# Relatório Técnico sobre o REPENTINO

REPositório para o reconhecimento de ENTIdades NOmeadas
Luís Sarmento, Porto, Maio de 2005


Resumo ..... 1
Introdução ..... 1
Entidades Nomeadas e Entidades Mencionadas ..... 2
O desenvolvimento do sistema de classificação ..... 3
A gestão inicial do recurso ..... 4
A recolha dos exemplos ..... 5
  Pesquisa de exemplos usando Padrões e Corpora ..... 6
  Pesquisa de exemplos na Web ..... 8
O sistema de classificação do REPENTINO ..... 9
  Abstracções ..... 9
  Arte / Media / Comunicação ..... 10
  Natureza ..... 11
  Eventos ..... 11
  Papeladas ..... 12
  Locais ..... 12
  Organizações ..... 14
  Produtos ..... 14
  Seres ..... 16
  Substâncias ..... 16
  Outros ..... 16
O REPENTINO como ferramenta colaborativa ..... 17
Números do REPENTINO ..... 19
Planos futuros para o REPENTINO ..... 20
Conclusões ..... 21


## Resumo


Neste relatório descreve-se o REPENTINO, o REPositório para reconhecimento de ENTIdades NOmeadas - http://poloclup.linguateca.pt/repentino/ - um recurso léxico-semântico que armazena cerca de 450 mil exemplos de entidades nomeadas, e que se encontra preparado para receber contribuições adicionais do público em geral. Serão descritos os seus métodos de criação, assim como serão explicadas as razões que levaram ao seu desenvolvimento. Iremos também evidenciar a ligação do REPENTINO com o SIEMÊS, o sistema de reconhecimento de entidades nomeadas com o qual o Pólo do Porto participou no HAREM e que foi a principal motivação para a concepção do REPENTINO. Será feita uma descrição detalhada do seu sistema de classificação e serão apresentadas algumas estatísticas acerca do seu conteúdo actual. Finalmente, serão propostas algumas possibilidades de melhoria do recurso e de eventuais aplicações.


## Introdução

O Pólo do Porto da Linguateca participou como concorrente na primeira edição do HAREM, a avaliação conjunta de sistemas de reconhecimento de entidades mencionadas (REM). Apesar de não possuir nenhuma experiência na área do REM, o Pólo do Porto possuía, contudo, algum tecnologia simples para a detecção de terminologia em texto técnico (instalada no Corpógrafo) que poderia servir de base à

tarefa de REM e que pensava inicialmente em adaptar. No entanto, ao fim de algumas tentativas de efectuar pequenas adaptações ao sistema de extracção de terminologia, foi notória a dificuldade em fazer as adaptações necessárias para que o sistema fosse cumprir a tarefa de REM. De facto, se por um lado a fase de identificação / delimitação das entidades mencionadas poderia ser facilmente alcançada com um grau de precisão e abrangência elevada, a fase de classificação semântica apresentava dificuldades muito grandes, nomeadamente na necessidade de recorrer à análise de evidências internas e externas acerca da classe e papel semântico das entidades.
Por esse motivo, o Pólo decidiu abandonar a adaptação do seu sistema de extracção terminológica para as tarefas de REM e decidiu-se pela implementação de um sistema de raiz, sabendo à partida que teria algumas limitações de recursos para tal efeito, dada a existência de outros projectos em que se encontrava envolvido. Para poder fornecer o máximo de reutilização ao sistema a ser desenvolvido, optámos por desenvolver um sistema de REM de largo espectro de forma a poder ser integrado noutras aplicações como, por exemplo, o Corpógrafo.
Tendo também já participado na discussão das regras do HAREM, a equipa do Pólo chegou às seguintes conclusões:

1. o problema do REM pode realmente ser muito abrangente quando se ultrapassam as categorias básicas que tradicionalmente são definidas nas avaliações (Pessoa, Local, Organização, Produto/Obra). Há possibilidade de incluir várias outras categorias quer por especialização das anteriores, quer considerando outras que normalmente não são tidas em conta, como Abstracções, Documentação, etc.
2. a construção de um sistema de grande abrangência poderá envolver o desenvolvimento de uma enorme base de regras sobre evidências internas/ externas e/ou a construção de almanaques. A construção de regras e de sistemas que as implementam é normalmente complexa, sendo a sua manutenção habitualmente difícil. Os almanaques para português não abundam, não tendo a nossa equipa encontrado qualquer recurso público, para além de algumas listas dispersas contendo nomes de empresas ou de pessoas.

Nestas circunstâncias, a primeira decisão tomada foi a de tentar recolher o máximo de exemplos de entidades nomeadas com o objectivo de melhorar a compreensão do problema. Com suficiente amostra, seria eventualmente possível a construção de regras de identificação e classificação de entidades, ainda que focadas apenas na análise de evidências internas. Por outro lado, e dada a carência de recursos desta natureza em português, pensámos que esta recolha seria uma boa oportunidade de contribuir para o REM independentemente da nossa capacidade de construção de um sistema de REM. Da convergência de todos estes objectivos surgiu o REPENTINO, com o objectivo de ser a melhor contribuição que o Pólo do Porto poderia realizar em tempo útil para a área.

## Entidades Nomeadas e Entidades Mencionadas

Um ponto basilar na nossa aproximação consiste na premissa da diferença entre entidade nomeada e entidade mencionada. Entendemos que uma entidade nomeada se refere a uma entidade que possui um nome próprio que é usado na sua individualização, ainda que esse nome próprio seja utilizado por outras entidades, podendo gerar situações ambíguas. Uma entidade nomeada possui propriedades semânticas intrínsecas independentes de contexto, que são constantes relativamente ao seu papel semântico no

discurso. Por exemplo, "Porto" nomeia uma cidade, isto é, uma entidade que é um local geográfico/administrativo. O nome "Porto" nomeia, portanto, um local particular (ainda que possa nomear outras entidades), cuja propriedade semântica intrínseca é essa e é constante ou pelo menos estável num dado âmbito temporal. Poderemos eventualmente não concordar com a classificação utilizada ou até com a etiqueta "local geográfico/administrativo", mas qualquer que seja o sistema de classificação ou a etiqueta usada para a entidade nomeada, a propriedade semântica intrínseca da referida entidade não se altera. Por outro lado, quando se fala de "entidades mencionadas" o problema é bem diferente. E aqui encontramos vários casos bem mais complexos. De facto podemos encontrar situações em que uma determinada entidade é mencionada num contexto que lhe atribui um determinado papel semântico diferente do que lhe é intrínseco, bem como podemos encontrar situações em que um nome de uma determinada entidade bem definida é usado para mencionar implicitamente outras entidades (ex: "Porto candidata-se ao Jogos Olímpicos", "A presença do Porto em Macau", "Tenho o Porto no coração", "Porto é convidado de honra no Salão do Livro").
O REPENTINO pretendeu logo desde o início ser apenas um armazém de exemplos de entidades nomeadas, pelo que não inclui nenhuma informação relativa ao contexto. Este objectivo pode parecer uma aproximação demasiado ingénua, mas a alternativa implicava um conhecimento do problema que não possuíamos à partida. Além disso, obrigava a um sistema de classificação que fosse capaz de organizar as entidades não só relativamente às suas propriedades intrínsecas, mas também relativamente às várias possibilidades de menção. Consideramos que um recurso como o REPENTINO, por muito simples que fosse, poderia ser útil em algumas situações. Na pior das hipóteses o REPENTINO poderia servir para construir um sistema de REM minimalista que apenas realizaria operações de consulta e marcaria as entidades encontradas em texto livre com a classificação usada para as armazenar.

## O desenvolvimento do sistema de classificação

Uma questão que também se colocou logo à partida foi: que sistema de classificação usar para organizar os exemplos recolhidos? Por outras palavras: mesmo assumindo que iríamos organizar os exemplos recolhidos discriminando-os segundo propriedades intrínsecas, que categorias e subcategorias deveríamos considerar para classificar os exemplos? Esta questão encontrava-se, e ainda se encontra, em aberto. Esta questão é quase equivalente a "como classificar os objectos do mundo". Por esse motivo decidimos adoptar não um sistema de classificação em si, mas sim uma estratégia para desenvolver um sistema de classificação apropriado ao contexto em causa: o REM. Assim, partindo de um conjunto base de categorias de topo, iguais ou muito próximas daquelas que foram propostas pelo HAREM iríamos especializando o nosso sistema de classificação por criação de novas subcategorias sempre que, e apenas quando, se encontrassem exemplos de entidades em número suficiente que permitissem preencher significativamente essa categoria. Ou seja, para além de um conjunto de categorias semânticas de topo que poderiam facilmente ser consideradas consensuais, apenas iríamos considerar subcategorias cujos exemplos que nelas se enquadrassem tivessem alguma representatividade. A representatividade por sua vez poderia ser estimada através de pesquisas em corpora e de pesquisas na rede, ainda que estes processos sejam sempre falíveis e dados a fenómenos de sub-amostragem. Contudo, pensamos que criar categorias que depois cobrem apenas uma fracção reduzida de exemplos apenas porque nos parece semanticamente correcto incorre também em problemas de reduzida adaptação do repositório à realidade de um sistema de REM.

