Novos rumos para a recuperação de informação geográfica em português

Nuno Cardoso
Universidade de Lisboa, Faculdade de Ciências, Laboratório LaSIGE
ncardoso@xldb.di.fc.ul.pt

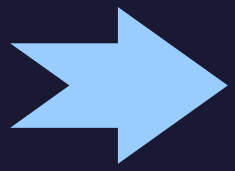
Encontro 10 anos da Linguateca – 11 de Setembro de 2008, Aveiro, Portugal

Motivação

- Estamos a entrar numa terceira geração de motores de busca. [Broder, 2007]
- "Dá-me o que eu quero", em vez "dá-me o que eu disse". [Singhal, 2008]
- Maior foco às necessidades de cada um. [Belkin, 2008]
  - Mistura de textos com imagens, mapas, vídeos, ...
  - Histórico de consultas, perfil do utilizador.
  - Respostas mais inteligentes (sumarização documentos, resposta automática a perguntas).

Motivação

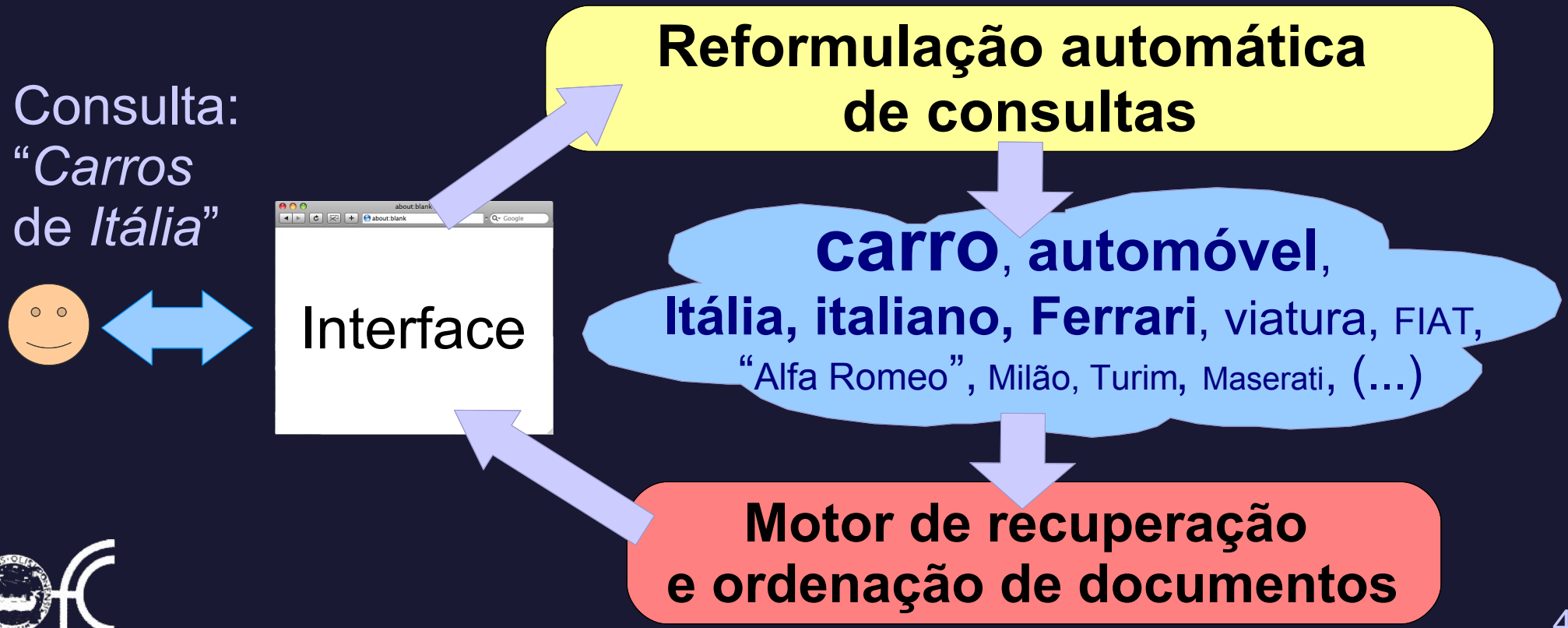
Encontrar os termos da consulta



Compreender as intenções do utilizador

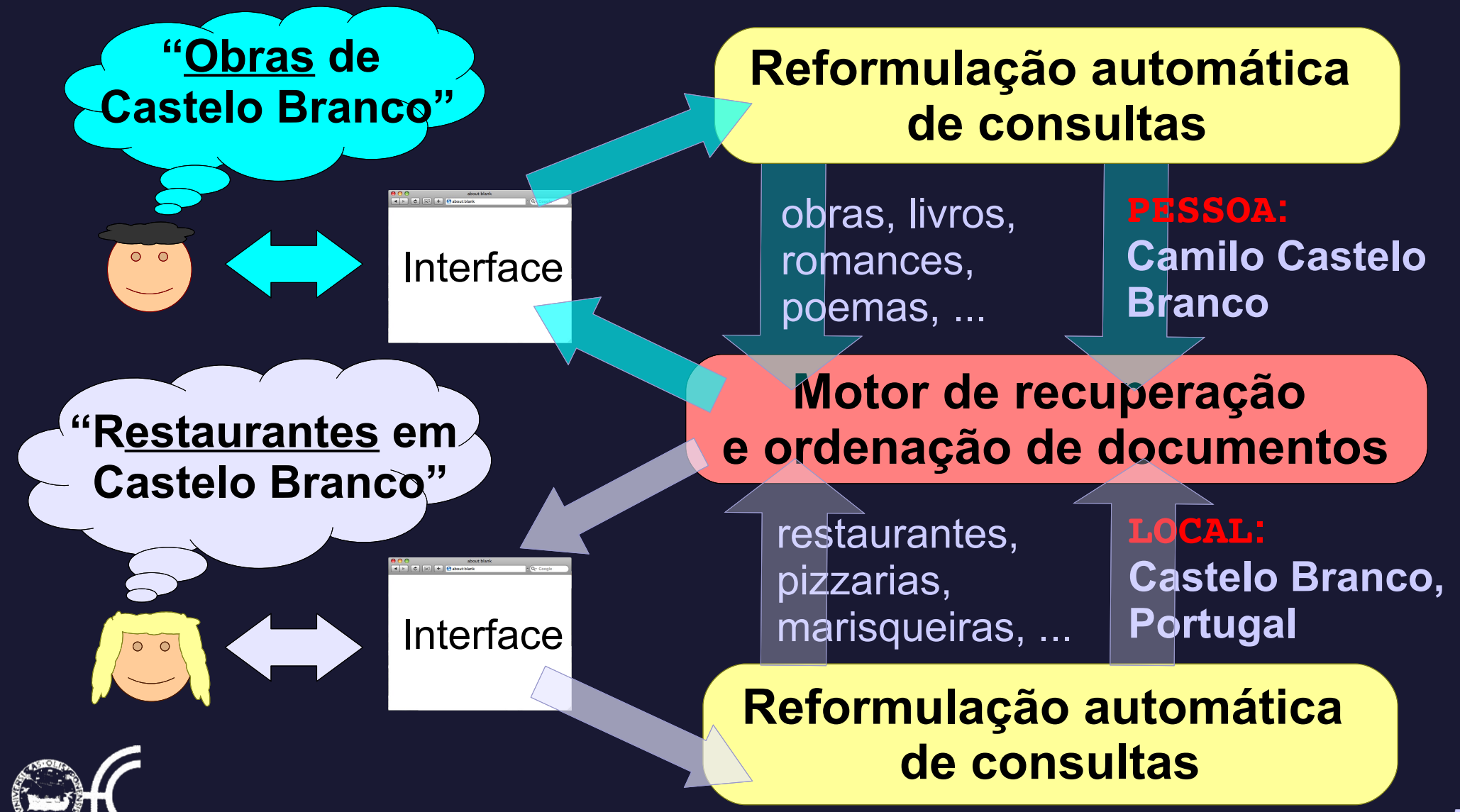
Objectivos da minha tese

- Desenvolver novas formas semânticas de reformulação automática de consultas dos utilizadores, e aplicar na recuperação de informação geográfica em português.



