

Aplicações na Web

Relatório de Projecto (Etapa 3)



Grupo 006

MBBC

sexta-feira, 3 de Junho de 2016

Alunos:

- Ana Filipa Vale, nº 18120
- Isabela Mott, nº 48391
- Pedro Barros, nº 48385
- Pedro David, nº 40403

Índice

1. Introdução.....	3
2. Planeamento	4
3. Arquitetura.....	6
4. Back-end.....	7
1. Tecnologias usadas	7
2. Fontes de Informação usadas.....	7
3. Estrutura de Dados	7
4. Componentes Implementados	8
5. Web Service.....	9
1. Tecnologias usadas	9
2. Lista de Métodos	9
3. Exemplos	10
6. Front-end (admin e user)	12
1. Tecnologias usadas	12
2. Web Services usados.....	12
3. Screenshots	12
7. Discussão.....	19
1. Problema 1 encontrado	19
2. Problema 2 encontrado	19
3. Problema 3 encontrado	19
4. Problema 4 encontrado	19
5. Problema 5 encontrado	20
Crítica Global	20
8. Anexos.....	20

1. Introdução

O presente relatório acompanha o desenvolvimento de ResearchBond, uma aplicação Web que permite criar uma visualização inovadora do percurso científica de um investigador, através das suas publicações e colaborações. A aplicação utiliza fontes de informação disponíveis na Web, tais como o Google Scholar, DBpedia e CrossRef, de onde são extraídas listas de instituições, publicações, citações e co-autores, para criar um perfil de cada investigador. Este perfil é representado em formato de gráfico de rede (*network*) onde estão representados o investigador e os co-autores que publicaram em conjunto com este. Adicionalmente, informações como o *h*-index do investigador e a percentagem de artigos do deste que também são comuns a cada co-autor serão também representadas. Esta visualização é interativa, sendo possível identificar o impacto de um dado co-autor no perfil do investigador considerando a proporção de artigos publicados em conjunto.

O input que lança a aplicação é o nome do investigador ou de uma instituição de investigação/universidade.

O output gerado é uma rede de representação das ligações entre o autor principal e a todos os seus co-autores. O autor principal (que é pesquisado) é representado por um círculo central, cuja área será diretamente proporcional ao impacto do investigador (*h*-index). As ligações entre autores têm um tamanho gerado aleatoriamente. A espessura da ligação é proporcional à percentagem de publicações feitas em conjunto e o tamanho dos nós que representam os co-autores é proporcional ao número de citações total que as publicações conjuntas receberam.

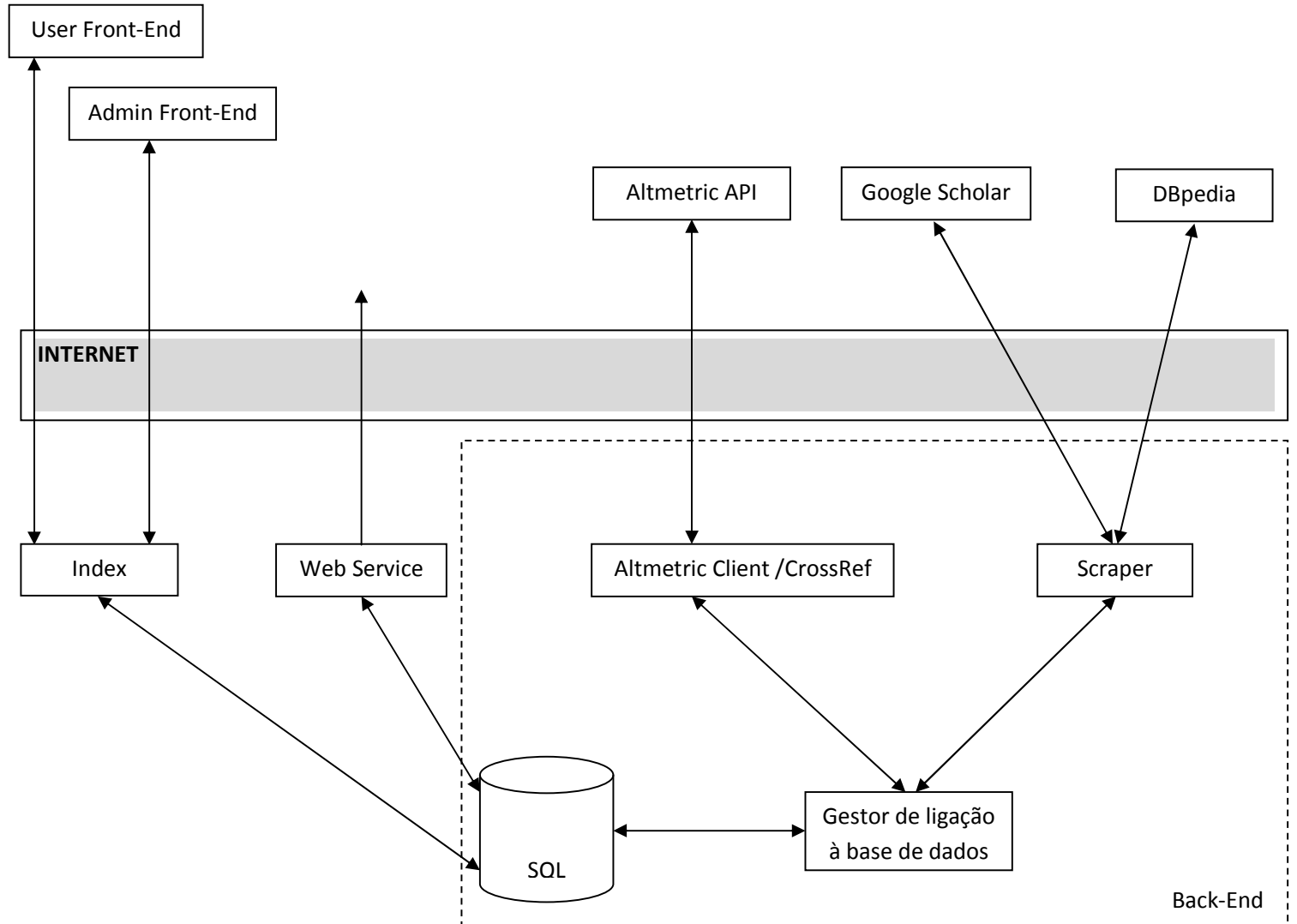
Um trailer representativo do site pode ser visto aqui: <https://youtu.be/maTZHYilAxs>