Esta estratégia de construção teve várias consequências no desenvolvimento do sistema de classificação do REPENTINO, o que gerou algumas diferenças relativamente ao sistema de classificação adoptado pela organização do HAREM (que resultou da combinação de vários consensos). Em primeiro lugar, foi-nos possível especializar mais detalhadamente algumas categorias. Por exemplo, no que se refere a locais, foi-nos possível distinguir entre 16 subcategorias, quase todas elas com mais de 100 exemplos e muitas com alguns milhares, o que é um número de subcategorias muito superior ao proposto pela organização do HAREM. Adicionalmente, encontrámos exemplos que permitiram criar categorias de topo completas, e que depois acabaram por ser divididos detalhadamente. Um destes casos é a categoria "Papeladas" que foi possível preencher com cerca de 4500 exemplos divididos por oito subcategorias. Em muitos destes casos, as categorias e subcategorias encontradas dificilmente seriam consideradas na construção em abstracto de um sistema de classificação, porque parecem pouco importantes. No entanto, foi possível encontrar em grande quantidade exemplos referentes a tais categorias, tanto em corpora como executando pesquisas em motores de pesquisa, o que na nossa opinião demonstra a importância das aproximações "empiristas" durante a fase da conceptualização.
Actualmente, o sistema de classificação do REPENTINO prevê 11 categorias de topo que agregam no total 102 subcategorias, o que demonstra o grau de detalhe a que foi possível chegar. As implicações que este nível de detalhe tem sobre sistemas de REM que usem este recurso serão descritas noutro documento. Uma descrição mais detalhada do sistema de classificação do REPENTINO encontra-se numa próxima secção.

# A gestão inicial do recurso

Para assegurar uma recolha e organização sustentada, desenvolvemos localmente uma base de dados para armazenar os exemplos encontrados. Foi também desenvolvida uma interface Web que simplificava a gestão de todas as entidades recolhidas, assim como a organização do sistema de classificação. Através desta interface inicial, que depois de vários desenvolvimentos se veio a transformar na actual interface de administração do REPENTINO, tornava-se possível:

1. acrescentar, remover, alterar os exemplos recolhidos;
2. executar com um simples clique pesquisas sobre motores de pesquisa Web (Google e Tumba) para poder encontrar ocorrências de uma dada entidade para proceder à sua validação;
3. criar, fundir, remover e alterar categorias e subcategorias.

Durante todo o desenvolvimento do REPENTINO esta interface Web mostrou-se fundamental, pois permitiu que grande parte do processo de construção do recurso pudesse ser executado por colaboradores da área da Linguística, já que não exigia a necessidade de conhecimentos de programação para o seu desenvolvimento. Desta forma o esforço de desenvolvimento do REPENTINO encontrava-se dividido entre a equipa do Pólo sem criar "engarrafamentos" do ponto de vista da engenharia. Actualmente, é a bolseira da FLUP Ana Sofia Pinto que assume a responsabilidade de curadora do recurso, podendo desenvolver toda esta tarefa usando a interface de administração desenvolvida.
Adicionalmente, foi possível manter o desenvolvimento do projecto REPENTINO sem prejudicar o objectivo inicial que era o do desenvolvimento de um sistema de REM. Como veremos já em seguida, os dois projectos entraram em simbiose perfeita.


Figura 1 - Uma vista sobre o interface de administração do REPENTINO

# A recolha dos exemplos

A recolha de exemplos não foi feita de uma forma arbitrária, mas sim orientada pelas necessidades de desenvolvimento do SIEMÊS. De facto, o funcionamento do SIEMÊS passa precisamente pela tentativa de detectar semelhanças entre entidades que encontra em texto livre e outras que façam parte da sua base de conhecimento. Sem entrar em grande detalhes acerca do SIEMÊS neste documento, poderemos dizer que o SIEMÊS aumenta a sua capacidade de detectar e classificar correctamente uma dada entidade mencionada quando possui na sua base de conhecimento um exemplo de uma entidade nomeada próxima lexicalmente falando (possivelmente até igual). Com essa informação e com informação sobre contexto analisada posteriormente (daí a estratégia siamesa), o SIEMÊS formula um juízo de classificação.

Com o desenvolvimento de versões minimamente funcionais do SIEMÊS foi possível realizar os primeiros testes, inicialmente usando o material fornecido pela organização do HAREM e, mais tarde, usando texto retirado aleatoriamente do WPT03 e do CETEMPúblico. Por uma questão de simplicidade, o REPENTINO foi usado pelo SIEMÊS desde o início como base de conhecimento, apesar de isso não ser obrigatório e ser até prejudicial do ponto de vista da performance computacional. Houve desde o início do desenvolvimento planos para compilar uma base de conhecimento optimizada para o SIEMÊS a partir da informação do REPENTINO, mas isso acabou por não ser realizado.

No entanto, e a partir dos testes realizados ao sistema SIEMÊS+REPENTINO foi possível guiar todo o processo de recolha de exemplos. Sempre que o SIEMÊS não era capaz de classificar correctamente uma entidade mencionada por falta de evidências que deveriam supostamente estar na sua base de conhecimento, eram efectuadas recolhas de exemplos de entidades que depois eram acrescentadas ao REPENTINO para colmatar a falha do SIEMÊS. Normalmente, a detecção de uma falha de "conhecimento" do SIEMÊS levava à adição de várias dezenas de exemplos ao REPENTINO. Por exemplo,

supondo que o SIEMÊS não foi capaz de classificar "Faculdade de Letras da Universidade do Porto", por falta de dados na sua base de conhecimento (para o efeito, o REPENTINO) então seriam pesquisados e adicionados ao REPENTINO vários exemplos de Faculdades (ex: "Faculdade de Medicina da Universidade de Coimbra") e eventualmente de outras entidades que fossem próximas semanticamente (ex: "Escola Superior de Educação do Instituto Politécnico do Porto". Esta política de aquisição de exemplos permitiu guiar o processo de recolha mais focadamente.
A recolha propriamente dita dos exemplos de entidades nomeadas foi realizada empregando dois métodos distintos. Por um lado, e aproveitando o facto de termos disponível localmente corpora de grandes dimensões, como o CETEMPúblico, e colecções de texto massivas, como o WPT03, foram efectuadas pesquisas sobre texto livre usando padrões lexicais. Por outro lado, foram procurados em sítios Web temáticos listas de exemplos de entidades nomeadas, sendo os sítios Web encontrados quer usando os motores de pesquisa habituais, quer a partir do conhecimento dos membros da equipa do Pólo. Estas duas aproximações eram de certa forma complementares relativamente à capacidade que tinham de encontrar novos exemplos. Note-se que em ambos os casos foi sempre necessário algum esforço manual no tratamento, validação e organização dos candidatos a exemplo recolhidos.

## *Pesquisa de exemplos usando Padrões e Corpora*

Relativamente ao primeiro método, a extracção de exemplos a partir de texto livre essa foi realizada maioritariamente sobre o WPT03. Para que o processo de extracção pudesse ser realizado com alguma eficiência e acima de tudo sem demasiado trabalho manual de tratamento e validação, decidimos executar pesquisas focadas numa determinada subcategoria. Assim, tentámos explorar três propriedades que se verificam para certas categorias de entidades e que facilitam a sua detecção em texto livre:

1. a presença de uma "cabeça" descriminatória. Grande parte das entidades, normalmente organizações, possuem uma estrutura relativamente bem definida, começando por um conjunto de palavras que permite a identificação da subcategoria de uma forma simples. Por exemplo: "*Universidade* do Porto", ou "*Junta* de Freguesia de Ramalde".
2. uma forte probabilidade de ocorrência num determinado contexto "típico". Há certos contextos relativamente bem conhecidos e fáceis de formular que permitem a extracção de certas entidades. Por exemplo, se estivermos a pesquisar locais a instanciação de um padrão como "localizado na XXX" ou "próximo da XXX" indicia fortemente que "XXX" se tratará de um local.
3. a presença de um sufixo descriminatório. Certas entidades, normalmente organizações, possuem como sufixo partículas típicas como "Lda." ou "S.A".

