Tirando o chapéu à Wikipédia: A coleção do Págico e o Cartola

Alberto Simões, Luís Costa e Cristina Mota

Coimbra
17 de Abril de 2011

Introdução

- O uso da Wikipédia (real) como coleção para o Págico é impensável:
  - inconstância nos conteúdos;
  - liberdade de edição;
- Imprescindível preparar uma coleção oficial para os participantes;
- Sintaxe MédiaWiki complexa e pouco comum, pelo que preferível disponibilizar coleção num formato mais comum.

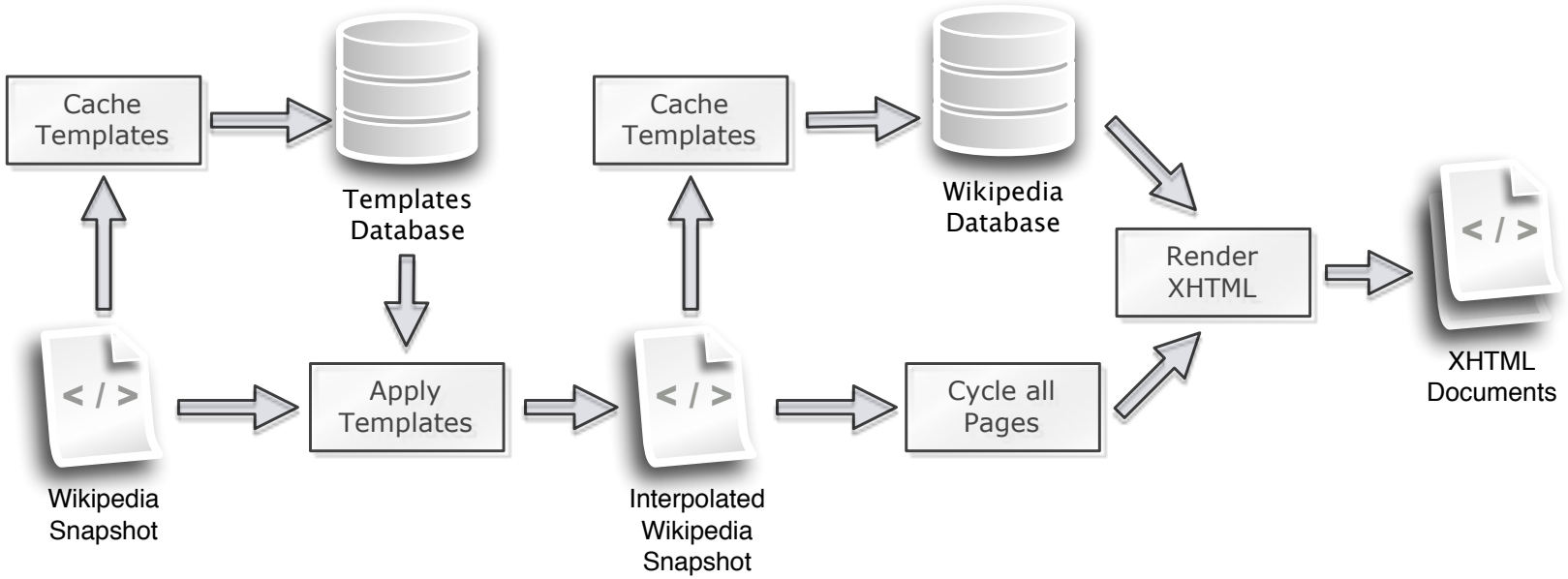
Construção da Coleção: Fonte

- Cópia Estática de 25 de Abril de 2011;
- Disponível em:
  `http://dumps.wikimedia.org/ptwiki/20110425/`
- Versão da Wikipédia num único documento XML;

Construção da Coleção: Ferramentas

- `MediaWiki::DumpFile` para processar, e percorrer os artigos constantes no documento XML;
- `mwlib` para a transformação em XHTML;
- Ferramentas *caseiras* Perl;

Construção da Coleção: Fluxo de Dados



Problemas

- As ferramentas disponíveis estão preparadas para a versão inglesa da Wikipédia;
- A versão portuguesa traduz alguns *namespaces*, o que faz com que essas ferramentas não saibam processar macros ou como gerar URL.
- O tratamento completo de macros é complicado e moroso, pelo que se optou por esquecer alguns macros complicados (nomeadamente, infoboxes).