2. Planeamento

Período	Tarefa	Responsável	Observações
15/02 – 06/03	Modelação do Projecto: Definição da arquitectura; Ideias para a interface e webservices	Todos	
07/03 – 03/04	Exploração da API do Altmetric; Desenvolvimento de código relevante para o projecto que faça uso da API Utilização da API CrossRef.	Filipa Vale	
	Desenvolvimento da interface provisória, somente para visualização dos dados já existentes na base de dados	Isabela Mott, Mariana Nave	
	Exploração do Google Scholar; Desenvolvimento do Google Scholar Scraper	Pedro Barros	
	Desenvolvimento do WebService;	Pedro David	
	Desenvolvimento de uma base de dados relacional temporária em SQL	Pedro David, Mariana Nave	
04/04 – 01/05	Desenvolvimento de uma interface de ligação à base de dados e respectiva documentação; Redefinição do WebService e desenvolvimento da sua documentação online	Pedro David	
	Desenvolvimento de um novo modelo ER para armazenamento de dados em SQL; Ajuste dos Scrapers para o Google Scholar	Pedro Barros	
	Desenvolvimento de script para utilizar API CrossRef para armazenar dados em SQL	Filipa Vale	
	Ajuste dos Scrapers para o Google Scholar	Pedro Barros, Isabela Mott	
	Desenvolvimento do Front-end de administrador: login, opções para adicionar, editar ou eliminar dados da Base de Dados. Ajuste dos scrips anteriores.	Isabela Mott	

Período	Tarefa	Responsável	Observações
15/02 – 06/03	Conclusão geral do projecto; Troca de ideias finais e resolução de problemas em conjunto	Todos	
	Criação do logotipo e scraping de fotos dos autores; Função em PHP para cálculo de percentagens de publicações conjuntas e número de citações	Pedro Barros	
	Query SPARQL à DBpedia para inserção de novas instituições na base de dados	Pedro Barros, Pedro David	
	Desenvolvimento da rede de interacção Autor/Co-autor (D3); Definição da semântica (RDFa)	Pedro David	
	Ajuste ao código da página de administrador do Front-end, para entrar apenas com login e password; Função em PHP para comparação das strings Título da publicação encontradas entre Google Scholar e CrossRef e inserção na base de dados da percentagem de semelhança	Filipa Vale	
	Conclusão do desenvolvimento de todas as páginas do Front-end. Criação de página novas de Front-end (adminOptions, authorInfor, fullInfo, gethintAuth, gethintInst, removeAuthor, removeInstitution, searchRemove), utilizando HTML, CSS, JavaScript e PHP; Queries SQL à base de dados para apresentar nas páginas do Front-end	Isabela Mott	

3. Arquitetura



4. Back-end

1. Tecnologias usadas

O conjunto de tecnologias a serem utilizadas na construção do back-end desta Aplicação Web irá incluir:

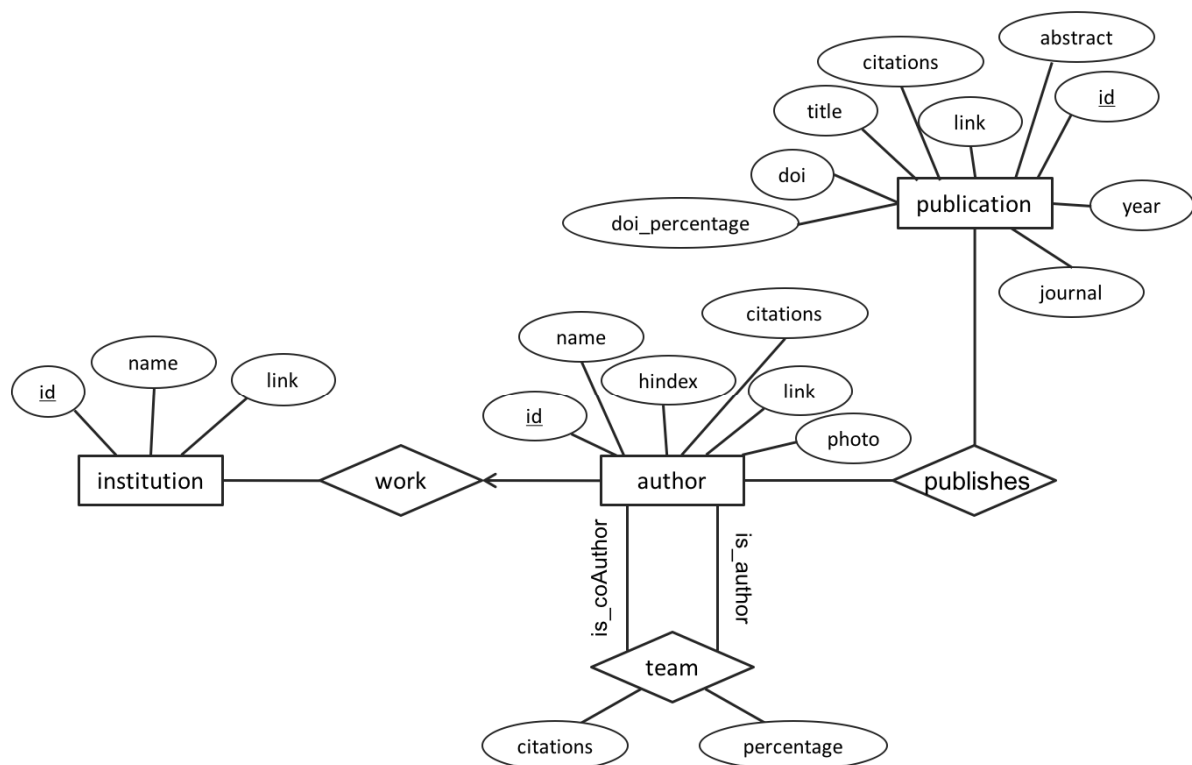
- PHP
- MySQL

2. Fontes de Informação usadas

- Altmetric: Componente social (citações na Wikipedia, twitter, blogs, etc.)
- Google Scholar: Nome dos autores, co-autores, artigos publicados, abstracts, Instituições, citações, h-index
- DBpedia: Instituições
- CrossRef: identificação única do artigo (Digital Object Identifier, DOI) e, por vezes, nome do jornal onde trabalho está publicado

3. Estrutura de Dados

O modelo ER desenvolvido para a Base de Dados SQL está representado abaixo:



4. Componentes Implementados

O input da aplicação é o nome do investigador ou de uma instituição de investigação/universidade. O output é uma rede de representação das ligações entre o autor principal e a todos os seus co-autores. Todos os autores são representados por círculos, cuja área será directamente proporcional ao impacto do investigador (h-index) ou o número de citações das publicações conjuntas no caso dos co-autores. As ligações entre autores têm um tamanho definido aleatoriamente. A espessura das ligações é proporcional ao número de artigos publicados em conjunto. No output é mostrado também uma representação da API do Altmetric para indicar referências nas redes sociais de cada publicação.