Poderia afirmar-se que o conhecimento destes padrões esvazia a necessidade da construção do próprio REPENTINO com o objectivo de auxiliar sistemas de REM: se sabendo os padrões se consegue recolher e classificar entidades, então porquê armazenar os exemplos, ainda por cima excluindo os contextos. Na verdade isto não é bem assim, porque o problema é um pouco mais complexo.
Em primeiro lugar, o facto de existirem certos contextos típicos ou regularidades morfológicas (cabeças ou sufixo) que indiciam a classe de uma determinada entidade, isto não permite só por si resolver o problema do REM. De facto, os contextos típicos ocorrem com uma frequência muito reduzida, ou sofrem grandes variações em torno de uma versão base, pelo que só em certas ocasiões em que a forma do discurso é

controlada, ou em situações onde existe muita redundância, é que se torna possível fazer uso efectivo de tais padrões. Além disso, as entidades nomeadas admitem normalmente muitas variações (supressões de prefixos, transformação parcial em acrónimos, e.g.: "C.M. Lisboa") e que dificultam bastante a sua detecção e classificação. Ou seja, a utilização exclusiva de padrões e de regras baseadas em certas regularidades morfológicas, mesmo que não ambíguas, nem sempre é exequível do ponto de vista prático pela incapacidade de lidar com todas as variações válidas e possíveis.
Em segundo lugar, a utilização de padrões permite, e apenas em certas circunstâncias como referido anteriormente, eventualmente diferenciar semanticamente as entidades até um certo nível de detalhe. Ou seja, um determinado padrão poderá indiciar que uma entidade se trata de uma Organização, mas só em casos muito particulares é que será possível discriminar entre uma empresa ou uma associação recreativa. Ainda a este nível, as regularidades morfológicas permitem algo mais a este nível, mas há tantos casos que a codificação das regras necessita de uma base de exemplos vasta, pois a sua codificação manual a partir da intuição pessoal de quem as constrói é inviável. A combinação das duas informações parece ainda mais complexa de codificar.
E esta condição leva a um argumento final relativamente à construção do REPENTINO. Um recurso depois de construído pode ter múltiplos usos e ser usado para desenvolver sistemas sem ter de partir do vazio. Por exemplo, parece-nos bastante viável a construção de uma boa base de padrões lexicais altamente descriminatórios a partir da compilação dos contextos encontrados quando se pesquisa em corpora cada uma das entidades do REPENTINO já previamente classificadas. Ou a construção automática de regras baseadas em regularidades morfológicas que permitam a análise expedita das possíveis categorias de uma dada entidade.
O processo de recolha pode então ser decidido em 6 etapas:

1. escolher uma categoria para a qual se pretende pesquisar exemplos de entidades
2. decidir qual a estratégia mais apropriada para a pesquisa de exemplos: pesquisa por "cabeça", por contexto ou por sufixo);
3. construir o respectivo padrão e executar a pesquisa usando scripts em Perl desenvolvidos para o efeito. As pesquisas eram efectuadas sobre o WPT03.
4. validação manual dos candidatos obtidos tendo em conta a categoria pretendida. Contudo, alguns resultados inválidos obtidos poderiam ser apropriados para outras categorias, pelo que muitas vezes sugeriam certos padrões de pesquisa interessantes para as referidas categorias.
5. introdução dos candidates positivos na base de dados
6. se necessário, era criada uma nova categoria/subcategoria que aumentava assim o sistema de classificação

A próxima figura resume o processo descrito:

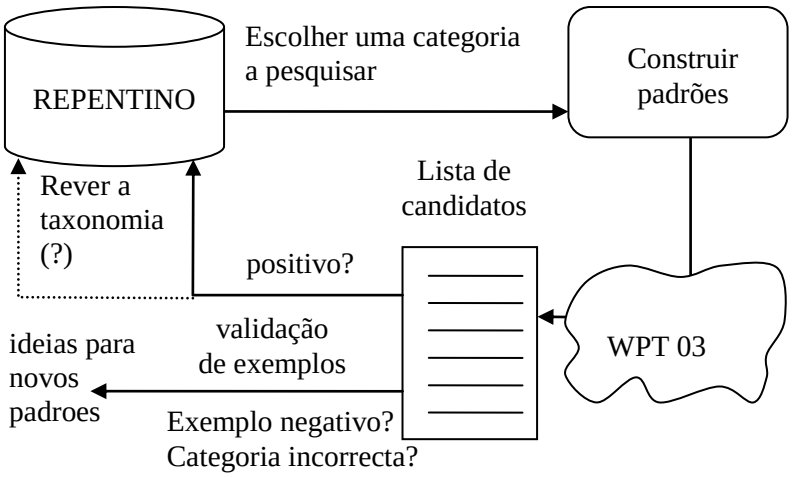


Figura 2 - O processo de recolha de exemplos usando pesquisa de padrões sobre corpora

Cada ronda completa sobre a totalidade do WPT03 demorava inicialmente entre 2h30 e 3h00, já incluindo a revisão manual dos exemplos recolhidos. Normalmente, por cada ronda é possível encontrar cerca de 500 a 1000 exemplos. O processo de pesquisa sobre o WPT03 foi entretanto optimizado, sendo agora possível pesquisar a totalidade do WPT03 em menos de 15 minutos. Contudo o tempo de revisão manual dos exemplos recolhidos eleva o tempo total de cada ronda para cerca de 1h.

Durante o processo de extracção usando padrões foram muitas vezes encontrados contextos em torno de certos exemplos positivos que considerámos importante armazenar. Estes contextos têm quase sempre uma função de modificação dos exemplos recolhidos que podem ser utilizados em futuras análises ou com o objectivo de criar regras de classificação. Por exemplo, se encontrássemos "margem do Rio Douro" adicionaríamos ao REPENTINO tanto a entidade base "Rio Douro" como uma versão do exemplo com o modificador "margem do Rio Douro". Não conseguimos ainda avaliar o impacto desta decisão na qualidade geral do recurso, mas podemos dizer que estes casos são relativamente raros relativamente à totalidade de exemplos armazenados. Em todo o caso, o sistema de administração do REPENTINO permite uma rápida alteração ou eliminação destes casos, se isso se tornar necessário.

## *Pesquisa de exemplos na Web*

Relativamente ao segundo método, pouco há a dizer tirando o facto se ter recolhido informação de mais de 120 sítios Web. Este método é mais apropriado para obter exemplos de entidades cuja pesquisa em corpora é muito mais difícil, quer pela variedade de contextos em que podem ocorrer, quer pelas variadíssimas formas que podem assumir. Por exemplo, nomes de músicas, software, filmes e outros produtos culturais possuem estas características: por um lado é difícil especificar a sua forma e os eventuais contextos, por outro são facilmente localizáveis em sítios Web temáticos que apresentam longas listas de exemplos.

Em termos de variedade de entidades recolhidas um dos principais destaques vai para o sítio Web da Wikipedia onde foi possível encontrar exemplos para diversas entidades. O sítio de onde foi possível recolher mais exemplos foi o do Tribunal Superior Eleitoral do Brasil, (http://www.tse.gov.br) de onde pudemos recolher mais de 257 mil nomes de candidatos das listas eleitorais dos vários estados do Brasil. Para mais detalhes acerca das fontes usadas remete-se o leitor para a página de estatísticas do sítio oficial do REPENTINO.

# O sistema de classificação do REPENTINO

Como referido anteriormente, o sistema de classificação do REPENTINO foi crescendo e foi sendo adaptado à medida que eram encontrados exemplos de novas entidades, o que por sua vez era motivado pela detecção de falhas de cobertura no SIEMÊS. Por esse motivo o sistema de classificação do REPENTINO apresenta importantes diferenças relativamente à hierarquia de classificação semântica proposta pela organização do HAREM, apesar de ter sido fortemente inspirada nessa proposta. O sistema de classificação actual do REPENTINO inclui 11 categorias de topo e 102 subcategorias, um valor bastante diferente das XX subcategorias previstas no HAREM distribuídas também por 11 categorias de topo, embora distintas. As categorias de topo do sistema de classificação do REPENTINO são:

1. Abstracções
2. Arte / Media / Comunicação
3. Natureza
4. Eventos
5. Papeladas
6. Locais
7. Organizações
8. Produtos
9. Seres
10. Substâncias
11. Outros

O mapeamento das categorias do REPENTINO para o HAREM é normalmente de N para 1, embora haja alguns casos de subcategorias previstas no REPENTINO que não foram previstas no HAREM, e outras cujo mapeamento não é consensual havendo sobreposição com mais do que uma subcategoria alvo. Estas diferenças foram depois notórias durante a transferência dos resultados do SIEMÊS do seu sistema de classificação directamente derivado do REPENTINO para o sistema de classificação da avaliação conjunta.
Em seguida iremos descrever em detalhe cada uma das categorias e as respectivas subcategorias fornecendo exemplos ilustrativos em cada um dos casos. Convém desde já referir que apesar dos esforços desenvolvidos o sistema de classificação não é totalmente ortogonal, havendo em alguns casos situações de sobreposição entre duas ou mais subcategorias. Em todo o caso, espera-se com o evoluir do sistema resolver alguns destes problemas de ambiguidades, sabendo à partida que qualquer solução será sempre uma solução de compromisso.