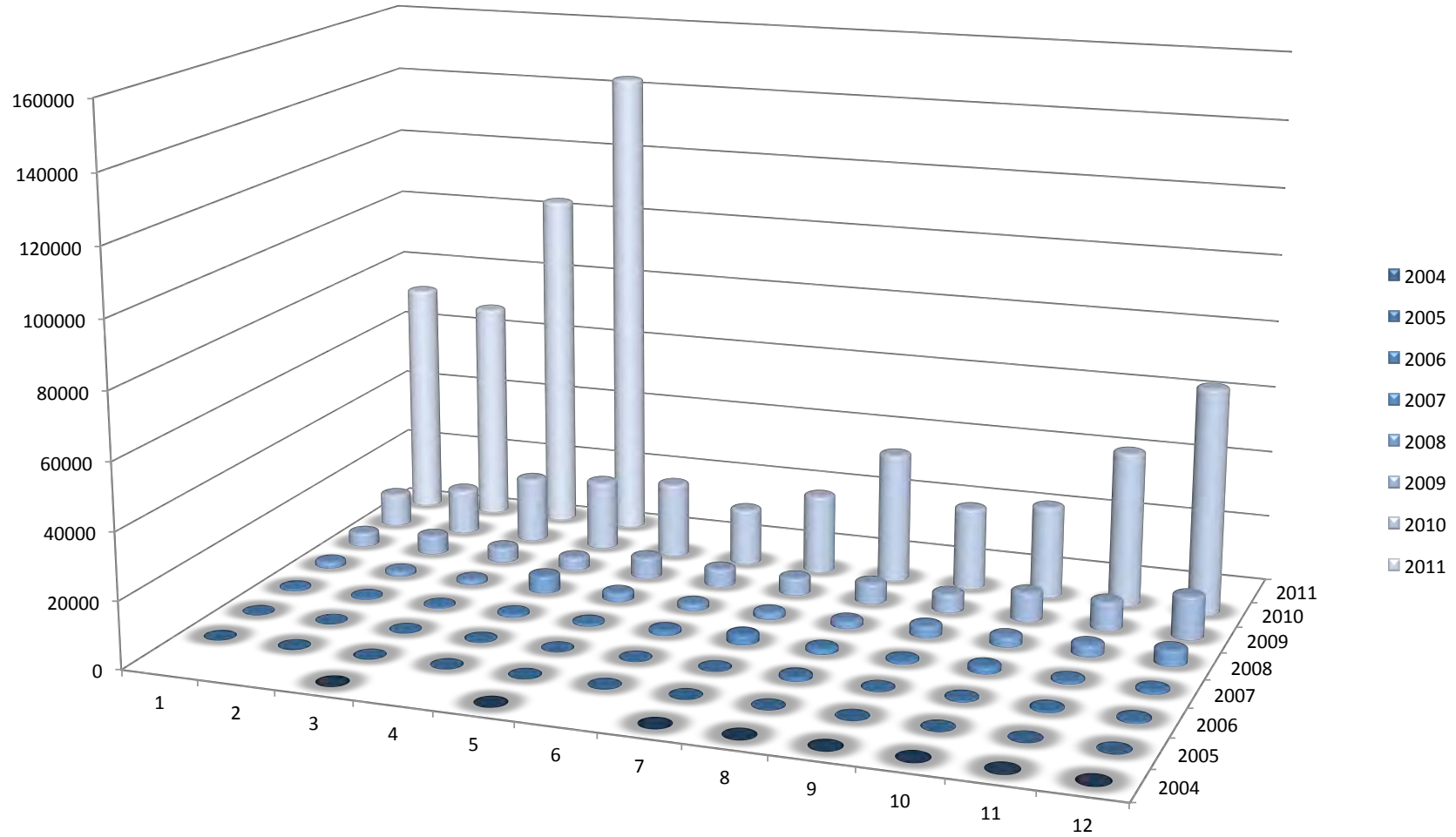
Distribuição por Tipo de Documento

Tipo	Nº de documentos
Páginas de pré-definição	32 900
Páginas de desambiguação	5 006
Páginas de redireção	574 077
Páginas de audiovisuais	9 678
Artigos (e anexos)	856 005

Tamanhos dos artigos

nº de formas	nº docs	percentual
$]0, 5]$	1	0.00%
$]5, 1042[$	541 628	78.54%
$]1042, 2075[$	87 789	12.73%
$]2075, 3108[$	26 527	3.85%
$]3108, 4141[$	11 931	1.73%
$]4141, 5176[$	6 501	0.94%
$]5176, 6232[$	3 946	0.57%
$]6232, 7378[$	2 711	0.39%
$]7378, 8707[$	1 989	0.29%
$]8707, 10256[$	1 691	0.25%
$]10256, 12439[$	1 447	0.21%
$]12439, 15585[$	1 256	0.18%
$]15585, 21968[$	1 139	0.17%
$]21968, \infty]$	1 063	0.15%

Constante atualização da Wikipédia



O Cartola

Pacote de *recursos públicos*
produzidos no decurso da organização do *Págico*.