A base de dados (BD) SQL para armazenamento de dados para a aplicação encontra-se agora na sua versão final. Um conjunto de funções de suporte à gestão (insert, delete, update, etc) da informação na BD foi também desenvolvido (bd_functions.php)

O webservice também já se encontra completo, estando a sua documentação situada no servidor, aqui: <http://appserver.di.fc.ul.pt/~aw006/Webservice/documentation/>

A página de cada autor possui informação semântica, definida em RDFa.

A API Altmetric permite recolher citações dos trabalhos científicos que são feitas de forma menos convencional, recorrendo por exemplo a citações em agências de notícias de ciências, social media, entre outros. A API Altmetric permite o acesso programático a dados sobre artigos e conjuntos de dados reunidos por Altmetric. O input da API Altmetric poderá ser: DOI (digital object identifier), arXiv ID, Handle, PubMed ID, URI ou Altmetric ID. Porém, nenhum destes inputs é dado pelo Google Scholar, razão pela qual é utilizada a API CrossRef para obter DOI a partir do título e autores. A API CrossRef pode ainda ser usada para obter o nome da revista onde foi publicado o trabalho. Assim foi desenvolvido um script em PHP para recolher estas informações da API CrossRef e armazenar na base de dados, de acordo com as especificações do WebService desenvolvido.

Dois *scripts* foram desenvolvidos em PHP para *Web scraping* para recolher informações a partir de uma página do Google Scholar. Estes *scrapers* fazem uso da ferramenta cURL, para transferir os dados HTML de uma página de perfil do Google Scholar referentes a uma determinada Instituição, Autor ou Publicação. Estes dados são depois transferidos para um Document Object Model (DOM) e processados de acordo com as especificações do WebService desenvolvido. Ambos os *sripts* utilizam agora funções existentes em bd_functions.php para inserir e actualizar informação na BD. O *scraper* para a Instituição recebe o nome de uma Instituição como *input* e adiciona na base de dados o nome da Instituição e dos autores (Investigadores) registados naquela instituição. Adiciona também as informações relativas aos autores, nomeadamente o nome, número total de citações, h-index, Instituição, títulos de publicações (e correspondente número de citações) e a os respectivos co-autores. As ligações entre Instituição-Autor, Autor-CoAutor e Autor-Publicações são garantidas através de tabelas de relação *work*, *team* e *publishes*, que utilizam as chaves primárias de cada elemento.

O *scraper* para Autor receberá como *input* o nome e instituição de um Autor e adiciona na base de dados informações do autor incluindo: nome, número total de citações, h-index, Instituição, títulos de publicações (e correspondente número de citações) e a lista co-autores. Este script servirá para complementar o scraper para Instituição, permitindo incluir autores que não se encontram associados directamente a um perfil de uma Instituição no Scholar.

Ambos os scrapers utilizam a função de chamada da API do CrossRef para adicionar o código DOI e a Revista e completar informação em falta relativamente à Revista. Tendo em conta que muitas publicações podem não ser encontradas pela API (por exemplo, devido a problemas de formatação da string contendo o título da publicação), foi definida a utilização da referência sempre que o título devolvido pela API fosse semelhante ao título original em pelo menos 80% da sua composição. A percentagem de semelhança é adicionada à tabela *publications* na coluna *doi_percentage*.

Ambos os *scrapers* na sua função default adicionam apenas o Título, número de citações, ano de publicação e link uma vez que incluir a informação adicional (nomeadamente Abstract e Revista) sobrecarrega os *scrapers* e aumenta a probabilidade de bloqueio por parte do Google Scholar. Assim foi criada a opção no *scrapers* para um autor de se fazer esta pesquisa mais aprofundada. A activação dos *scrapers* é feita na interface Admin).

Foi adicionada uma função aos scrapers que também guardam a foto de cada autor no Scholar. Esta é guardada no servidos, e a path fica guardada como string na coluna *photo* da tabela *authors*.

5. Web Service

1. Tecnologias usadas

- REST
- XML
- JSON

2. Lista de Métodos

Recurso	Método	Parâmetro	Resultado
/author	GET		Lista Autores
/author/{A_ID}	GET	Id do Autor	Devolve um Autor
/author/{A_ID}/name	GET	Id do Autor	Devolve nome do Autor
/author/{A_ID}/citations	GET	Id do Autor	Devolve o número de citações do Autor
/author/{A_ID}/h-index	GET	Id do Autor	Devolve o h-index do Autor
/author/{A_ID}/link	GET	Id do Autor	Devolve o url do Google Scholar do Autor
/author/{A_ID}/institution	GET	Id do Autor Id da Instituição	Lista as Instituições do Autor
/author/{A_ID}/institution/{I_ID}	GET	Id do Autor Id da Instituição	Devolve uma Instituição do Autor
/author/{A_ID}/publication	GET	Id do Autor Id da Publicação	Lista as Publicações do Autor

/author/{A_ID}/publication/{P_ID}	GET	Id do Autor Id da Publicação	Devolve uma Publicação do Autor
/author/{A_ID}/co_author	GET	Id do Autor Id do Co-Autor	Lista Co-Autores do Autor
/author/{A_ID}/co_author/{A_ID}	GET	Id do Autor Id do Co-Autor	Devolve um Co-Autor do Autor
/institution	GET		Lista Instituições
/institution/{I_ID}	GET	Id da Instituição	Devolve uma Instituição
/institution/{I_ID}/name	GET	Id da Instituição	Devolve nome da Instituição
/institution/{I_ID}/link	GET	Id da Instituição	Devolve o url do Google Scholar da Instituição
/institution/{I_ID}/investigator	GET	Id da Instituição	Lista Investigadores da Instituição
/institution/{I_ID}/investigator/{A_ID}	GET	Id da Instituição Id do Investigador	Devolve um Investigador da Instituição
/publication	GET		Lista Publicações
/publication/{P_ID}	GET	Id da Publicação	Devolve um Publicação
/publication/{P_ID}/title	GET	Id da Publicação	Devolve título da Publicação
/publication/{P_ID}/abstract	GET	Id da Publicação	Devolve abstract da Publicação
/publication/{P_ID}/citations	GET	Id da Publicação	Devolve o número de Citações da Publicação
/publication/{P_ID}/journal	GET	Id da Publicação	Devolve revista da Publicação
/publication/{P_ID}/year	GET	Id da Publicação	Devolve o ano da Publicação
/publication/{P_ID}/doi	GET	Id da Publicação	Devolve o doi da Publicação
/publication/{P_ID}/link	GET	Id da Publicação	Devolve o url do Google Scholar da Publicação
/publication/{P_ID}/author	GET	Id da Publicação	Lista Autores da Publicação
/publication/{P_ID}/author/{A_ID}	GET	Id da Publicação Id do Autor	Devolve um Autor da Publicação

O output de cada método do webservice pode ser obtido no formato XML ou JSON.