## *Abstracções*

Esta categoria inclui entidades abstractas tais como disciplinas, ciências, teorias e outras formulações mentais. Períodos, movimentos, estados, doenças e índices também fazem parte desta categoria. As ‘Abstracções’ podem ser divididas nas seguintes subcategorias:

1. **Estado/Condição -** doenças e outros estados/condições permanentes ou temporários. Ex: 'doença de Alzheimer', 'SIDA', 'síndrome de Down'.
2. **Crime -** todo o tipo de crimes previstos na lei. Ex: 'peculato', 'tráfico de influências', 'homicídio simples', 'lenocínio', etc.

3. **Disciplina/Arte & Ofício -** onde se incluiriam, por exemplo, 'Inteligência Artificial', 'Belas-Artes', 'Ciências da Educação', 'Electrónica', 'Tapeçaria Oriental', 'Tai-chi', 'Futebol', etc.
4. **Período/Movimento/Tendência -** movimentos, tendências ou outros períodos que tenham ocorrido num dado período da História. Ex: 'Renascimento', 'Movimento Feminista', 'Iluminismo', 'Budismo', 'Art Déco', etc.
5. **Formulação Mental -** regras, teorias, formulações teóricas ou sistematizações. Ex: 'Lei de Murphy', 'Teoria da Relatividade', 'lei da gravitação de Newton', 'princípio da incerteza de Heisenberg', 'princípio da dualidade de De Broglie', etc.
6. **Era/Época -** períodos de tempo em que se divide a história do Mundo. Ex: 'Idade da Pedra', 'Era Mesozóica', 'Antiguidade Clássica', 'Jurássico', etc.
7. **Processo -** Ex: 'fotossíntese', 'pseudomorfose', 'sulfatação', 'diferenciação magmática', 'processo anaeróbio aláctico', 'osmose', etc.
8. **Símbolo -** todo o tipo de símbolos religiosos, políticos, esotéricos, heráldicos, etc. Ex: 'Estrela de David','Crucifixo', 'Pentagrama', 'Selo de Salomão', etc.
9. **Índice/Taxa -** índices, taxas e outros indicadores económicos. Ex: 'NASDAQ', 'PSI 20', 'Índice de Concentração Empresarial', 'taxa de inflação', 'taxa de abstenção', 'PIB', etc.
10. **Expressões Latinas –** expressões em Latim que ocorrem com alguma frequência e que podem aparecer maiusculadas. Ex: 'Ad hoc', 'Totus Tuum', 'Tabula rasa', 'Data venia', etc.
11. **Tipo/Classe -** conceito que representa um tipo ou classe de algo que não se encaixa em nenhuma das outras categorias ou subcategorias. Ex: 'Metro', 'Metropolitano', etc.

## *Arte / Media / Comunicação*

Nesta categoria incluem-se apenas entidades nomeadas que são produtos relacionados com arte, media e comunicação. Estes podem ser divididos nas seguintes subcategorias:

1. **Filme -** títulos de filmes. Ex: 'Indiana Jones em Busca do Templo Perdido', 'A Guerra das Estrelas', 'Dogville', etc.
2. **Livro -** todos os livros ou publicações não periódicas, que possam ser consideradas equivalentes a livros, incluindo exemplares únicos. Ex: 'Os Maias', 'Sonho de Uma Noite de Verão', 'As Minhas Receitas de Bacalhau', etc.
3. **Música -** títulos de músicas de todos os tipos, assim como peças musicais. Ex: 'Dunas', 'Jingle Bells', 'Bolero de Ravel', etc.
4. **Multimédia -** jogos de computador e outros produtos multimédia, tais como cd-roms ou sítios web. Ex: 'Final Fantasy IV', 'Flight Simulator', 'Tomb Raider II', etc.
5. **Periódico -** todo o tipo de publicações periódicas, normalmente impressas, tais como jornais e revistas. Ex: 'Expresso', 'Jornal de Notícias', 'Super Interessante', 'Exame', etc.
6. **TV/Radio/Teatro -** programas e canais de rádio e televisão, assim como peças de teatro. Ex: 'Jornal da Noite', 'Acontece', 'Oceano Pacífico', 'My Fair Lady', 'Cats', 'SIC Comédia', 'AXN', 'RFM', 'The Green Lounge', 'Club 977 The '80s Channel', etc.
7. **Arte & Design -** aqui incluem-se quadros, esculturas ou qualquer outro objecto de arte, único ou não. Ex: 'O Grito' de Munch, 'La Pietà', 'chaise longue de Le Corbusier', etc.

8. **Texto Académico/Científico -** incluem-se aqui títulos de teses, artigos, dissertações e outros textos académicos e/ou científicos. Ex: 'Influência da granulometria na cinética de dissolução de fármacos', 'Hidrólise biocatalítica da trioleina', ' Tese de Mestrado em Literatura Portuguesa', 'A Statistical Approach to Machine Translation', 'A Maximum Entropy Approach to Natural Language Processing', etc.

## *Natureza*

Nesta categoria incluem-se animais, vegetais, todos os elementos que constituem os organismos vivos, assim como os fenómenos naturais. Estes podem ser divididos nas seguintes subcategorias:

1. **Animal -** todos os animais e respectivas espécies/raças. Ex: 'São Bernardo', 'Tubarão Martelo', 'Lince da Malcata', 'Tyranossaurus Rex', 'gato do Maine', etc.
2. **Fisiologia -** todos os elementos que constituem os organismos vivos, como músculos, células, ossos, etc. Ex: 'Esternocleidomastoideu', 'aparelho de Golgi', 'células de Scwann', 'ADN', 'Perónio', 'Neurónio', etc.
3. **Micro-organismos -** micróbios, bactérias, parasitas, fungos, protozoários, etc. Ex. 'bactéria Staphylococcus aureus', 'vírus Ébola', 'fungo da ferrugem asiática', 'Giardia Lamblia', 'protozoário Plasmodium vivax', etc.
4. **Vegetal -** todos os vegetais e respectivas espécies. Ex: 'Alecrim', 'Pinus Pinaster', 'sabugueiro', 'couve de Bruxelas', 'pêra Rocha', 'camomila', 'salsa', etc.
5. **Fenómenos Naturais -** aqui incluem-se os ventos, marés, terramotos, maremotos, ciclones, tufões, furacões, tempestades, etc. Ex: 'El Niño', 'Terramoto de 1755', 'ciclone Nancy', 'Zéfiro', 'furacão Ivan', 'Tempestade de Inverno Vivian', 'tufão Andrew', etc.

## *Eventos*

Nesta categoria inclui-se todo e qualquer evento, cujo início ou duração estejam claramente definidos. Os eventos podem ser divididos nas seguintes subcategorias:

1. **Desportivo -** acontecimentos desportivos, tais como campeonatos, torneios, entre outros. Ex: 'Jogos Olímpicos', 'Campeonato Nacional de Futebol', 'Torneio de Hóquei em Patins da Mealhada', 'Open do Estoril', etc.
2. **Socio-Cultural -** reuniões ou ajuntamentos que envolvam qualquer actividade social ou cultural. Ex: 'Feira da Pêra Rocha', 'Feira da Senhora das Dores', 'Exposição Internacional Filatélica de Buenos Aires', 'Arraial D'Ajuda', 'Festival Internacional de Cinema', 'Festa de São Tomás de Aquino', 'Mostra de Artes Plásticas do Concelho de Loulé', etc.
3. **Efeméride -** acontecimento ou período que tem lugar num determinado ponto da História. Ex: 'Dia D', 'I Grande Guerra Mundial', 'Guerra Civil Americana', 'Batalha de Aljubarrota', etc.
4. **Científico -** acontecimentos científicos, tais como conferências, simpósios, etc. Ex: 'Congresso de Pneumologia e Tisiologia do Estado do Rio de Janeiro', 'Conferência Engenharia 2001', 'Colóquio Vitivinícola da Estremadura', 'Simpósio Internacional de Vulcanoespeleologia', etc.
5. **Cíclico -** acontecimentos cíclicos que normalmente celebram/comemoram ou estão relacionados com uma determinada efeméride. Ex. 'Natal', 'Páscoa', 'Véspera de Ano Novo', 'Dia da Independência', 'Solstício', etc.

