Em Direção à Construção de uma Ontologia Geográfica a partir de Textos em Português na Web Portuguesa

Marcirio Silveira Chaves

Simpósio doutoral - Linguateca

10 de abril de 2006

Ligação com o Simpósio doutoral 2005

- Plano do doutorado
- Construção de uma ontologia geográfica a partir de fontes de informação de autoridades
- GKB
- Ontologias

Estrutura da apresentação

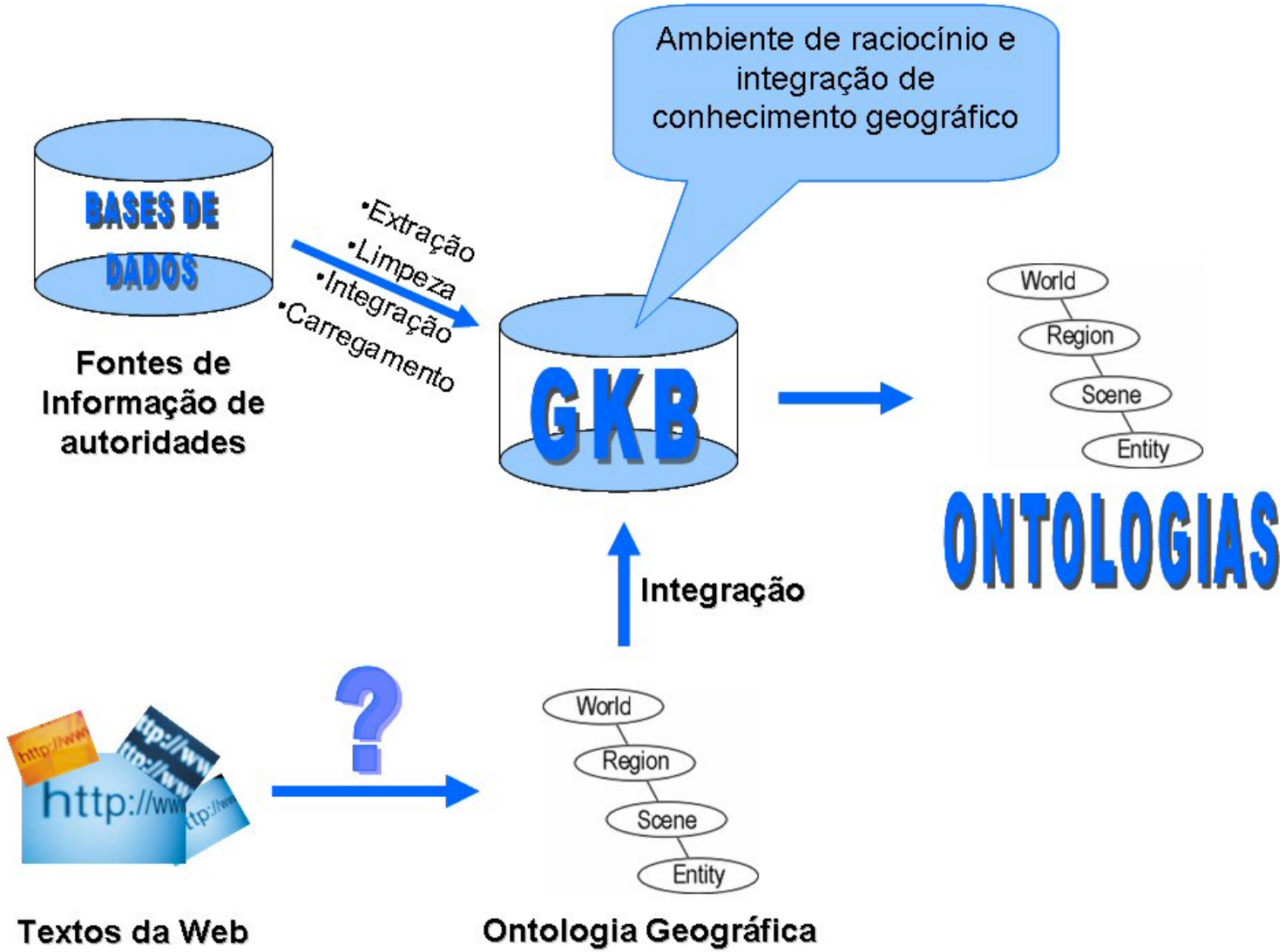
Motivação

Problemas de pesquisa:

- Coleta, identificação, limpeza, classificação, integração e formalização da informação geográfica (administrativa) sobre Portugal
- Informação Formal - Carência de informação geográfica integrada e detalhada formal
- Informação Informal - Nomes e relações geográficas informais

Motivação

- Carência de informação integrada, com qualidade e sem custo
- Ontologia: conceito fundamental na arquitetura da Web Semântica
- "***Assembling data*** *is* ***no*** *longer the biggest challenge. Instead,* ***the major hurdle these days is one of data integration***." Russ Altman, Stanford


Ambiente de raciocínio e integração de conhecimento geográfico
BASES DE DADOS
•Extração
•Limpeza
•Integração
•Carregamento
Fontes de Informação de autoridades
GKB
World
Region
Scene
Entity
ONTOLOGIAS
Integração
http://www
Textos da Web
World
Region
Scene
Entity
Ontologia Geográfica

Fases do Trabalho

1. Criação da GKB
2. Caracterização da "geograficidade" existente nos textos
3. Extração de conhecimento geográfico
4. Criação de ontologia geográfica
5. Integração do conhecimento obtido em 4.

Objetivos Subsequentes

- Caracterizar a informação geográfica que existe na web portuguesa
- Identificar termos e relacionamentos geográficos em textos
- Formalizar os termos e relacionamentos encontrados
- Construir a ontologia geográfica

Hipótese

Existe informação geográfica nos textos da web portuguesa que pode ser usada para estender uma ontologia geográfica derivada das fontes de informação oficiais.

Recursos disponíveis

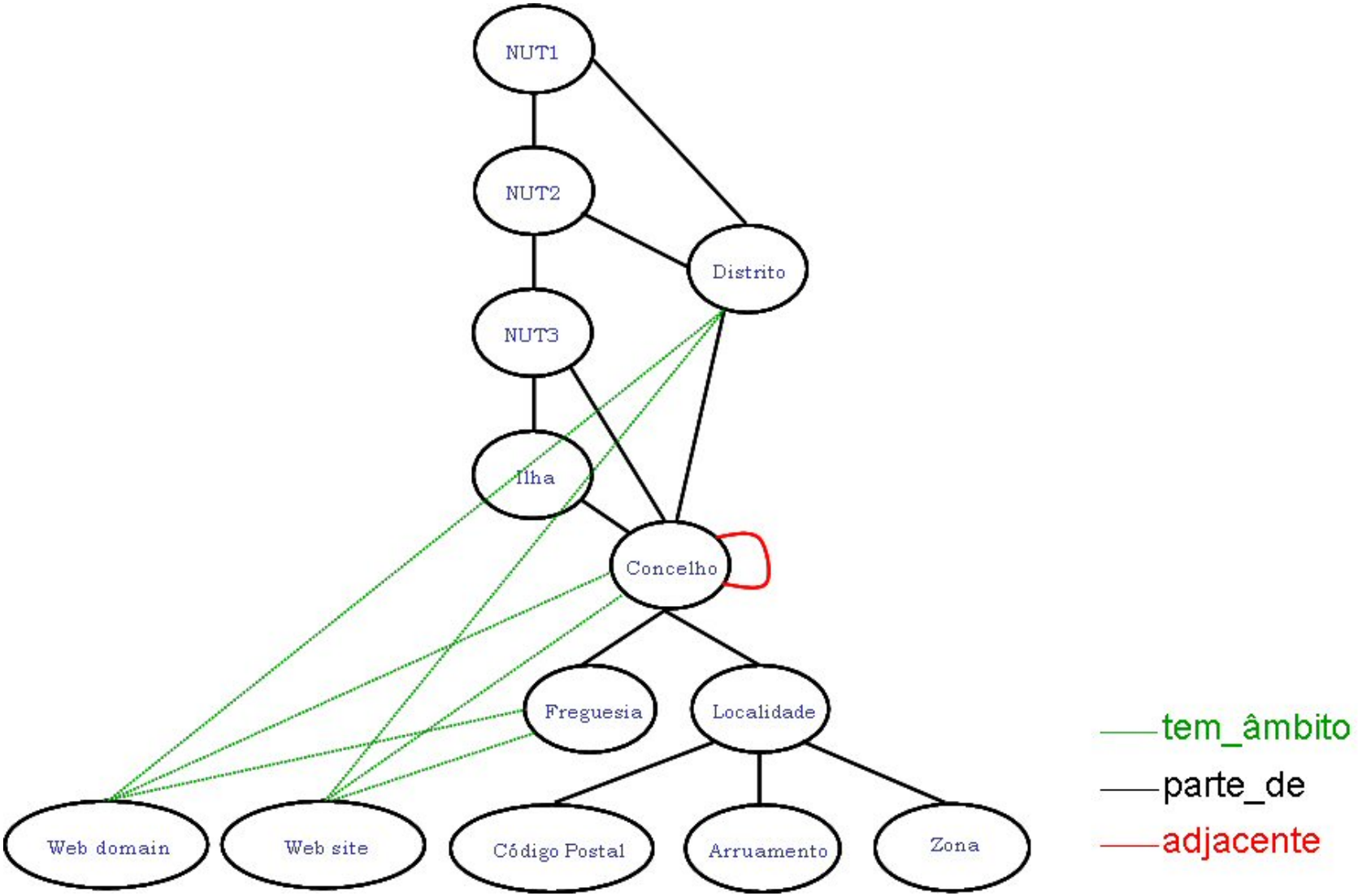
- GKB / Geo-Net-PT01
- WPT 03
- CAGE
- SIEMÊS
- BACO