Disponível em
`http://www.linguateca.pt/Cartola/`

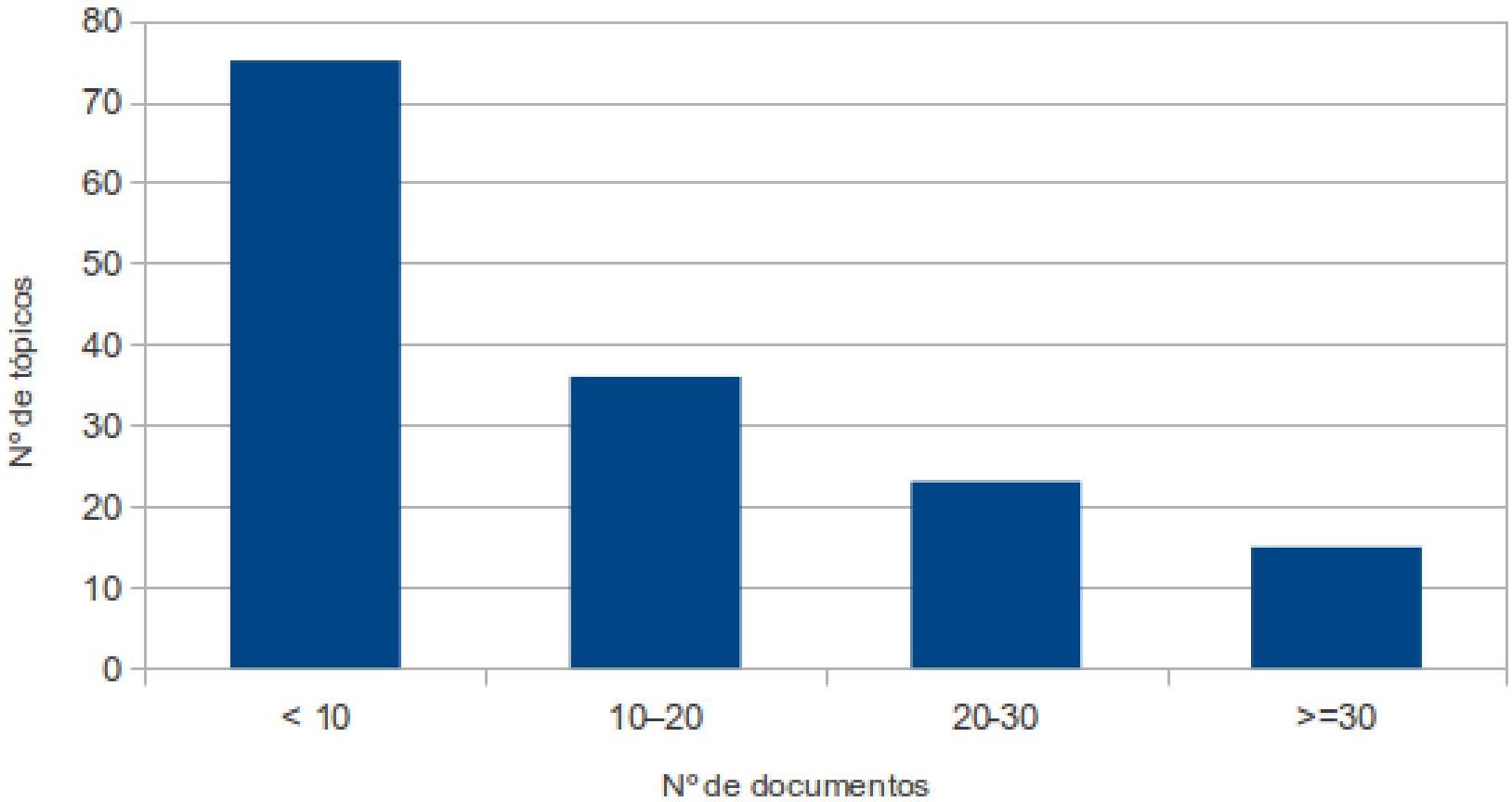
Cartola: conteúdo (I)

- coleção do Págico, de 681.058 documentos da wikipedia portuguesa de 25 de abril de 2011
- coleção de tópicos do Págico (xml, txt)
- monte das respostas avaliadas
- subcoleção do monte do Págico