6. **Político -** acontecimentos de carácter político, tais como cimeiras, congressos partidários ou comícios. Ex: 'Cimeira Mundial sobre Desenvolvimento Sustentável', 'Comício do Partido Comunista', 'Congresso do Partido Social Democrata', etc.
7. **Prémio/Galardão -** Ex: 'Prémio Nobel da Literatura', 'Prémio Valmor', 'Galardão PME Excelência', 'Grande Prémio da Literatura Policial de França', etc.

## *Papeladas*

Aqui incluem-se as leis, decretos, tratados, pactos, normas e planos. As entidades nomeadas que cabem nesta categoria podem ser divididas nas seguintes subcategorias:

1. **Lei -** leis e decretos, em vigor ou não, promulgados por uma dada organização (ex: um país ou uma empresa) e que regulam o comportamento dentro dessa organização. Ex: 'Prohibition Act', 'Artº 14º do Decreto Lei nº 43/46 de 20/01', 'Decreto Lei nº 445/91 de 20 de Novembro','Lei da Água', etc.
2. **Acordo -** tratados, pactos e outras formas de acordos entre duas ou mais partes. Ex: 'Tratado de Tordesilhas', 'Pacto de Varsóvia', 'Convénio de Genebra', 'Declaração de Bolonha', 'Liga de Augsburgo', etc.
3. **Norma -** normas acordadas oficialmente e que são estabelecidas com um objectivo funcional. Ex: 'Unicode Standard', 'ASCII', 'BETAMAX', 'ISO', etc.
4. **Certificações -** alvarás, certidões, certificados, licenciaturas, formações, pós-graduações, apólices, licenças, mestrados, doutoramentos, etc. Ex: 'Alvará de Construção', 'Apólice de Seguro Automóvel', 'Certidão de Nascimento', 'Certidão Predial', "Certificado de Aptidão Profissional', 'Curso de Árbitros', 'Licença de condução de ciclomotor', 'Licenciatura em Filosofia', 'Mestrado em Contabilidade', 'Pós-graduação em Direito Fiscal', etc.
5. **Impostos/Emolumentos -** impostos ou outro tipo de emolumentos, como propinas, juros, etc. 'Imposto da SISA', 'Contribuição Autárquica', 'Propina Máxima', 'Taxa Moderadora', 'juros de Mora', etc.
6. **Planos e Procedimentos -** todos os documentos que estabelecem um determinado plano de acção ou que determinam os procedimentos para a realização de um dado projecto/objectivo. Ex: 'Quinto Programa Quadro', 'Programa Sócrates/ERASMUS', 'Política de Privacidade', 'Plano Marshall', 'Orçamento do Estado para 2003', 'Plano Hidrológico Espanhol', etc.
7. **Documentos -** todo o tipo de documentos burocráticos, tais como termos, minutas, relatórios, editais, actas, autos, balanços, ofícios, etc. Ex: 'Auto de denúncia', 'Balancete da Despesa', 'Bilhete de Identidade', 'Carta Circular nº 03/2001', 'Declaração Amigável de Acidente Automóvel', 'Ofício-Circular n.º2/2005', etc.

## *Locais*

Esta categoria inclui todas as entidades nomeadas que são individualizadas sobretudo de acordo com a sua posição no universo. Os 'Locais' podem ser divididos nas seguintes subcategorias:

1. **Terrestre -** regiões geográficas da Terra não tendo em atenção critérios políticos/administrativos. Inclui regiões com dimensões variáveis. Ex: Europa, Península Ibérica, vale do Rio Douro, etc.

2. **Hidro -** locais geográficos constituídos fundamentalmente por água (ex: rios, lagos, oceanos, etc.). Ex: 'rio Guadiana', 'Lago Maggiore', 'Oceano Índico', etc.
3. **Espacial -** pontos ou áreas localizadas no espaço, como, por exemplo, planetas, galáxias ou constelações. Ex: 'Mercúrio', 'Via Láctea', 'Ursa Maior', 'Constelação de Capricórnio', etc.
4. **Endereço -** endereço explícito de um determinado local. Deverá conter informação suficiente para, pelo menos, identificar um edifício. Ex: 'rua Cândido dos Reis nº7 r/c 1495 - 030 - Algés', 'praça Almeida Garrett nº27', etc.
5. **Socio-cultural -** edifícios ou áreas onde se realizam eventos socio-culturais, como teatros, estádios, galerias, museus, etc. Ex: 'Teatro Nacional D. Maria', 'Museu Soares dos Reis', 'Estádio do Dragão'.
6. **Religioso -** edifícios ou áreas onde se realizam eventos religiosos, como igrejas, catedrais, mosteiros, mesquitas, sinagogas, etc. Ex: 'Sé de Braga', 'Igreja da Lapa', etc.
7. **Endereço Alargado -** parecido com 'Endereço', mas não requer uma informação tão detalhada. Exemplos de endereços alargados são ruas, zonas como a 'Baixa' ou a 'Ribeira', praças ou outros locais dentro de uma área urbana. Ex: 'Rua de Cedofeita', 'Praça do Marquês', etc.
8. **País/Estado -** países e uniões ou federações de países de acordo com fronteiras geopolíticas, assim como estados com alguma independência administrativa. Ex: Portugal, Emirados Árabes Unidos, Estados Unidos da América, Massachussets, Alabama, Maranhão, etc.
9. **Povoação/Região/Div. Administrativa -** cidades, vilas, aldeias, freguesias, regiões, povoações, etc. Ex: Porto, Trás-os-Montes e Alto Douro, freguesia de Mafamude, A-dos-Cunhados, Vila da Marmeleira, Albufeira, etc.
10. **Civil/Administração/Militar -** edifícios, complexos ou áreas relacionadas com organizações civis, militares ou administrativas, tais como quartéis, centros de saúde, bases militares, etc. Ex: 'Quartel dos Bombeiros Voluntários de Valadares', 'Base militar de Tancos', 'Centro de Saúde de Vilar do Paraíso', 'Aeródromo de Tires', 'Hospital de São João', etc.
11. **Património/Monumento -** edifícios e outras construções consideradas património ou que são importantes por alguma razão simbólica. Ex: 'Torre Eiffel', 'Pirâmides Teotihuacán', 'Monumento aos Descobrimentos', 'Capela Sistina', etc.
12. **Propriedade -** quintas, herdades ou outras propriedades que têm um nome próprio e cuja posse não é facilmente divisível. Ex: 'Quinta das Rosas', 'Herdade da Baracha', 'Solar dos Albuquerques', 'Casa do Sal', etc.
13. **Mitológico/Ficcional -** locais que não existem de facto, mas que são referidos na literatura, cinema , etc. Ex: 'Atlântida', 'Isengard', etc.
14. **Comercial/Industrial/Financeiro -** locais como parques industriais, minas, centros comerciais ou outros edifícios ou áreas que se destaquem essencialmente pela existência associada de actividades comerciais, industriais e/ou financeiras. Ex: 'Centro Comercial Vasco da Gama', 'Parque Industrial de São Domingos', 'zona industrial da Maia', 'Taguspark', 'loja FNAC em Matosinhos', 'Hotel Mercure Batalha', etc.
15. **Infraestrutura -** todo o tipo de construções cuja finalidade seja essencialmente práctica, embora não possa ser definida claramente sob um ponto de vista comercial/industrial e financeiro. Incluem-se nesta categoria todas as infraestruturas de uso generalizado (estradas, pontes, barragens, etc.), bem como outras construções, cuja utilização seja múltipla ou indefinida, como é o caso de

edifícios ou empreendimentos com nome próprio. Ex: 'Ponte da Arrábida', 'Edifício Les Palaces', 'A1', 'nó de Francos', 'barragem do Alqueva', etc.