GKB – *Geographic Knowledge Base*

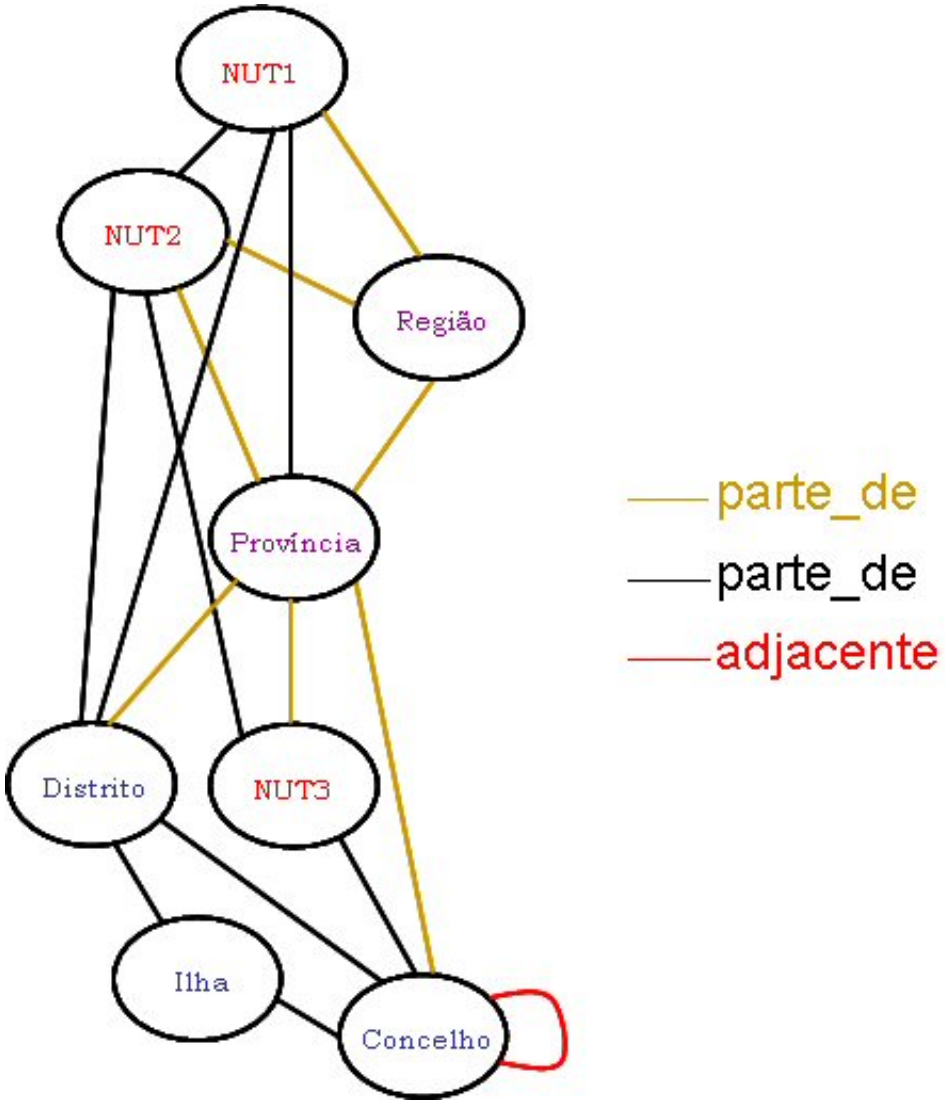
- KB formada por fontes de informação distintas, heterogêneas e complementares
- Informação geográfica e de rede
- Mais de 800.000 registros
- Exportada como ontologias
- Geo-Net-PT01