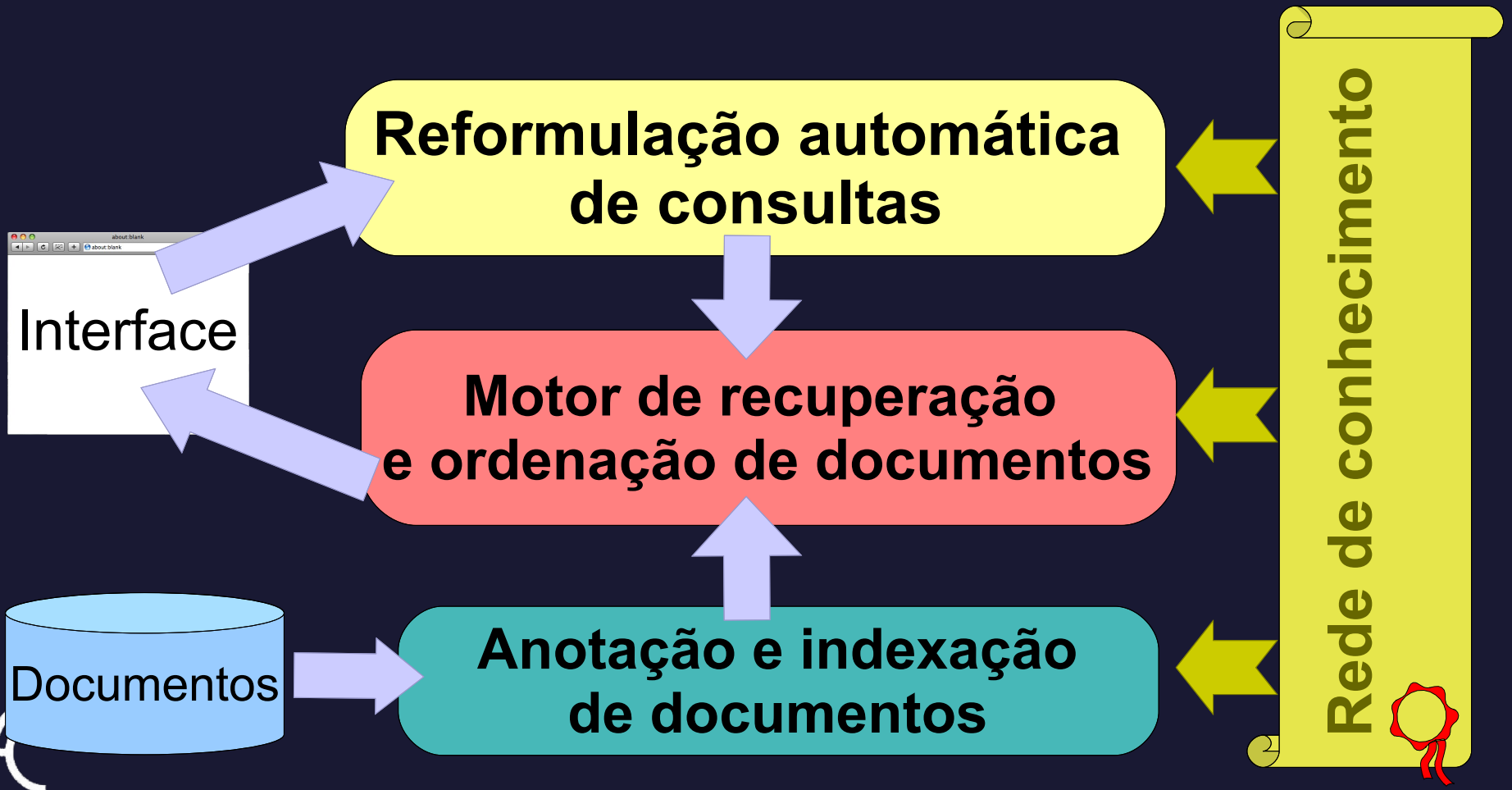
Compreendendo as consultas...

Exemplo típico: consultas com “Castelo Branco”



Rede de Conhecimento

- Rede semântica composta por fontes de informação em português.
- Base para a extracção de conhecimento.



Fontes de informação

Ênfase em fontes ricas no domínio geográfico.

Sítios:
1. www.cm-lisboa.pt
2. pt.wikipedia.org/wiki/Lisboa
3. www.atl-turismolisboa.pt/

Títulos:
”Câmara Municipal de Lisboa”, “Lisboa”, “Associação de Turismo de Lisboa”, ...

Co-ocorrências entre consultas:
“cidade”, “hotéis”, “benfica”, “pousadas”, “farmácia”, “Lisboa”, “metro”, “turismo”, ...

Escolhas dos utilizadores:
1. www.cm-lisboa.pt
2. www.metrolisboa.pt
...

Área: 84.8 km²

Coordenadas:
38°42' N, 9°11' O

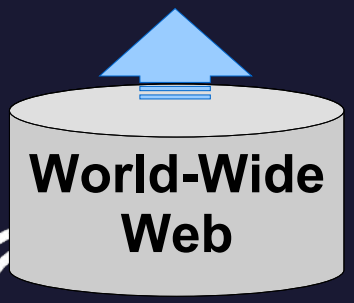
População: 564,477

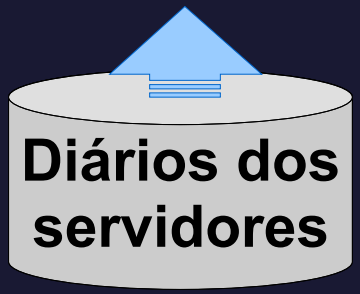
Listas de freguesias, castelos, universidades, museus, parques, ...