## *Organizações*

Esta categoria inclui todas as organizações, isto é, todas as entidades constituídas por mais de uma pessoa que existem e funcionam como um todo. As organizações, normalmente, têm objectivos, um conjunto de regras que as rege e uma estrutura interna, o que não se verifica em simples grupos de pessoas ou ajuntamentos. As 'Organizações' podem ser divididas nas seguintes subcategorias:

1. **Civil-Militar -** organizações com propósitos militares e/ou civis/sociais, como é o caso dos bombeiros, hospitais, exército, etc. Ex: 'Bombeiros Voluntários da Sertã', 'Guarda Nacional Republicana', '2º Esquadrão de Aviação do Exército', 'Cruz Vermelha', etc.
2. **Clubes -** esta sub-categoria destina-se exclusivamente a clubes desportivos. Ex: 'Futebol Clube do Porto', 'Sport Clube Dragões Sandinenses', 'Mocidade Invicta Futebol Clube', 'Sport Lisboa e Benfica', 'Sporting Clube de Portugal', etc.
3. **Desportiva -** organizações directamente ligadas ao desporto, como federações, associações desportivas, etc. Ex: 'Futebol Clube do Porto SAD', 'Federação Portuguesa de Andebol', 'Associação de Patinagem do Porto', 'Associação Butokukai de Karate-do', 'Liga de Clubes', etc.
4. **Empresa -** organizações claramente com fins lucrativos. Ex: 'TAP Air Portugal', 'SEIKO EPSON Corporation', 'Plátano Editora', 'Alfa Romeo', 'Agência Associated Press', etc.
5. **Ensino/I&D -** organizações, públicas ou privadas, ligadas à educação ou à investigação e desenvolvimento. Ex: 'Universidade do Porto', 'Faculdade de Economia', 'Escola Secundária de Barcelinhos', 'Observatório Afonso de Chaves', 'Laboratório Nacional de Luz Síncrotron', etc.
6. **Governamental/Administrativa -** organizações oficiais criadas por governos, como é o caso dos Ministérios, Departamentos, Secretarias, Câmaras, etc. Ex: 'Câmara Municipal da Azambuja', 'Consulado da Holanda', 'Embaixada da República de São Tomé e Príncipe', 'Junta de Freguesia de Bemposta', 'Ministério do Comércio Externo', 'Secretaria de Estado das Pescas', etc.
7. **Grupos de Interesse -** organizações criadas para proteger os interesses de um qualquer grupo social ou económico, como é o caso dos sindicatos, partidos políticos, etc. Ex: 'Ordem dos Médicos', 'Associação Comercial de Lisboa', 'Cooperativa dos Operários da Arrábida', 'PSD', 'Sindicato dos Professores do Norte', etc.
8. **Religiosa -** todas as organizações e/ou grupos religiosos. Ex: 'Ordem das Carmelitas Descalças', 'Igreja Católica', 'Testemunhas de Jeová', 'Liga dos Servos de Deus', 'Associação Cristã de Moços', etc.
9. **Socio-Cultural -** organizações vocacionadas para assuntos sócio-culturais ou que trabalham para uma causa pública, como fundações, associações, etc. Ex: Companhia de Teatro 'A Barraca', 'Fundação Calouste Gulbenkian', 'Liga de Amigos da Fundação Aurélio Amaro Diniz', etc.

## *Produtos*

Esta categoria inclui todo o tipo de produtos: comerciais, financeiros, farmacêuticos, industriais, etc. Esta categoria pode ser confundida com a categoria 'Organização-Empresa', no entanto, há uma diferença importante entre elas, já que o 'Produto' deverá

referir-se a um modelo específico, enquanto que a 'Organização' referir-se-á ao produtor do produto. Isto é, no REPENTINO "Ford" seria armazenada na categoria ORGANIZAÇÃO-EMPRESA, enquanto que "Ford Mustang" seria armazenado na categoria PRODUTO-VEÍCULO, uma vez que se refere a um produto comercial específico. Os produtos poderão ser divididos nas seguintes subcategorias:

1. **Ferramentas/Instrumentos -** nesta sub-categoria incluem-se as ferramentas e instrumentos que possuem uma funcionalidade práctica latente e que não estão directamente associados a uma marca, como é o caso de sondas, ferramentas de medição, armamento, etc. Ex: 'contador Geiger', 'Voyager', 'chave Philips', 'detector Geiger-Müller', 'Espectómetro Planetário de Fourier', 'Mars Express', etc.
2. **Consumíveis -** produtos normalmente para consumo próprio que têm uma data limite para serem consumidos, como perfumes, produtos de beleza, bebidas, produtos alimentares, etc. Ex: 'Martini', 'óleo Fula', 'Allure', 'CHANEL Nº 5', 'Coca-Cola Diet', 'creme Nivea', 'Café da Normandia', 'Atum Ramirez', etc.
3. **Electrónica/Electrodomésticos -** aqui incluir-se-ão telemóveis, computadores, televisores, ecrãs, máquinas de barbear, escovas de dentes eléctricas, leitores de CD/DVD, máquinas de lavar, secadores, auscultadores, fogões, microondas, batedeiras, cafeteiras, etc. Ex: 'iPod', 'Nokia 6600', 'Philishave Cool Skin', 'Oral-B Professional Care 7500', 'ecrã LCD', 'HP iPAQ hx2110', etc.
4. **Financeiro -** produtos normalmente fornecidos por entidades bancárias ou seguradoras, como créditos, contas Poupança, seguros, etc. Ex: 'Títulos de Crédito', 'Certificados de Aforro', 'Crédito Habitação BPI', 'Seguro de Vida Allianz', MBNet', etc.
5. **Formato -** formatos e linguagens informáticas. Ex: 'PDF', 'CD', 'DVD', 'VHS', 'XLS', 'TXT', 'Perl', 'Java', etc.
6. **Gastronomia -** sub-categoria onde se incluem todos os elementos da gastronomia nacional e internacional. Ex: 'Açorda Alentejana', 'Pastéis de Belém', 'Pão-de-ló', 'Bacalhau à Zé do Pipo', 'Tripas à Moda do Porto', 'Lasanha', 'Sushi', 'Moqueca de Camarão', 'Moussaka', etc.
7. **Inspecção/Exame -** inspecções, vistorias, assim como exames médicos. Ex: 'Vistoria Sanitária', 'TAC', 'Inspecção Técnica de Veículos', 'Teste do Pezinho', 'Biópsia', 'Ecografia', etc.
8. **Médico/Farmacêutico -** sub-categoria onde se incluem todos os produtos médico-farmacêuticos. Ex: 'Aspirina-C', 'Viagra', 'Ben-u-ron', 'Nimed', 'Actifed', 'Sargenor 5', etc.
9. **Marcas -** Ex: 'AEG', 'Agros', 'Adidas', 'Alardo', 'Aprilia', 'Cadbury's', 'Seiko', 'Armani', 'Auchan', 'Ermenegildo Zegna', 'Miele', 'HP', 'Apple', 'Aquafresh', etc.
10. **Serviços e Recursos -** serviços como linhas telefónicas, Internet ou televisão por cabo, assim como recursos disponíveis online ou em suporte informático, como por exemplo, bases de dados. Ex: 'ADSL', 'NetCabo', 'Projecto Vercial', 'Wikipédia', etc.
11. **Sistemas Informáticos e Aplicações -** sistemas e aplicações informáticas, que normalmente pressupõem um processo de instalação. Ex: 'Adobe Reader 6', 'Internet Explorer 5', 'Microsoft Windows XP', 'Microsoft Word 2003', 'McAfee Anti-Virus', 'Nero6', etc.
12. **Tarefa Manual/Artesanato -** qualquer produto fabricado manualmente. Ex: 'tapetes de Arraiolos', 'Lenços dos Namorados', 'galo de Barcelos', etc.

13. **Vestuário/Utilidades -** peças de vestuário, calçado, acessórios, mas também outras utilidades como brinquedos, material escolar, etc. Ex: 'Barco Pirata Playmobil', 'vestido Fátima Lopes', 'sapatos Jimmy Choo', 'esferográfica Bic', 'cola UHU Stick', 'óculos de sol Valentino', 'relógio Citizen', 'Swatch Ursinhos', etc.
14. **Veículos -** aqui incluem-se todo o tipo de veículos desde os automóveis aos aviões, passando pelos tanques de guerra, pelas motas, bicicletas, trotinetes, barcos, helicópteros, etc. Ex: 'Boeing 747', 'Audi TT 1.8T Roadster', 'Harley Davidson XL 1200C Sportster', 'Flybridge P56', 'helicóptero Alouette II', etc.