Feature: Um objeto com significado no domínio selecionado do discurso [ISO19109].
Ex.: países, cidades e localidades

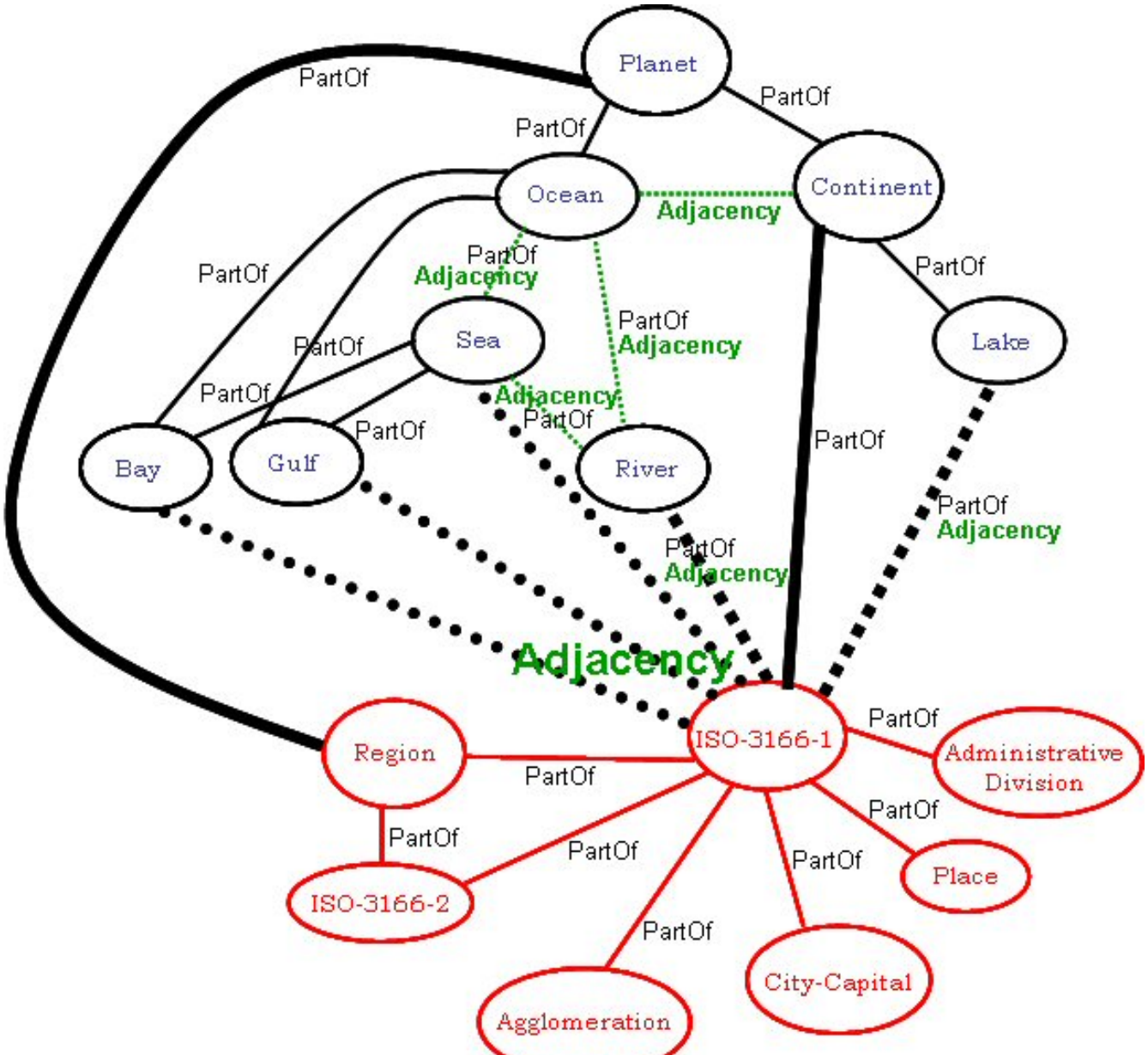
GKB


NUT1
NUT2
Distrito
NUT3
Ilha
Concelho
Freguesia
Localidade
Web domain
Web site
Código Postal
Arruamento
Zona
tem_âmbito
parte_de
adjacente