3. Exemplos

Alguns exemplos de outputs XML gerados pelo WebService são apresentados em baixo:



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0"?>
<authors>
  <author>
    <id>2</id>
    <name>Wolfgang Brandner</name>
    <citations/>
    <hindex/>
    <link>
      <url>https://scholar.google.pt/citations?user=wPxAfH8AAAAJ&hl=pt-PT&oe=ASCII</url>
    </link>
    <publications/>
    <institutions/>
    <co-authors/>
  </author>
</authors>
```



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0"?>
<authors>
  <author>
    <name>Antonio Amorim</name>
  </author>
</authors>
```



This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
<?xml version="1.0"?>
<authors>
  <author>
    <id>2</id>
    <name>Wolfgang Brandner</name>
    <citations/>
    <hindex/>
    <link>
      <url>https://scholar.google.pt/citations?user=wPxAfH8AAAAJ&hl=pt-PT&oe=ASCII</url>
    </link>
  </author>
</authors>
```



6. Front-end (admin e user)

1. Tecnologias usadas

O conjunto de tecnologias a serem utilizadas na construção do front-end desta Aplicação Web irá incluir:

- PHP
- JavaScript
- D3
- AJAX
- CSS (com recurso à Bootstrap framework)

2. Web Services usados

É utilizado o Web Service desenvolvido no âmbito do projecto e o webservice da Altmetric.

3. Screenshots

São apresentados seguidamente alguns screenshots da aplicação ResearchBond.

A página inicial (Figura 1) apresenta um layout simples e agradável onde se encontram duas caixas de texto onde que recebem o input do utilizador. O botão *Search!* inicia a aplicação. Ambas as caixas têm implementadas funções de sugestões automáticas que permitem um acesso mais rápido a um autor ou instituição específico/a.

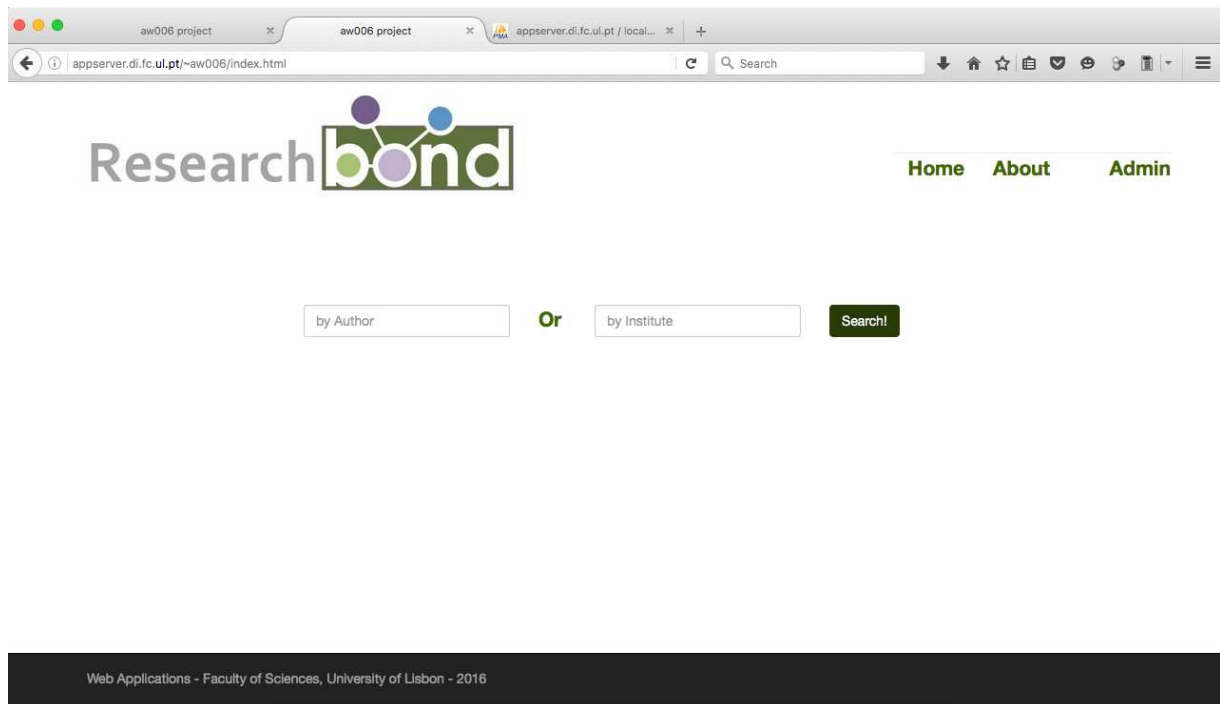


Figura 1 - Página inicial

Na parte superior da página inicial existem três campos *Home*, *About* e *Admin* que permitem aceder à página inicial, a informação sobre os detalhes da aplicação e ao painel do administrador, respetivamente. O acesso ao Home pode ser feito também clicando no logotipo no canto superior esquerdo.

