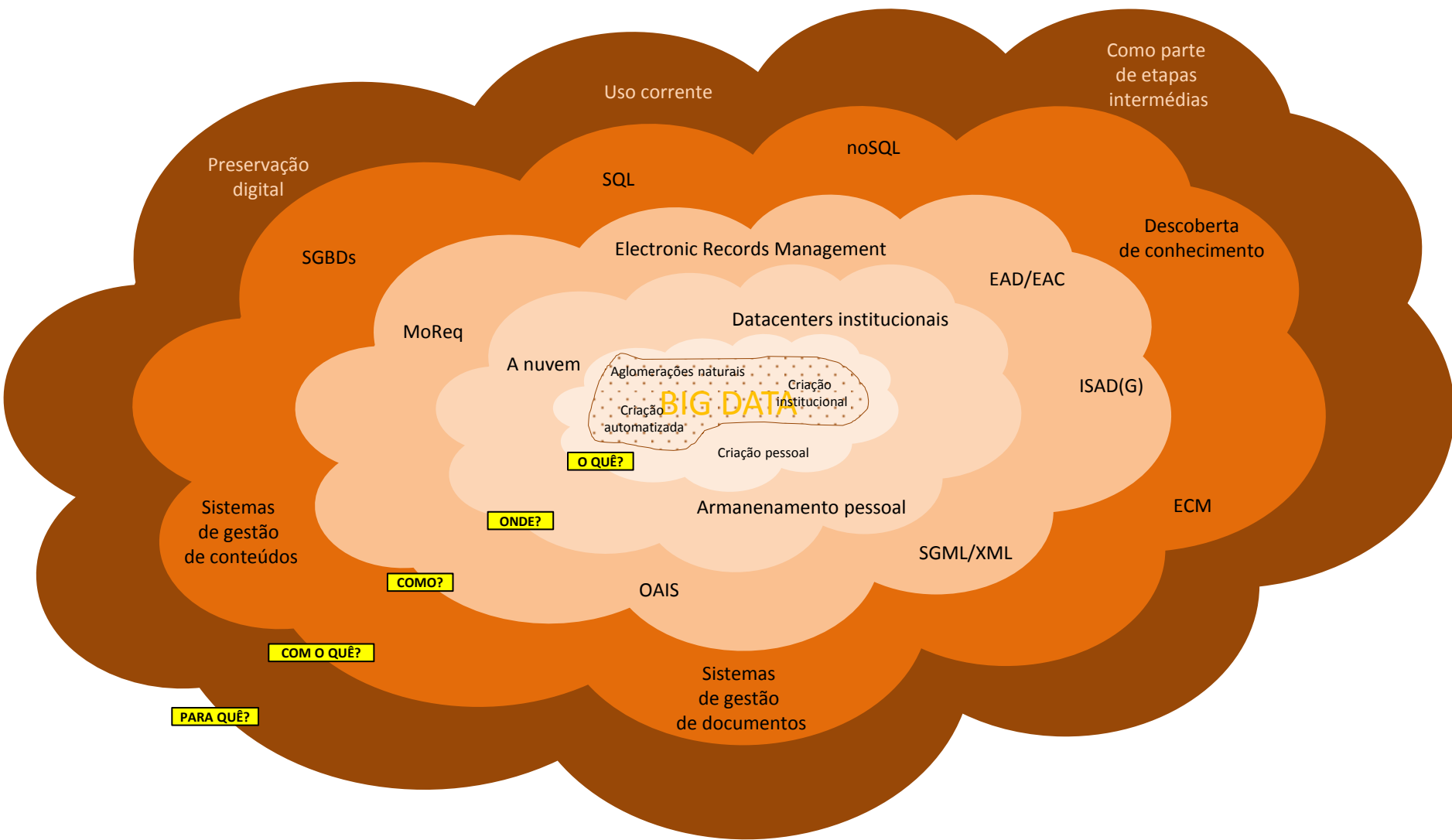


ARQUIVOS ELECTRÓNICOS

João Paulo
Amado
21ª aula
2011-12-17

ARQUIVOS ELECTRÓNICOS: Uma proposta de articulação de componentes

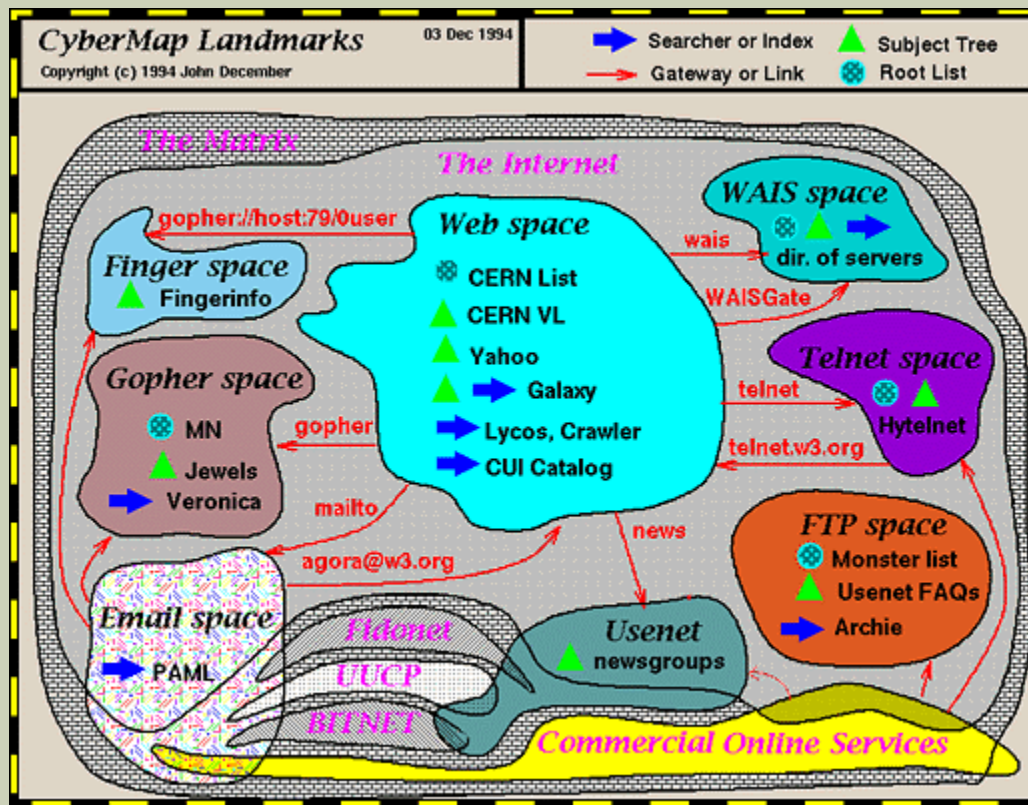


A MAIOR BASE DE DADOS DE TODAS (1)

- A maior base de dados existente ainda não é, verdadeiramente, uma base de dados;
- Estamos a falar de um aglomerado de informação que se afasta deste conceito em particular;
- Referimo-nos à **WWW**.

A MAIOR BASE DE DADOS DE TODAS (2)

- NB: A WWW não é, nem nunca foi, a única vertente da Internet.



John December, CyberMap Landmarks, trabalho de 1994

A MAIOR BASE DE DADOS DE TODAS (3)

- Mas se não é verdadeiramente uma base de dados, então o que é?
 - Falemos antes de uma vasta aglomeração de conteúdos, que se diria estruturado em texto livre;
 - Contém inúmeros blocos discretos de informação, que se dividem entre os não-descritos, os mal-descritos e os indescritíveis.

A MAIOR BASE DE DADOS DE TODAS (4)

- Não esqueçamos, igualmente, que não há qualquer obrigatoriedade quanto à correcta descrição e identificação dos conteúdos.
 - Um web site apenas tem que «mostrar» coisas e valer por si só, sem ter que, forçosamente, se importar com o resto da Internet.

A MAIOR BASE DE DADOS DE TODAS (5)

- Não esqueçamos ainda que o *software* que apresenta estes conteúdos (maioritariamente *web browsers*), é bastante tolerante a falhas e consegue lidar, de forma robusta, com situações de falta de informação.