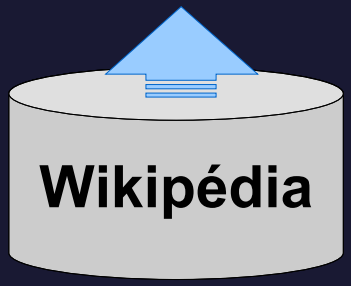
Parte de:
Portugal (*tipo*: país)

Adjacente a:
Tejo (*tipo*: rio)

Contém:
Portela
(tipo: aeroporto)

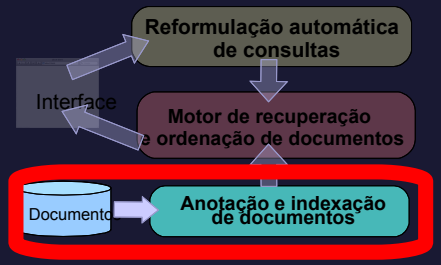






Ponto da situação: Reformulação automática de consultas

- QuerCol: Módulo de reformulação de consultas.
- Estratégias diferentes para termos geográficos e termos não-geográficos.



Ponto da situação: Anotação de documentos

- **REMBRANDT**: Reconhecimento de entidades mencionadas.
- Usa a Wikipédia + regras gramáticas para reconhecer EM.

Wikipédia

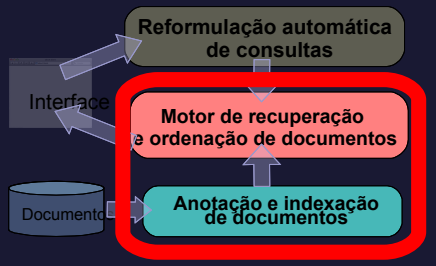
Documentos iniciais

Eu visitei a Torre dos Clérigos, num passeio que fiz ao Porto.

REMBRANDT

Documentos anotados

Eu visitei a **<CONSTRUÇÃO>** Torre dos Clérigos**</CONSTRUÇÃO>** ***<LOCAL*>***Porto ***</LOCAL*>***, num passeio que fiz ao **<LOCAL>**Porto**<LOCAL>**.



Ponto da situação: Indexação e recuperação geográfica de docs.

- **MG4J**: Indexação e ordenação de documentos.
- Indexa separadamente as EM anotadas do **REMBRANDT**.

Documentos anotados

Eu visitei a **<CONSTRUÇÃO>** Torre dos Clérigos**</CONSTRUÇÃO>** ***<LOCAL*>Porto</LOCAL*>***, num passeio que fiz ao**<LOCAL>**Porto**<LOCAL>**.

MG4J

Índices

textual

a, ao, clérigos, dos, eu, fiz, num, passeio, porto, que, torre, visitei.

geográfico: implícito

porto

geográfico: explícito

porto

Outros trabalhos em curso

- **RENOIR** – Módulo de experiências para a geração de consultas semânticas.
  (ex: PESSOA? presidente de LOCAL:Portugal)
- Aproveitar as anotações do REMBRANDT para enriquecer automaticamente as consultas.
- Usar conhecimento (ex: DBpedia [Auer et al, 2007]) para assistir a reformulação das consultas.

- **RENOIR: xldb.di.fc.ul.pt/Renoir**
- **REMBRANDT: xldb.di.fc.ul.pt/Rembrandt**

Considerações finais

Consultas com âmbitos geográficos são frequentes. Os motores de busca precisam de se adaptar a esta realidade. Como?

- Compreendendo as consultas do utilizador.
- Explorando diversas fontes de informação, extraindo conhecimento, raciocinando sobre o domínio geográfico.
- Personalizando os resultados de acordo com o contexto de pesquisa de cada utilizador.

Fim.

Questões?

Novos rumos para a recuperação de informação geográfica em português

Nuno Cardoso
Universidade de Lisboa, Faculdade de Ciências,
Laboratório LaSIGE
ncardoso@xldb.di.fc.ul.pt

Encontro 10 anos da Linguateca – 11 de Setembro de 2008, Aveiro, Portugal

Referências

- Andrei Broder, “The Next Generation Web Search and the Demise of the Classic IR model “, ECIR 2007, Roma, Itália, Abril 2007
- Amit Singhal. “Web Search: Challenges and Directions”, ECIR 2008, Glasgow, Escócia, Abril 2008.
- Nicholas J. Belkin. “Some(what) Grand Challenges for Information Retrieval”, ECIR 2008, Glasgow, Escócia, Abril 2008.
- Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak e Zachary Ives, “DBpedia: A Nucleus for a Web of Open Data”, ISWC 2007 + ASWC 2007, Busan, Coréia, 2007