GKB


NUT1
NUT2
Região
Província
Distrito
NUT3
Ilha
Concelho
parte_de
parte_de
adjacente

GKB


Planet
Ocean
Continent
Sea
Lake
Bay
Gulf
River
Region
ISO-3166-1
Administrative Division
ISO-3166-2
Place
City-Capital
Agglomeration
PartOf
Adjacency

GKB — *Geographic Knowledge Base*

Tipo de local	# ocorrências distintas	# de multi-palavras	Sobreposição	Exemplos
NUT1	3	2	3	Continente, Açores, Madeira
NUT2	7	2	7	Norte, Centro, Algarve
NUT3	30	22	11	Grande Porto, Grande Lisboa, Alentejo Central
distrito	**18**	3	**18**	Porto, Setúbal, Beja
concelho	**308**	121	**308**	Lisboa, Sintra, Lagos
ilha	**11**	11	**11**	Ilha das Flores, Ilha do Pico, Ilha da Graciosa
freguesia	3,595	1,462	2,876	Meca, Pego, Mina
localidade	26,924	16,073	7,584	Igreja, Cabana, Horta
zona	3,594	2,392	1,737	Santana, São Bento, Forca
arruamento	75,946	51,087	27,805	Travessa Azenha, Rua Azenha, Beco das Flores
Total	110,436	71,175	-	

Estatísticas sobre as ontologias criadas

Estatística	Portugal	World
# de features	**418.065**	**12.293**
# de relacionamentos	**419.867**	**12.258**
# de relacionamentos parte-de	**418.340 (99,83%)**	**12.245 (99,89%)**
# de relacionamentos de equivalência	395 (0,09%)	2.501(20,40%)
# de relacionamentos de adjacência	1.132 (0,27%)	13 (0,10%)
Média de features mais abrangentes por feature	1	1,07
Média de features mais específicas por feature	10,56	475,44
Média de features equivalentes por feature com equivalente	1,99	3,82
Média de features adjacentes por feature com adjacente	**3,54**	6,5
# de features sem ascendentes	3 (0%)	1(0.00%)
# de features sem descendentes	374.349 (89,54%)	12.045 (97,98%)
# de features sem equivalentes	417.867 (99,95%)	11.819 (96,14%)
# de features sem adjacentes	417.739 (99,92%)	12.291 (99,99%)

Frequência dos tipos de arruamentos no WPT 03

Mais frequentes			Menos frequentes	
Tipo	**ocorrências distintas (%)**	**Freq. WPT 03**	Tipo	ocorrências distintas
Rua	91.310 (62,36)	410.576	Ruela	18
Travessa	18.150 (12,40)	288.045	Carreira	20
Largo	7.284 (4,97)	237.234	Acesso	24
Praceta	3.749 (2,56)	213.643	Adro	30
Avenida	3.630 (2,48)	194.700	Recanto	42
Beco	3.426 (2,34)	181.721	Cais	43
Estrada	2.317 (1,58)	138.988	Ponte	46
Bairro	2.009 (1,37)	129.609	Campo	46
Caminho	1.450 (0,99)	129.169	Lugar	47
Praça	1.358 (0,93)	93.852	Via	56

Experimentos

Descrição:

- 32.000 documentos etiquetados pelo SIEMÊS
- Pessoas, Organizações e Locais

Resultados das EMs detectadas em uma amostra aleatória de 32.000 documentos do WPT 03

MP = multi-palavra
EMD = entidade mencionada distinta
GN = Geo-Net-PT01

	# de EMs (%)	# de EMDs	# of MP EMs (%)	# de MP EMDs (%)	# EMDs MP contendo um nome na GN (%)	# de EMDs ocorrendo na GN (%)
PES	250.585 (26,48)	77.228	140.155 (55,93)	58.991 (76,39)	24.105 (31,21)	521 (0,67)
ORG	418.915 (44,27)	114.353	214.698 (51,25)	89.790 (78,52)	26.789 (23,43)	462 (0,40)
LOC	276.775 (29,25)	47.972	90.018 (32,52)	36.395 (75,87)	22.959 (47,86)	4.576 (**9,53**)
Total	**946.275** (100)	239.553	444.871 (47,01)	185.176(**77,30**)	73.853 (30,83)	5.559 (2,32)