A MAIOR BASE DE DADOS DE TODAS (6)

- Qualquer uma das regras subjacentes às estruturas de bases de dados não pode ser aplicada a este universo informativo.
 - Por exemplo, um conjunto de páginas web desenvolve, muito naturalmente, falhas de integridade entre si com o passar do tempo (porque os links se alteram, porque as páginas mudam de nome e desaparecem, etc.);
 - A WWW é efémera – em 2004 estimava-se que cerca de 20% da totalidade das páginas web existentes não estariam disponíveis daí a um ano. Claro que entretanto surgiram muitas mais.

A MAIOR BASE DE DADOS DE TODAS (7)

- Então, porque é que se diz que é uma base de dados?

A MAIOR BASE DE DADOS DE TODAS (8)

- Os principais responsáveis são os grandes motores de pesquisa da Internet:
 - Habitúamo-nos a olhar para a Internet como um conjunto de coisas que podem ser pesquisáveis;
 - A ideia de «coisa pesquisável» está mais relacionada com a existência de bases de dados que com outra coisa;

A MAIOR BASE DE DADOS DE TODAS (9)

■ O que existe hoje:

- A unidade básica de registo é a página *web*;
- Uma página *web* não é equiparável a um registo dentro de uma tabela de uma base de dados; também não podemos pensar num web site como uma base de dados e cada página *web* que o compõe como uma tabela;
- Os conceitos do mundo das bases de dados não se plasmam directamente sobre esta realidade;
- As bases de dados verdadeiras estão imbricadas no tecido da WWW, fazendo parte das fontes do seu conteúdo.

A MAIOR BASE DE DADOS DE TODAS (10)

- Uma página *web* vale pelo seu conteúdo:
 - É ele que é pesquisável, porque indexado;
 - É esse o conteúdo que verdadeiramente interessa aos utilizadores.
- No entanto, podem haver outros níveis de conteúdo presentes numa página *web*:
 - Directivas de meta-informação no cabeçalho do código que compõe a página e que servem para fornecer dados de identificação;
 - Outras indicações meta-informativas, associadas a blocos de conteúdo da página (imagens, por exemplo);
 - Estes meta-dados podem ser aproveitados para dar mais riqueza a uma indexação por conteúdos.

A MAIOR BASE DE DADOS DE TODAS (11)

- Há quem tente fazer «cópias de segurança» desta base de dados:
 - Projectos de âmbito universal para preservar conteúdos da Internet, o mais famoso e eficaz dos quais está consubstanciado no *web site* archive.org;
 - Uma quantidade de projectos geograficamente mais restritos, normalmente à escala de um país, tendo por objectivo preservar o conjunto de páginas web desse país; em Novembro de 2011 eram 52 um pouco por todo o mundo;
 - O projecto português está disponível a partir de www.arquivo.pt;
 - Não há verdadeiramente uma abordagem comum e uniforme.

A MAIOR BASE DE DADOS DE TODAS (12)

- **O que se pretende que venha a existir:**
 - A web semântica, a web das coisas, a web 3.0.
 - Os nomes são muitos, o objectivo é um só: ter dados online, bem descritos e interligados, de forma a que a partir de algo que se encontre, seja fácil encontrar informação relacionada.
 - Por si só, este objectivo não difere muito do modelo teórico que está subjacente à existência da WWW. O que muda é a abordagem e os conceitos.

A MAIOR BASE DE DADOS DE TODAS (13)

■ Componentes desta abordagem:

- O conjunto de especificações RDF (Resource Description Framework), definido pelo World Wide Web Consortium e que se destina a suportar a descrição conceptual da informação disponibilizada através da WWW.
- A descrição RDF de um artigo da Wikipedia pode assumir esta forma:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://pt.wikipedia.org/wiki/Lisboa">
    <dc:title>Lisboa</dc:title>
    <dc:publisher>Wikipedia</dc:publisher>
  </rdf:Description>
</rdf:RDF>
```

A MAIOR BASE DE DADOS DE TODAS (14)

■ Componentes desta abordagem:

- O conjunto de especificações RDF (Resource Description Framework), definido pelo World Wide Web Consortium e que se destina a suportar a descrição conceptual da informação disponibilizada através da WWW.
- A descrição RDF de um artigo da Wikipedia pode assumir esta forma:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://pt.wikipedia.org/wiki/Lisboa">
    <dc:title>Lisboa</dc:title>
    <dc:publisher>Wikipedia</dc:publisher>
  </rdf:Description>
</rdf:RDF>
```