A página *About* (Figura 2) apresenta informação geral relativa à funcionalidade da aplicação e aos autores. É também descrita a representação criada no output da aplicação.

A página *Admin* (Figura 3) permite efetuar o login pelo utilizador e aceder às funcionalidades de update da base de dados.

Após o login, o administrador é direcionado para uma página (Figuras 4, 5) onde são apresentadas diversas funções que permitem activar os scrapers ou remover informação. Deste modo, novas instituições e autores podem ser adicionados ou atualizados na Base de dados. O botão Add/Remove (Figura 4) permite aceder às funções que fazem o scraper do Google Scholar por instituição ou por um autor específico. O botão Load Data Base permite actualizar toda a Base de dados. Este botão actualiza um scraper que inicia com uma pesquisa SPARQL à DBpedia, para adicionar todas as instituições portuguesas anotadas. Seguidamente, adiciona informações relativas a todos os autores registados nessas instituições. O botão Remove (Figura 4) permite procurar um autor ou instituição por nome a remover toda a informação existente na base de dados.

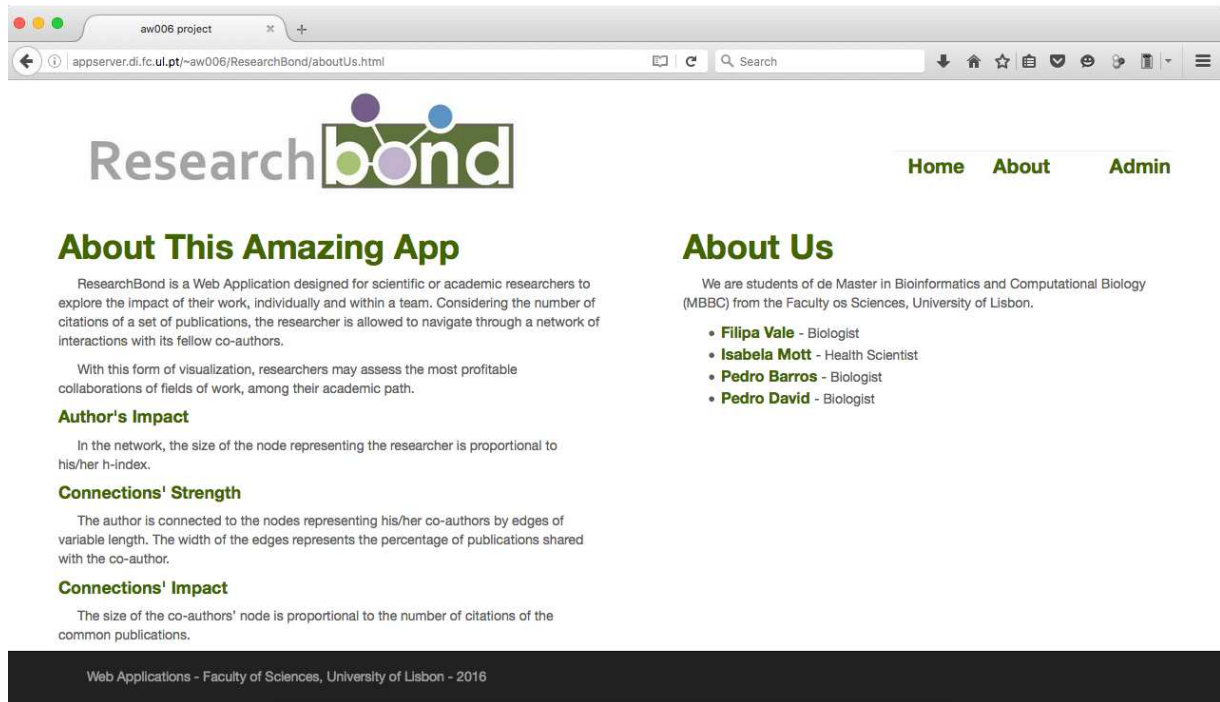


Figura 2 - Página About

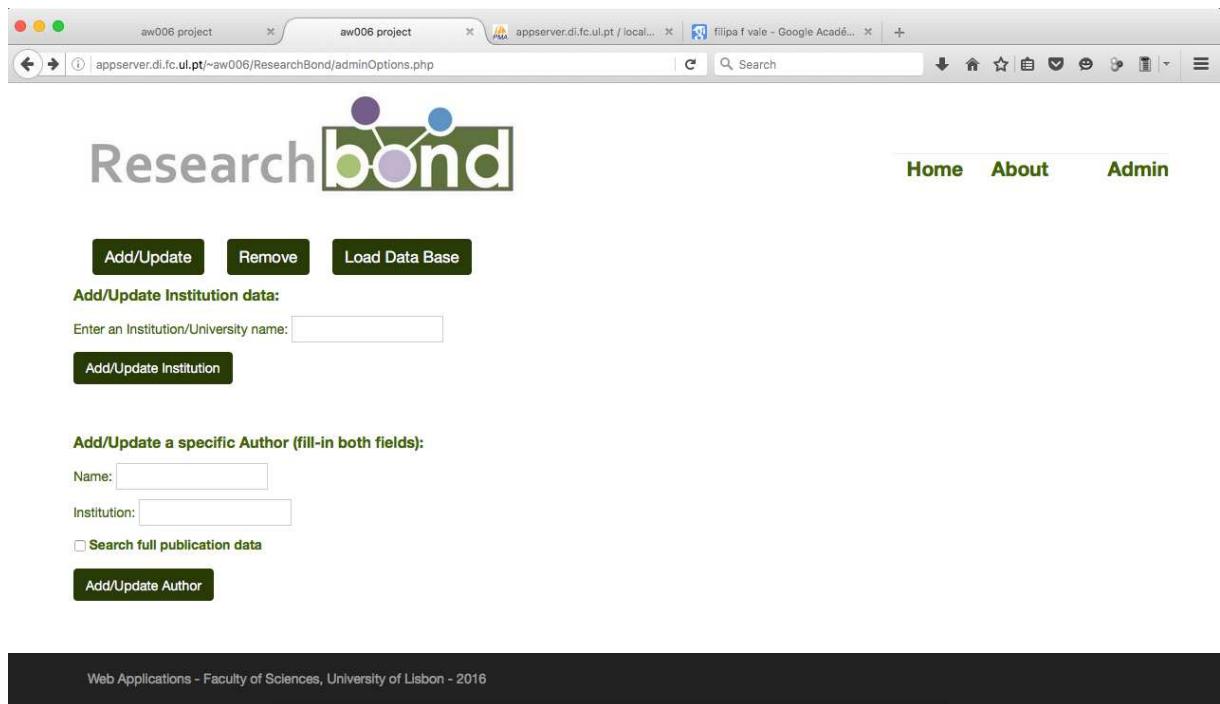


Figura 3 - Página Admin

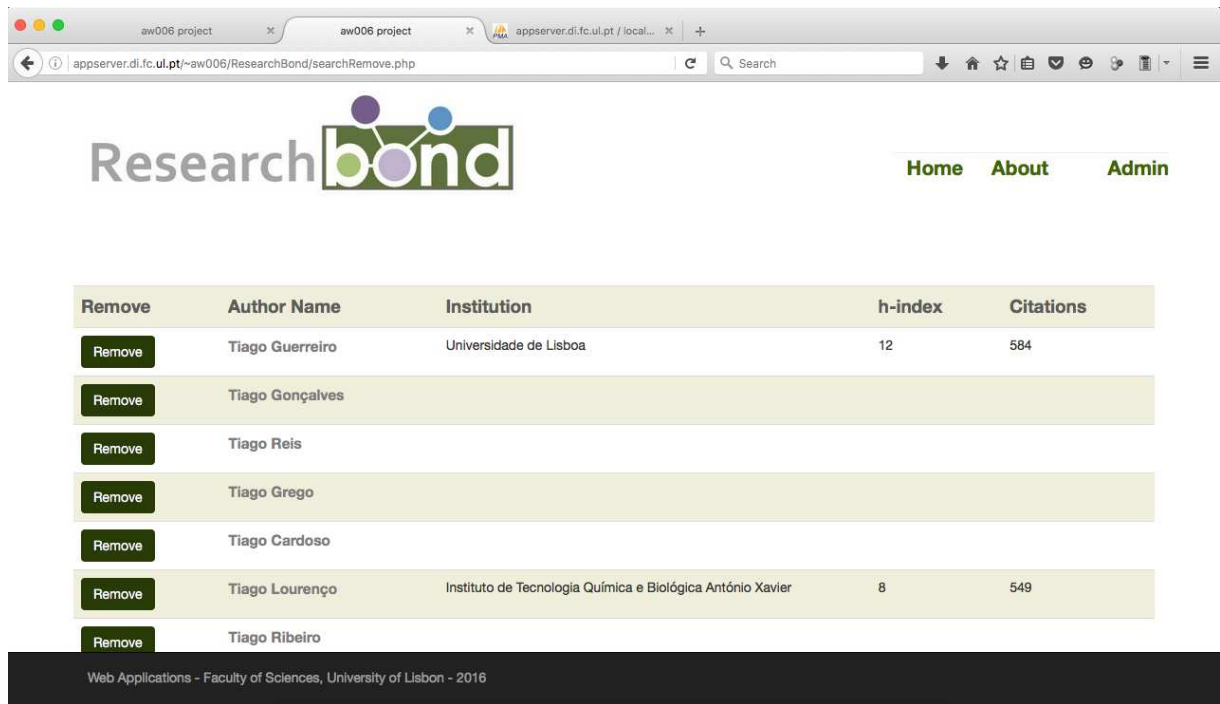
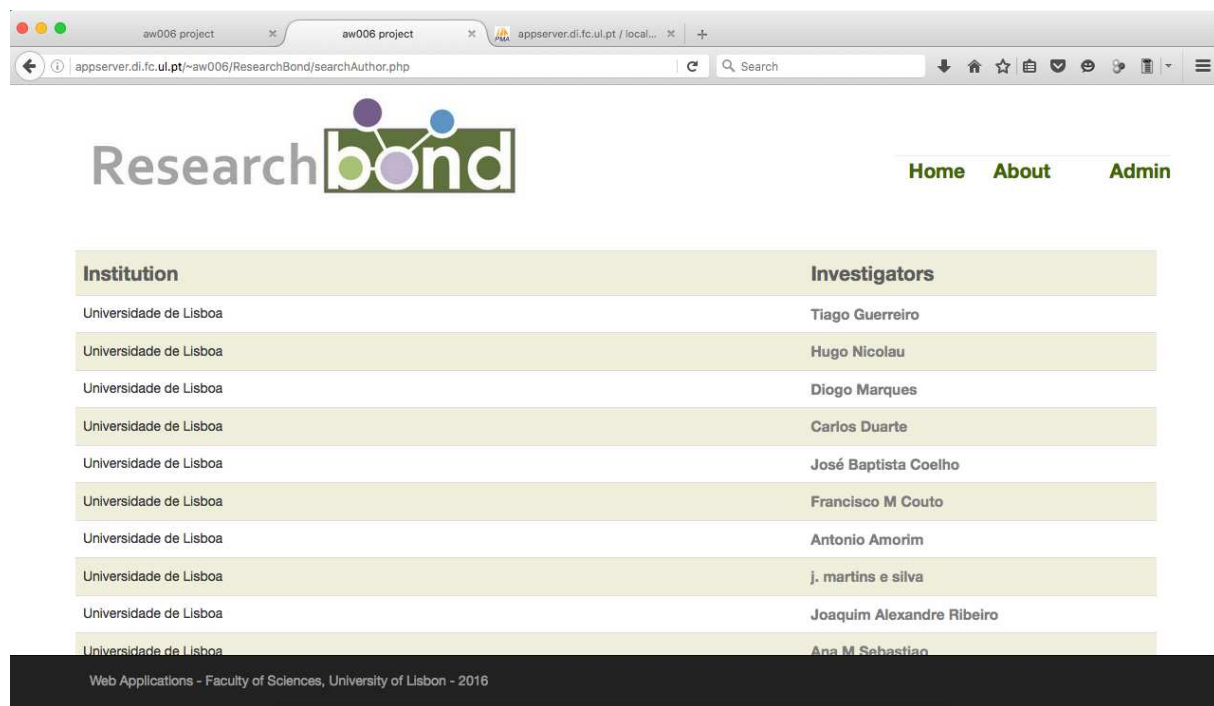


Figura 4 - Página interna do *Admin* após clicar no botão Remove, pesquisar por nome específico e Search!. Ao clicar num botão Remove, o autor correspondente será removido da Base de dados

Quando é feita uma pesquisa por instituição, o utilizador é encaminhado para uma página que mostra uma tabela (Figura 6) com todos os autores registados nessa instituição. Clicando nos nomes dos autores, é gerada uma rede para o autor. Se a pesquisa for feita por um nome parcial um autor, é gerada uma tabela com vários autores que podem corresponder a esse nome. Estes autores podem também ser seleccionados para gerar a visualização.