Análise

- Perto de 1 milhão de EMs anotadas em três categorias, 30% Locais
- Mais de 75% das EMDs são multi-palavra
- Locais se repetem mais do que nomes de pessoas
- Sobreposição com a Geo-Net-PT01
  - Ambigüidade com nomes de pessoas e organizações menor do que 1%
  - Entretanto, 31.21% das pessoas e 23.43% das organizações contêm um nome geográfico na Geo-Net-PT-01 (27.855 nomes da Geo-Net-PT-01 utilizados. Não consideramos nomes de arruamentos e códigos postais).
  - Somente cerca de 10% dos Locais reconhecidos estão na Geo-Net-PT-01

Distribuição das EMs na amostra de 32.000 documentos

- Número de documentos com ao menos uma EM: 31.489 (98.4% da amostra).
  - Pessoas reconhecidas em 21.499 (67,18%) documentos
  - Organizações reconhecidas em 30.328 (94,77%) documentos
  - **Locais reconhecidos em 24.468 (76,46%) documentos**

Informação geográfica é transversal ao domínio de conhecimento do texto

Tabela: Estatística descritiva sobre o experimento

	Total	Distintas		Total	Distintas
Média PESs. por doc. com PESs.	11,65	7,82	Mediana LOCs	4	3
Média ORGs. por doc. com ORGs.	13,81	9,78	Desvio Padrão LOCs	149,7	57,54
Média LOCs. por doc. com LOCs.	11,31	**7,34**	# docs. com 1 LOC	5,443	6.184
Média EMs por doc. com EMs	30,04	**20.47**	# docs. $>$ 3 LOCs	12.913	11.640
# máximo de LOCs em 1 doc.	20.594	6.472	# docs. $>$ 30 LOCs	1.483	713

- Ps.: Os valores da coluna Distintas medem as EMDs dentro de cada documento.

Análise

- Cada documento contendo ao menos uma EM contém em média cerca de 20 EMDs,
- das quais mais de sete são locais
- e cerca de 50% dos documentos contendo Locais contém mais de três Locais

Experimento: SIEMÊS e CAGE

Descrição:

- 1.000 documentos selecionados aleatoriamente
- Sistemas REM: CAGE (CAGE-WPT e CAGE-PT) e SIEMÊS

Tabela: Estatística descritiva: comparação entre sistemas REM

		CAGE-WPT	CAGE-PT	SIEMÊS
Total	Total LOCs	6.701	4.395	2.635
	LOCs distintas	958	640	**981**
	# de docs. com LOCs	690	450	592
	Média LOCs por doc.	6,70	4,40	2,63
	Média LOCs por doc. com LOC	9,70	9,74	4,45
	Mediana LOCs	3	2	2
	Desvio padrão LOCs	28,67	32,40	7,5
	# máximo de LOCs	244	232	79
Distintas	Média LOCs por doc.	0,95	0,64	0,98
	Média LOCs por doc. com LOC	**7,11**	**7,49**	**3,41**
	Mediana LOCs	2	2	2
	Desvio padrão LOCs	23,74	27,6	4,2
	# máximo de LOCs	208	198	44

Análise ...

- SIEMÊS reconheceu mais LOCs distintas
- SIEMÊS e CAGE-WPT reconheceram LOCs em pelo menos 60% dos documentos
- Existe, em média, no mínimo 3 LOCs distintas por documento com LOC considerando os 3 sistemas
- Considerar *overtagging*

Aprofundando a análise do experimento: sobreposição entre os sistemas REM

Tabela: Sobreposição entre sistemas de REM

	O_1	% CAGE-WPT	% SIEMÊS	O_2	% CAGE-PT	% SIEMÊS
# LOCs distintas em comum	175	**18,26**	**17,83**	98	**15,31**	**9,99**
# docs. com LOCs distintas em comum	214	31,01	36,15	131	29,11	22,13
Média de LOCs comum por doc.	1,98	-	-	1,86	-	-
Mediana LOCs	1	-	-	1	-	-
Desvio Padrão	1,96	-	-	1,91	-	-
# máx. de LOCs	9	-	-	9	-	-

Análise ...

- O número de LOCs comuns em O_1 e O_2 é baixo
  - Vocabulário de nomes é diferente (quanto?)
- Menos de 20% das LOCs distintas em O_1 e O_2 estão sobrepostas
- Considerar *overtagging*

Tabela: Distribuição das LOCs comuns nos documentos

		# de LOCs comuns									
		1	2	3	4	5	6	7	8	9	Total
# de docs.	CAGE-WPT/SIEMÊS	146	29	8	6	6	2	10	3	4	214
	CAGE-PT/SIEMÊS	99	11	4	2	2	1	10	1	1	131

Construção de ontologia a partir de textos da web

Abordagens semi-automáticas utilizadas

- Extração baseada em padrões léxico-sintáticos
- Regras de associação
- Agrupamento conceitual
- Aprendizagem automática

Tarefas envolvendo ontologias

- Construção
- Extensão
- População
- Mapeamento
- ...