## *Seres*

Nesta categoria incluem-se todos os seres reais, ficcionais ou mitológicos, assim como os mitos. É também nesta categoria que se inserem os grupos de pessoas que não constituam claramente uma organização, tais como grupos étnicos e geopolíticos. Os 'Seres' poderão ser divididos nas seguintes subcategorias:

1. **Colectivo Humano -** grupos de humanos (reais ou ficcionais) conhecidos normalmente como grupo, isto é, grupos cuja identidade de grupo é mais forte do que a identidade individual dos seus membros, como é o caso de equipas, bandas, duos, famílias, etc. Ex: 'The Rolling Stones', 'Bonnie and Clyde', 'os irmãos Dalton', 'os Sete Anões', 'família Melo Campos', 'os Vieira de Mello', etc.
2. **Geopolítico/Étnico/Ideológico -** grupos de pessoas (reais ou ficcionais) que partilhem a mesma identidade geográfica, política, étnica ou ideológica, embora não pertençam a uma organização estruturada. Ex: 'Incas', 'Budistas', 'Dadaístas', 'Nudistas' 'Marcianos', 'Atlantes', 'Visigodos', etc.
3. **Humano -** qualquer pessoa (real, ficcional ou mito) viva ou morta. Ex: 'Mr. Bean', 'Othello', 'Branca de Neve', 'Brad Pitt', 'Papa', 'Rainha de Inglaterra', 'Ulisses', 'Mick Jagger', 'Vladimir Putin', 'Adalberto Alves', etc.
4. **Mitológico -** toda e qualquer entidade mitológica. Ex: 'Pégaso', 'Minotauro', 'Ícaro', 'Adamastor', 'Afrodite', 'Cupido', etc.
5. **Não-Humano -** qualquer ser (real ou ficcional) que não seja humano, vivo ou morto, como é o caso dos animais de estimação, monstros, etc., com excepção das entidades mitológicas. Ex: 'Laika' (primeira cadela no espaço), 'Bambi', 'Lassie', 'Winnie the Pooh', 'Monstro do Lago Ness', 'Samwise Gamgee', 'Pantera Cor-de-Rosa', 'Gollum', 'Donald', etc.

## *Substâncias*

Nesta categoria incluem-se substâncias, elementos e minérios. As 'Substâncias' poderão ser divididas nas seguintes subcategorias:

- **Grupo -** aqui incluem-se grupos de substâncias. Ex: 'álcool', 'cetonas', 'aldeídos', 'monossacarídeo', 'esterol', 'glicose', etc.
- Minério - incluem-se aqui todos os minérios, assim como pedras preciosas. Ex: 'urânio', 'Ágata cornalina', 'rubi', 'opalina', 'pirite', 'pedra de Moleanos', etc.
- **Substância -** Ex: 'Paracetamol', 'H2O', 'anilina', 'penicilina', 'ácido ascórbico', 'acetilsalicilato de lisina', 'boldenona', 'hematoxilina', 'lecitina de soja', 'lidocaína', 'Mebendazol', 'nandrolona', 'Oxibutinina', etc.

## *Outros*

Nesta categoria inserem-se dois tipos de elementos:

- exemplos que não foram encaixados em nenhuma das categorias anteriores e que aguardam a existência de uma categoria própria; ou
- particulas e unidades léxicais que podem ser úteis para tarefas de classificação de entidades nomedas / mencionadas, como por exemplo nomes de unidades de medida, moedas, profissões, títulos pessoais, etc.

# O REPENTINO como ferramenta colaborativa

Um dos objectivos principais do REPENTINO é o de fornecer uma amostra representativa de exemplos de entidades nomeadas que sirva de base ao desenvolvimento de sistemas de reconhecimento de entidades nomeadas. A maior parte dos almanaques disponíveis para este efeito são compilados tematicamente e apresentam normalmente dois tipos de problemas:

1. Cobertura restrita à área base em que foram compilados
2. Enorme desproporção entre o número de entidades armazenadas e aquelas que é normalmente possível encontrar realmente em texto.

Para minimizar estes dois problemas, e para evitar um enviesamento do recurso relativamente à nossa perspectiva, foi decidido abrir a construção do REPENTINO ao público em geral, sendo que para isso se construiu uma interface Web destinada à recolha de sugestões via rede. Procurou-se assim resolver vários problemas.
Em primeiro lugar, os exemplos sugeridos podem permitir a abertura de novas categorias que não tivessem sido previstas por nós. As ideias e os conhecimentos espalhados pela comunidade são obviamente muito mais abrangentes que aquilo que seríamos capazes de idealizar sozinhos, permitindo assim recolher uma colecção muito mais abrangente de exemplos. Por outro lado, as sugestões realizadas pela comunidade são tendencialmente mais úteis, já que reflectem o conhecimento das entidades relevantes e frequentes. As sugestões realizadas serão previsivelmente e principalmente entidades sobre as quais realmente se produz referência, já que são estas que populam o conhecimento colectivo da comunidade. Desta forma, diminui-se o segundo problema apontado anteriormente. Por outro lado, se o volume de sugestões for elevado, poderemos até manter um desenvolvimento distribuído deste recurso o que facilitará a sua evolução sustentada. Este objectivo parece um pouco mais difícil mas ultimamente tem-se assistido a vários exemplos de construção colectiva de recursos, nomeadamente o Wikipedia, pelo que não seria surpreendente se um fenómeno semelhante, embora a uma escala muito menor, ocorresse com o REPENTINO.
Durante o período de redacção deste relatório foi feita alguma publicidade informal ao REPENTINO via e-mail, focando sobre o conjunto de contactos próximos dos elementos do Pólo. Nestes contactos foi explicado informalmente o objectivo do REPENTINO e foi também pedida colaboração com a sugestão de alguns exemplos de nomes de entidades que fizessem essencialmente parte do domínio de conhecimento do colaborador. Este detalhe é importante, pois optimiza a capacidade de recolha de exemplos de entidades nomeadas associadas a domínios específicos que normalmente são de difícil compilação.
Não temos ainda neste momento resultados que possam medir o impacto destas medidas publicitárias mas estamos optimistas relativamente às ideias que deverão surgir em função das sugestões feitas pelos eventuais colaboradores.


Figura 3 - Página de abertura da interface de utilizador do REPENTINO

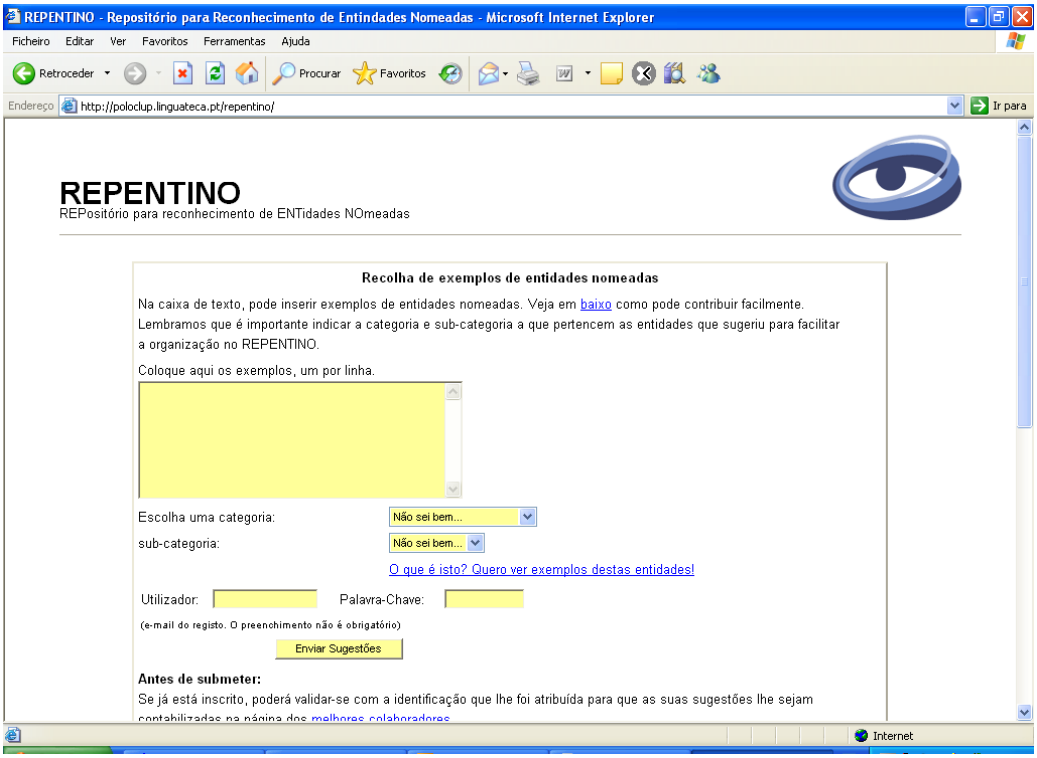


Figura 4 - A  interface de submissão de sugestões

## Números do REPENTINO

Neste momento, o REPENTINO armazena cerca 450 mil exemplos de entidades armazenados pelas 11 categorias de topo. Na próxima tabela são apresentados os valores da distribuição dos exemplos.