Quando é seleccionado um autor a aplicação encaminha para a página de informação do autor (`authorInfo.php?author=`), que se encontra subdividida em duas partes. Na parte esquerda está a informação sobre o autor, instituição e co-autores e publicações mais relevantes. São mostrados os cinco primeiros co-autores com maior co-participação nas publicações e as cinco publicações com maior número de citações. Na parte direita da página é mostrada a representação construída em D3 da rede do autor e dos seus co-autores (Figura 7). Esta página tem a particularidade de estar anotada com anotações semânticas, definidas em RDFa. Após a pesquisa de um autor, escolhendo o botão *See Full Information*, é disponibilizada a lista de publicações e, sempre que disponível, é apresentado para cada publicação a avaliação feita pela aplicação Altmetric. Esta API utiliza a tecnologia Ajax, que permite uma comunicação assíncrona. As interações do Ajax são iniciadas pelo código JavaScript. Quando a interação do Ajax é concluída, o JavaScript atualiza o código-fonte HTML da página (`fullInfo.php?author`) com a informação completa do autor. É importante referir que a API Altmetric é muito restricta quanto ao que constitui um método *call back* válido, sendo que apenas letras, números e subtraços (underscore) são permitidos. Desta forma, o método *call back*, que permite a actualização automática da página, só é concluído quando esta regra é cumprida, o

que resulta na ausência do símbolo Altmetric para as publicações cujo DOI não respeita esta regra, nomeadamente devido à presença de uma barra. Nos casos em que a aplicação CrossRef não recuperou o DOI da publicação, por o mesmo ser inexistente ou pela semelhança de títulos encontrados ser inferior a 80%, o símbolo Altmetric surge com fundo cinzento e um ponto de interrogação (Figura 8).



Institution	Investigators
Universidade de Lisboa	Tiago Guerreiro
Universidade de Lisboa	Hugo Nicolau
Universidade de Lisboa	Diogo Marques
Universidade de Lisboa	Carlos Duarte
Universidade de Lisboa	José Baptista Coelho
Universidade de Lisboa	Francisco M Couto
Universidade de Lisboa	Antonio Amorim
Universidade de Lisboa	j. martins e silva
Universidade de Lisboa	Joaquim Alexandre Ribeiro
Universidade de Lisboa	Ana M Sebastian