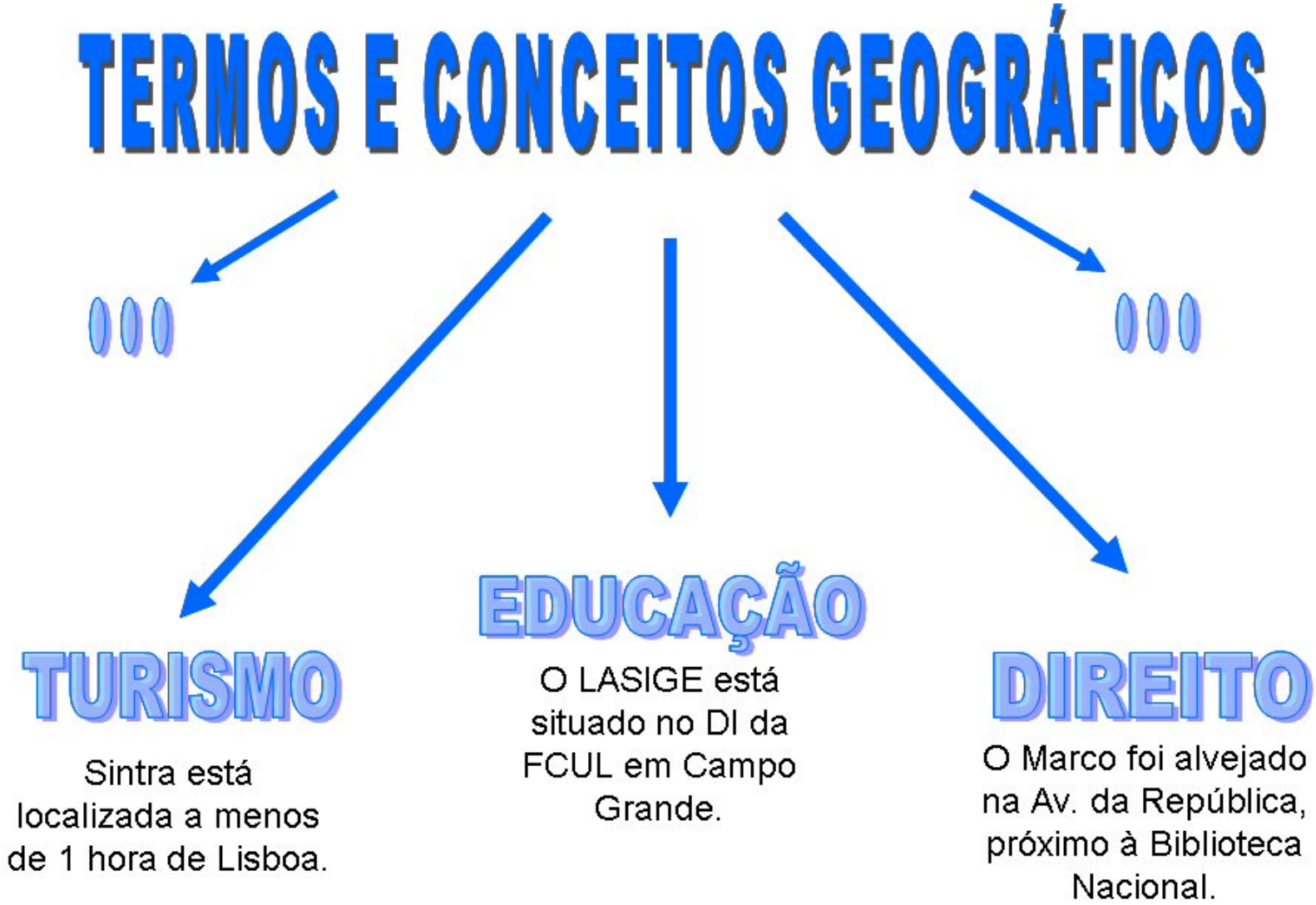
Tarefas envolvendo ontologias

- Construção de ontologias
  - Entrada: Conjunto de documentos
  - Saída: Ontologia com os documentos como ocorrências
- Extensão de ontologias
  - Entrada: Ontologia parcial e um conjunto de documentos como ocorrências
  - Saída: Ontologia com novos conceitos utilizando as ocorrências dadas
- População de ontologias
  - Entrada: Ontologia e um conjunto de ocorrências
  - Saída: para cada nodo da ontologia, listar o conjunto de documentos considerados como ocorrências do nodo
- Em todas as tarefas as ontologias são consideradas como taxonomias na forma de árvores!

Construção de ontologia geográfica a partir de textos da web

Mais complexa, pois envolve as seguintes diferenças:

- Ocorrências são os termos encontrados nos documentos e não os documentos
- Termos (e não documentos) são relacionados a conceitos
- Relacionamentos entre conceitos não são restritos a uma hierarquia
- A ontologia construída é um grafo e não uma árvore
- As ocorrências podem ser encontradas em textos pertencentes a inúmeros domínios de conhecimento, ao contrário de ontologias de domínio


TERMOS E CONCEITOS GEOGRÁFICOS
TURISMO
Sintra está localizada a menos de 1 hora de Lisboa.
EDUCAÇÃO
O LASIGE está situado no DI da FCUL em Campo Grande.
DIREITO
O Marco foi alvejado na Av. da República, próximo à Biblioteca Nacional.

Construção de ontologia geográfica a partir de textos da web

Exemplos:

- A vila de Ansião está próxima de Penela.
- Conceito (**Vila**)
- Relacionamento (**próximo**)
- Vila não é utilizado no vocabulário formal das fontes de informação
- **próxima(Ansião,Penela)**.
- Adjacentes, porém pertencentes a distritos distintos (Leiria e Coimbra, respectivamente)

Construção de ontologia geográfica a partir de textos da web

Exemplos:

- A aldeia à qual pertence esta escola, Bujões, fica situada acerca de 16 Km de Vila Real.
- Conceito (**Aldeia**)
- Ocorrências - **Bujões** e **Vila Real**
- Relacionamento indicando distância - **acerca de 16 Km**

Detectando termos geográficos em **todos** os documentos do WPT 03 em português

Objetivos:

- Utilizar uma abordagem escalável para encontrar ocorrências geográficas relacionadas
- Explorar o uso de padrões freqüentemente sucedidos por nomes geográficos
- Verificar a sobreposição dos nomes geográficos encontrados com aqueles na Geo-Net-PT01

Estratégia: Uso de padrões

P1: loc: localizad[ao]s? [a-z]+ [A-Z]

P2: sit: situad[ao]s? [a-z]+ [A-Z]

Tabela: Padrões com verbos

	# docs.	# sentenças	# Total de EMs	# EMs distintas	Sobreposição com a GN
P1: loc	817	918	877	522	128 (24,5%)
P2: sit	1.289	2.061	1.899	767	188 (24,5%)

Análise ...

- 75% das LOCs distintas estão fora da GN
- Nomes fora de Portugal (*Sul da Faixa de Gaza*, *Europa*, *Estado de São Paulo*, ...)
- Nomes informais - províncias e regiões (*Beira-Interior*, *Douro Litoral*, *Noroeste Trasmontano*, ...)
- Ocorrências de organizações frequentemente precedem o padrão. Exemplos:
  - O **Solar de Lavos** é um **restaurante** localizado em **Santa Luzia de Lavos** , a **8 kms da Figueira da Foz**.
  - ... a **Pousada de Vale de Gaio** , localizada entre **Alcácer do Sal e Torrão ou o Torrão e Alcácer do Sal** ...
  - A **Quinta do Carvalhal** fica localizada em **Celorico de Basto** , num vale ...

Outros padrões úteis

- norte [a-z]+ [A-Z]
- sul [a-z]+ [A-Z]
- [l]este [a-z]+ [A-Z]
- oeste [a-z]+ [A-Z]
- Raramente precedidas por ocorrências geográficas
- Geralmente sucedidas por ocorrências geográficas
- Granularidade dessas ocorrências? (trabalho futuro)

Considerações Finais

- GKB / Geo-Net-PT01
- Experimentos com sistemas de REM
- A construção de ontologias geográficas a partir de textos é "levemente" diferente da construção de ontologias de domínio

"O principal problema na construção de ontologias não é construir uma hierarquia, mas sim, assumir que dois termos existem e determinar **qual é a natureza da relação entre eles** [Brewster e Wilks04]".