto de que o nível de ambiguidades/ignorâncias aparece mais na resposta dos sistemas do que na resposta de humanos.

2. Há situações (ao fazer a "formatação"a dois níveis) em que certas sub-árvores são facetas (quase independentes) levando a que faça sentido duas classificações, e que por vezes a solução oficial "perca"certas facetas e aspectos cruciais à caracterização da entidade em causa.

   Considere-se o seguinte exemplo da colecção dourada:

   `<LOCAL|OBRA TIPO="ALARGADO|ARTE">Biblioteca Pública</LOCAL|OBRA>`

   A referida biblioteca é um lugar, um edifício ou semelhante mas simultaneamente é património artístico, (é uma obra de arte). De certo modo, ser ou não obra de arte é uma faceta que poderemos querer aplicar a edifícios, livros, cidades e outras classes. Portanto constitui uma informação que deveria poder coexistir com a informação da classe a que se refere. Ou seja aquela biblioteca é simultaneamente um edifício e uma obra de arte[4].

3. genericamente a existência de herança múltipla complica certas zonas da estrutura classificativas.

   Considere-se o seguinte exemplo teórico. Se a minha maneira de ver o mundo considerar que:

   ```
   palácio    é uma subclasse   obras de arte
   palácio    é uma subclasse   edifícios
   ```

   (ou seja palácio tem dois pais, ou tem herança múltipla dessas duas classes) uma marcação em taxonomia a dois níveis (e já agora usando uma notação semelhante à do HAREM) tenderá a ver uma ambibuidade artifical entre

   ```
   <ObraDeArte tipo="palácio">...
   <Edifício   tipo="palácio">...
   ```

   Em situações como esta o uso de **palácio** (sem obrigação de escolher qual dos pais) tenderia a simplicar as coisas[5].

[4] **Nota dos editores:** isso é precisamente o que a notação do HAREM quis dizer: que aquela ocorrência de Biblioteca pública é simultaneamente as duas coisas.

[5] **Nota dos editores:** essa é exactamente a filosofia do HAREM: não ver ambiguidades quando não existem. No caso em questão, seria **ambas** as coisas: `<OBRA | EDIFICIO>`. O HAREM nunca marca ambiguidade, porque assume que os humanos conseguem distinguir. O caracter ' '|' ' indica sempre vagueza.

4. por vezes o enquadramento das ferramentas concorrentes força estruturas classificativas diferentes das usadas e ligeiramente "antagónicas". Isto é apenas uma constatação que complica a participação e para a qual não há uma solução óbvia mas que ainda assim descrevemos:

   Considere-se o seguinte par de exemplos da colecção dourada:

   ```
   Visite o <OBRA TIPO="PRODUTO">DataGrama Zero</OBRA> a Revista
   Eletronica ( ...)
   A revista foi denominada <ABSTRACCAO TIPO="NOME">Medicina e
   Cultura</ABSTRACCAO> (...)
   ```

   Independentemente do contexto linguístico em que estas entidades possam estar a ser usadas, dum ponto de vista de representação de conhecimento pretende-se tirar partido de que esta duas revistas têm muito em comum (classes idênticas ou aparentadas) e será completamente inaceitável ignorar/esquecer que *Medicina e Cultura* é uma *revista*.

A granularidade e capacidade distintiva

Considere-se a questão ligada com os conceitos *Portugal*, *país*, *entidade geográfica*, etc:

O seguinte conjunto de relações binárias pode ser usado para descrever (algumas das) propriedades do conceito *Portugal*:

```
Portugal IOF  país
país     ISA  entidade geográfica
país     ISA  instituição administrativa
país     ISA  povo
...
```

Numa situação como a do IKF/RENA não dispomos de informação suficiente para resolver devidamente essa questão de escolher entre os vários pais possíveis e, assim, optámos por baixar a fasquia, crentes de que ter uma classificação que falhe 40% dos casos é pior do que dizer que é simplesmente um país.

Na visão IKF/RENA a nossa intenção corresponde a ir decorando a árvore de conhecimento com todos os atributos que conseguirmos obter (trata-se de uma finalidade específica nossa), ou seja pretendemos juntar a *Portugal* os atributos ligados a país nas suas várias acepções e usos (presidente da república, língua, rios, área, etc).

Esse tipo de junção e processamento de atributos, heranças, etc, cria restrições ao tipo de árvores classificativas a usar: a relação subclasse (nível1 – nível2 da estrutura HAREM) passa a ter maiores responsabilidades...

13.3.4 Sugestão para futuras edições

Em resumo, para futuras edições propunha:

- Etiquetagem mais prática:
  - uma única etiqueta Entidade `<ent ...>...</ent>`
  - um atributo *tipo* `<ent t="país">...</ent>`
  - com notação clara para alternativas `<ent t="t1|t2"> ...`
  - com notação clara para multiclassificação `<ent t="t1;t2"> ...`
  - para informação parcial = escolher um nó mais acima na árvore classificativa (caso extremo = topo = entidade)
  - um atributo de unificação para permitir ligar referências à mesma entidade
- Ontologia multi-nível de classes, com herança múltipla
- Identificadores de classe mais claros e únicos – a questão da clareza é crucial[6] para o contexto de extracção de informação onde o RENA se encaixa: dizer que *Palácio de Vila Flor* é um `LOCAL:ALARGADO` é inaceitável do ponto de vista de extracção de informação enciclopédica[7].

13.4 Conclusões e trabalho futuro

A participação no HAREM foi muito positiva, embora, por questões conjunturais, não tenha sido possível tirar partido de uma série de iniciativas.

A participação do RENA no HAREM seguiu uma abordagem que não visava maximizar o resultado final da avaliação, mas antes o tentar ajudar à evolução do RENA de acordo com os nossos objectivos imediatos (que por vezes não coincidiram com os do HAREM).

Apesar das evoluções conseguidas, o estado actual do RENA é de protótipo.

Ao nível do trabalho futuro, há genericamente o objectivo:

- melhorar as regras de inferência e unificação e resumo
- criar um processador estrutural
- melhorar o sistema de tratamento de nomes incluindo também dados estrangeiros
- documentar melhor a interface de biblioteca Perl.

[6] No geral, em teoria da classificação há a recomendação de que cada classificador deverá, sempre que possível, ter autonomamente uma leitura clara.

[7] Como dissemos, do nosso ponto de vista, palácio, monumento, etc, seria preferível. Classificações como `LOCAL`, localidade, edificação, são também claras; `LOCAL:ALARGADO` por si só é de leitura pouco clara e parece-me significar algo como *local que não se encaixa nas outras subcategorias*.