Web Applications - Faculty of Sciences, University of Lisbon - 2016

Figura 5 - *Página de output após pesquisa de um Instituição.*

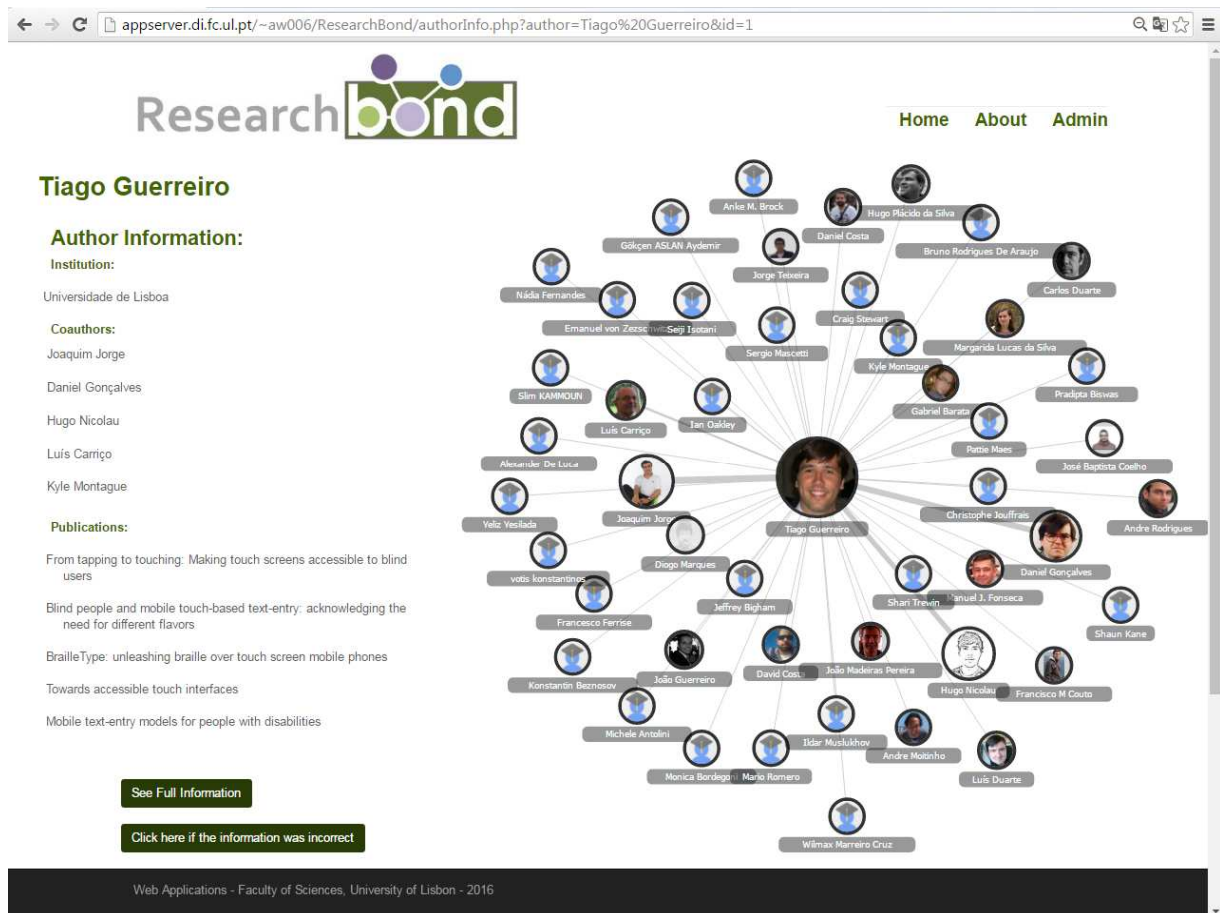


Figura 6 - Página de output após pesquisa de um Autor específico.

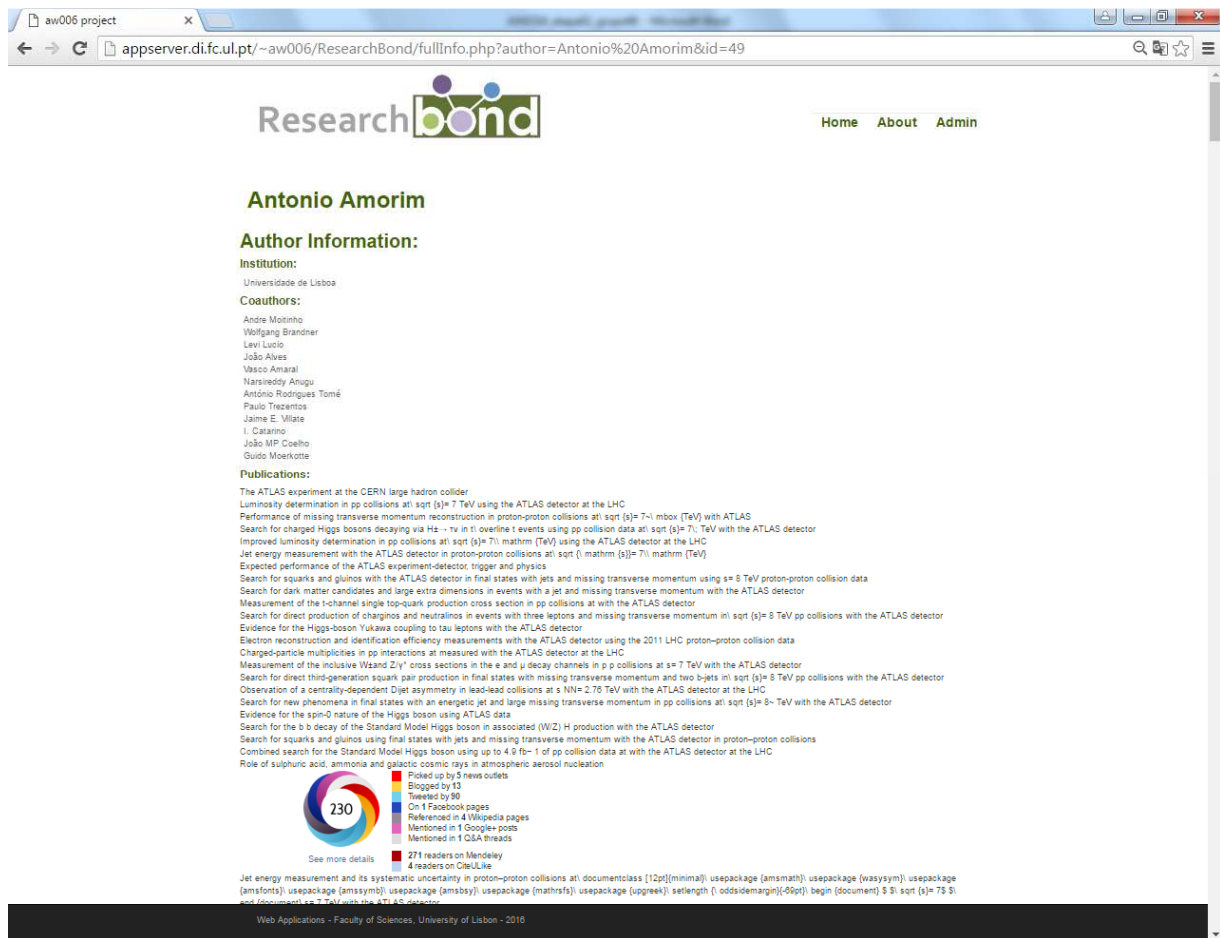


Figura 7 - Execução do método *call back* da API Altmetric na página *Full Information* sobre um autor.

7. Discussão

1. Problema 1 encontrado

Um problema encontrado relaciona-se com as limitações impostas pelo Google Scholar e pela API CrossRef. Estas não permitem uma busca intensiva de informação, levando à interrupção das funções de scrapper.

Uma forma de contornar esta questão passou pela criação de duas modalidades de scraping, por instituição e por autor. Foram também introduzidos tempos aleatórios de espera entre autores ou paginas através da função sleep em PHP.

2. Problema 2 encontrado

A estrutura da informação existente no Google Scholar é muitas vezes pouco estruturada. Por exemplo, enquanto muitos autores se encontram associados ao perfil da Instituição Universidade de Lisboa, muitos outros apresentam no seu perfil siglas ou referências à instituição em Inglês (University of Lisbon). Esta questão não é facilmente gerida pelos scrapers, levando à criação de diferentes entradas (redundantes) na base de dados. A resolução deste aspecto pode passar pelo feedback do utilizador.

Por outro lado, a pesquisa por co-autores apenas se limita aos co-autores adicionados no perfil do autor no Scholar, na barra lateral do perfil. Por cada pesquisa no perfil de um autor principal, a informação recolhida dos co-autores limita-se apenas ao seu nome e link (do Google Scholar). Deste modo, para aceder à informação dos co-autores e determinar qual a percentagem de trabalho feito em conjunto, estes co-autores também devem constar na base de dados com a sua informação atualizada. Ou seja, o scraping também teve de ser corrido para o perfil dos co-autores correspondentes, individualmente ou num scraping por instituição. As limitações do scraping no Google Scholar leva a que esta informação não se encontre completa para todos os autores da base de dados. Tendo e conta o volume de informação que as redes de autores implicam, a Base de Dados ainda necessita de actualização.