Tabela 1 - A distribuição do exemplos por categorias (valores absolutos)

| Categoria | # |
|---|---|
| Abstracções | 5807 |
| Arte/Media/Comunicação (A/M/C) | 15232 |
| Eventos | 25357 |
| Locais | 49451 |
| Outros | 1771 |
| Natureza | 867 |
| Organizações | 46869 |
| Papeladas | 4427 |
| Produtos | 9199 |
| Seres | 286297 |
| Substâncias | 1468 |
| Total | **446745** |

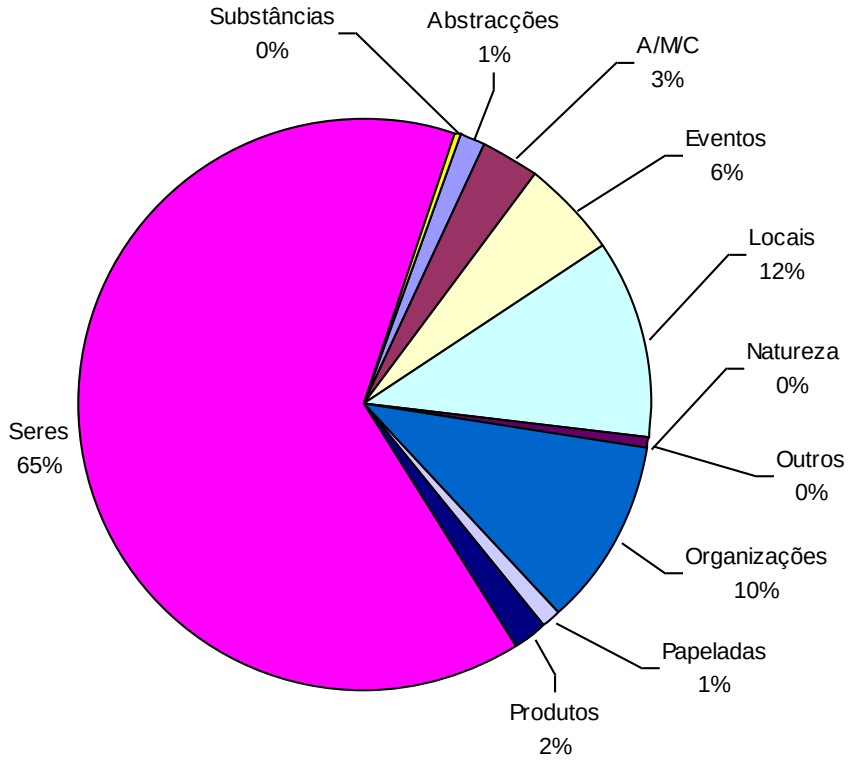


Figura 5 - A distribuição do exemplos por categorias (valores relativos)

Destes dados destaca-se a forte contribuição da categoria Seres (e em particular da subcategoria Humano) que totalizam cerca de dois terços do total de entidades armazenadas no REPENTINO. Também importantes são as contribuições das categorias Locais, Organizações e Eventos. Menos significativas mas ainda notórias são as categorias Arte/Media/Comunicação, Produtos e Papeladas e Abstracções. Quase sem expressão encontramos as restantes categorias (Natureza, Substâncias e Outros). Daqui

se pode concluir que há um forte desiquilibrio na distribuição de exemplos o que sugere que algumas medidas de correcção poderão ser necessárias. Uma dessa medidas poderá passar pela eliminação de uma porção significativa dos exemplos da subcategoria Seres::Humanos, que será neste momento certamente caracterizada por uma elevada redundância.
Dados actualizados e mais detalhados sobre cada uma destas categorias, incluindo a distribuição por subcategorias podem ser consultados na página de estatísticas do REPENTINO.

## Planos futuros para o REPENTINO

A génese do REPENTINO está intimamente ligada ao SIEMÊS mas pensamos que tem potencialidades para ser um recurso autónomo e interessante para outros usos que não apenas o actual SIEMÊS.
Considerando primeiro as possibilidades de interacção com o SIEMÊS, seria importante obter uma versão simplificada do actual REPENTINO que funcionasse como base de conhecimento do SIEMÊS. Esta versão simplificada seria composta por um número de exemplos bastante inferior ao actual, mas deveria manter ainda assim um elevado nível de abrangência e representatividade. Tal recurso, uma forma mínima do REPENTINO permitiria um significativo aumento do desempenho computacional do SIEMÊS sem alteração do seu modo de funcionamento, o que facilitaria a sua aplicação a grandes colecções. Um exemplo de uma possível simplificação seria a redução do número de exemplos de nomes próprios de pessoas armazenados no REPENTINO dos actuais cerca de 280 mil para um número muito inferior (20%?) que ainda assim manteria exemplos de todos os unigramas ou bigramas existentes nos exemplos. Talvez fosse possível executar uma operação semelhante em muitas das subcategorias do REPENTINO.
Relativamente à melhoria do próprio recurso, há vários acrescentos e melhorias que parecem ser importantes e que permitiriam enriquecer o recurso com informação útil para outros estudos.
Em primeiro lugar, seria importante obter informação acerca da frequência de cada um dos exemplos armazenados, tendo em conta as ocorrências em corpora ou na rede. Uma possibilidade de conseguir obter essa informação passaria por contar o número de ocorrências no WPT03 de cada uma das entidades, recorrendo ao BACO (BAse de Co-Ocorrências). Testes realizados com uma versão simplificada do BACO mostraram ser viável obter essa informação em tempo útil, já que as pesquisas sobre a totalidade de colecção podem ser realizadas em média em menos de 20 segundos por entidade. Apesar deste tempo poder parecer elevado (apenas 3 a 4 exemplos por minuto), especialmente tendo em conta o elevadíssimo número de entidades já armazenado no REPENTINO, não o será se considerarmos que esta informação é particularmente interessante apenas para algumas subcategorias do repositório, nomeadamente organizações, locais e eventos, que apesar de tudo representam um pequena porção do REPENTINO. Em todo o caso, esta informação pode ser obtida por fases usando uma estratégia distribuída (BACO instalado em várias máquinas do Pólo do Porto).
Uma outra possibilidade interessante, sob a qual já se realizaram algumas experiências, é a extracção de contextos existentes para cada um dos exemplos. Novamente, isso é possível concretizar usando o BACO, já que o processo de recolha dos contextos não difere muito do da obtenção das contagens, podendo inclusive ser realizado em simultâneo. O objectivo desta recolha de contextos é o de poder permitir estudar com algum detalhe situações interessantes e proveitosas para tarefas de identificação de entidades nomeadas / mencionadas. Por exemplo, o agrupamento dos contextos

associados a todos os exemplos de uma determinada subcategoria permite obter uma panorâmica das várias possibilidades de como as entidades dessa categoria podem ser mencionadas. A mesma informação pode também ser usada para criar padrões que permitam a identificação / recolha em corpora de novos exemplos da mesma subcategoria. Não foram realizadas ainda experiências a este nível mas parece-nos que existe aqui algum potencial por explorar.
Finalmente, parece ser relativamente simples executar um cruzamento entre o REPENTINO e o BACO e gerar um índice sobre o BACO usando as entradas do BACO. O BACO desta forma expandido permitiria todo um conjunto de pesquisas rápidas que envolvessem os exemplos, por exemplo testar co-ocorrências ou relações entre exemplos/entidades, possibilitando um novo ambiente de estudo para os tópicos associados ao reconhecimento de entidades nomeadas / mencionadas.

## Conclusões

O REPENTINO é um recurso em construção cujo interesse foi parcialmente validado através da sua utilização no SIEMÊS. O REPENTINO tem crescido e aposta numa estratégia de construção colaborativa, embora não esteja ainda provado que essa possibilidade seja eficaz e produtiva. Contudo, trata-se de um recurso que permite que desenvolvedores de sistemas de reconhecimento de entidades nomeadas / mencionadas para português não partam do zero, quer na modelização do problema, quer na necessidade de criar recursos auxiliares. O sistema de classificação do REPENTINO está relativamente evoluído e prevê mais de 100 subcategorias, o que o torna bastante abrangente e detalhado. O conteúdo do REPENTINO é também abundante e parece ser capaz de cobrir representativamente uma percentagem significativa do problema. Há uma série de possibilidades ainda por explorar que poderão ser úteis em diversos contextos num futuro próximo.

## Agradecimentos

Este trabalho foi financiado pela Fundação para a Ciência e Tecnologia, co-financiada pelo POSI, através do projecto POSI/PLP/43931/2001