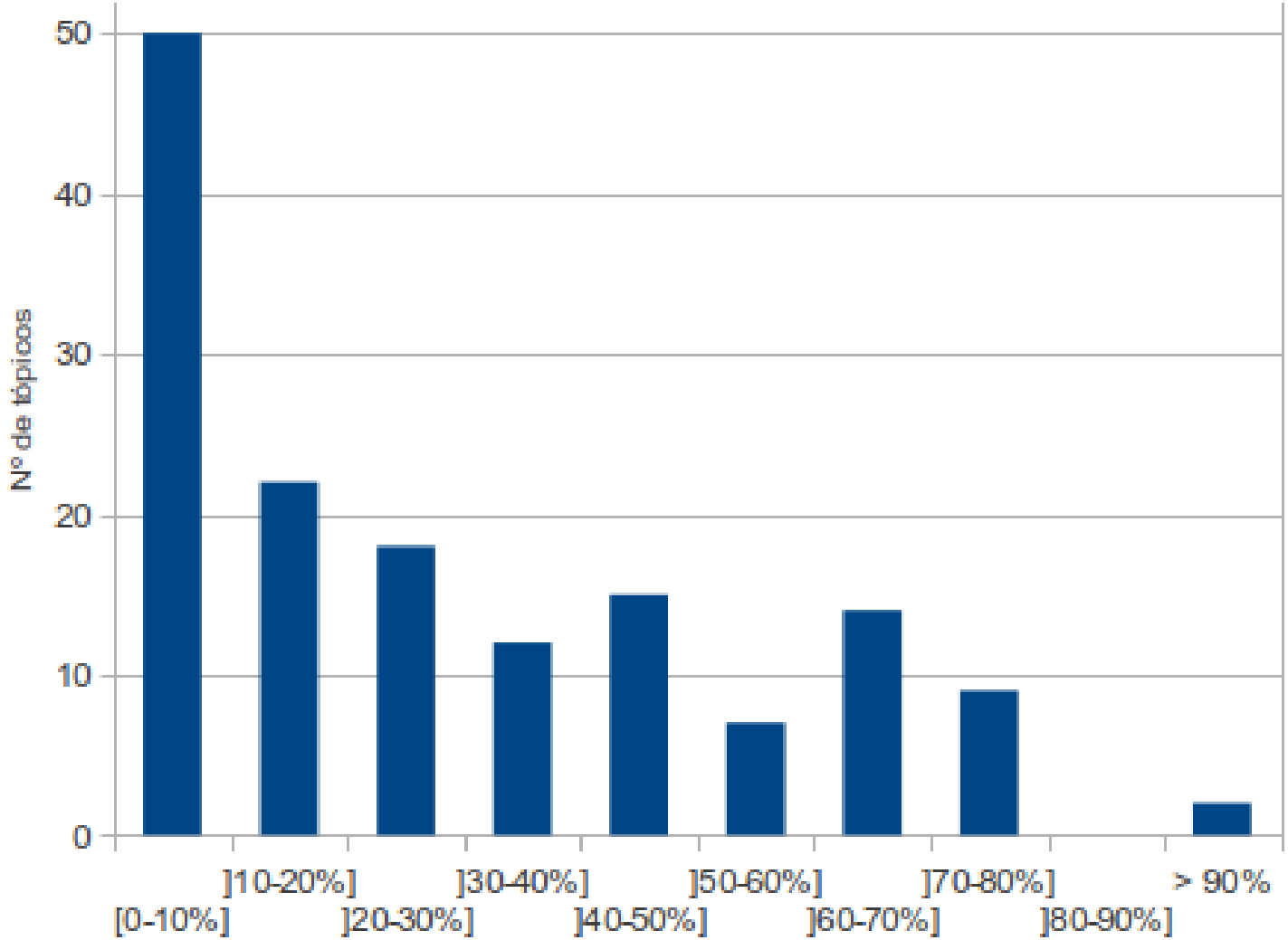
Cartola: conteúdo (II)

- lista de respostas corretas e justificadas, sem as respetivas justificações
- lista de respostas corretas e justificadas, com as respetivas justificações
- lista de respostas consideradas corretas independentemente de estarem bem justificadas

Documentos de resposta corretos por tópico



% documentos resposta corretos apenas na Wikipédia PT



Tópicos vs número de documentos de resposta corretos

ID	Tópico	# Docs
19	Tribos indígenas que vivem na Amazônia.	95
147	Museus em capitais de países lusófonos	62
144	Locais referidos n' "Os Lusíadas"	51
79	Povos indígenas brasileiros considerados extintos.	50
106	Vice-reis da Índia Portuguesa	48
	(...)	
110	Políticos da África lusófona que estudaram na União Soviética	2
54	Igrejas do Rio de Janeiro construídas por irmandades ou confrarias de negros.	1
132	Deputados da FRELIMO	1
116	Escritores moçambicanos que receberam o Prémio Camões	1
55	Escritores estrangeiros que visitaram Portugal no século XIX e que publicaram descrições das suas viagens	1

Conclusões

- Uma nova avaliação conjunta merece nova coleção, não só pela atualidade do conteúdo, como também pela possibilidade de se conseguir um melhor compromisso na geração dos documentos XHTML.
- Com a disponibilização do recurso Cartola (`http://www.linguateca.pt/Cartola/`) pretendemos que o trabalho e a experiência no Págico possa ser o mais proveitosa possível para quem estiver interessado nas áreas abordadas pelo Págico.