3. Problema 3 encontrado

A API Altmetric só conclui o método *call back* quando o DOI é constituído apenas por letras, números ou subtraços (*underscore*). Algumas publicações utilizam a barra na constituição do DOI, pelo que nestes casos não é possível apresentar o resultado da análise efectuada pelo Altmetric. Uma vez que o DOI é uma identificação única e não alterável de cada publicação, também não é possível remover ou substituir a barra por outro carácter aceite.

4. Problema 4 encontrado

O código entre os comentários “// Make picture circular” e “// Circle around Picture” só parece funcionar correctamente no Google Chrome. Depois de muito tentar perceber o erro, este não ficou resolvido. Continua a ser um mistério para nós o porquê de aquele pedaço de código não funcionar

correctamente noutros Browsers. O erro manifesta-se no D3, ao mostrar as fotografias dos co-autores com uma imagem de tamanho constante, em vez de variável conforme o tamanho que seria suposto.

5. Problema 5 encontrado

Durante o desenvolvimento do Front-end houve apenas um erro que foi salientado pelo validator.w3 e que não foi possível resolver. Este erro ocorreu nos script authorInfo.php, adminOptions.php e tem a descrição "The element button must not appear as a descendent of a element". Não foi possível corrigir este erro em tempo útil.

Crítica Global

No geral estamos contentes com a versão final do projecto. Conseguimos atingir a grande maioria dos nossos objectivos, mesmo tendo em conta o facto de termos perdido um membro do grupo e o facto de nenhum de nós ter um background extenso em programação (nenhum de nós é licenciado em informática) e nunca termos interagido a fundo com JavaScript, PHP, Ajax e D3.

Com mais tempo e mais experiência nestas linguagens conseguiríamos desenvolver um aplicação com aspecto e performance muito mais profissional

8. Anexos

Sketchs iniciais do site:



