LABORATÓRIO NACIONAL
DE ENGENHARIA CIVIL

CENTRO DE TECNOLOGIAS DA INFORMAÇÃO
Núcleo de Tecnologias da Informação
em Engenharia Civil

Proc. 1302/27

REPOSITÓRIO DIGITAL DO LNEC

Progressão entre 2009 e 2012

Lisboa • Fevereiro de 2012

OAC&T TECNOLOGIAS DA INFORMAÇÃO

RELATÓRIO 24/**2012 – CTI/NTIEC**

Repositório Digital do LNEC
Progressão entre 2009 e 2012

RESUMO

O Repositório Digital de Documentos do LNEC tem vindo a crescer continuamente, desde que entrou oficialmente em produção, em 2009, evoluindo desde esta altura, e cumprindo a sua função de acervo centralizado das publicações do Laboratório e facilitando a consulta das mesmas por parte dos funcionários. Este documento descreve a evolução da plataforma desde a sua entrada em produção, os seus problemas e o que se fez para os resolver, e apresenta uma visão crítica do estado atual da mesma, propondo alguns pontos chave para assegurar a continuidade do repositório no LNEC.

ABSTRACT

LNEC's Digital Repository has been growing continuously, since it was deployed in a production server, in 2009, evolving since then, and fulfilling its role in centralizing LNEC's publications, providing ease of access to all the employees. This document describes the evolution of the platform since its entrance in production, its problems and what has been done to overcome them, and it presents a critical vision of the repository's current state, pointing out some key points to assure its continuity in LNEC.

Índice

Figuras

1 Introdução

O projeto de Repositório Digital do LNEC, suportado pela aplicação de software livre DSpace, conta já com cerca de 3 anos de existência como Repositório Institucional oficial, e com cerca de 5 anos desde a sua primeira fase de testes. Desde a sua fase inicial até à data atual, foram introduzidos no Repositório cerca de 10.300 documentos, sabendo-se que este número irá ter um crescimento contínuo, fruto das atividades de investigação e projetos por contrato que continuarão a existir no Laboratório.

Ao longo deste tempo, foram necessárias algumas intervenções, a diversos níveis, a fim de afinar pormenores funcionais da plataforma, e a assegurar um acesso facilitado à aplicação. De facto, o principal problema na utilização da plataforma, transversalmente aos perfis de utilizador existentes no Laboratório, tem sido a ocorrência de erros de acesso via *browser* ao Repositório, meio essencial para a sua utilização. Alguns outros problemas prendem-se com o "ruído" da informação introduzida no Repositório, ou mais formalmente, a sua falta de normalização. Este ruído, como facilmente se entende, põe em causa a correta e expedita utilização da plataforma, dificultando, por exemplo, as pesquisas e obrigando os utilizadores a realizar mais passos para contemplar eventuais duplicações de dados, que levam a resultados distintos aquando da pesquisa.

Este documento descreve as ações continuamente empreendidas para tornar o Repositório Digital mais fácil e fiável de usar.

2 Migração ANDROMEDA – HECATE

Este secção descreve os problemas que originaram a migração do Repositório do servidor *Andromeda* para o servidor *Hecate*, iniciada nos finais de 2009 e terminada no início de 2010.

2.1 Problemas com o servidor Andromeda

Desde a entrada em produção, o Repositório foi replicado da sua instância de teste, no servidor *Bootes*, para um servidor de produção já existente, o servidor *Andromeda*. No entanto, com o aumento da intensidade de utilização, o servidor onde foi alojado o Repositório sofreu uma sobrecarga nas suas capacidades, dificultando em grande medida a utilização da plataforma. Estas dificuldades eram manifestadas sob a forma de lentidão a apresentar resultados de pesquisas, e até mesmo na navegação no *site* do Repositório. Sendo este servidor partilhado por outros serviços indispensáveis e críticos, a alternativa foi iniciado o estudo de uma possível migração do Repositório para um servidor dedicado.

2.2 Migração para o servidor Hecate

Com o intuito de resolver o problema da lentidão que tornavam penosa a utilização do Repositório, definiu-se um plano de migração para um servidor dedicado ao Repositório.

Existindo já experiência na virtualização de serviços dentro do CTI, e pela facilidade que o processo apresentava face à aquisição de novo *hardware*, ficou estabelecido que o novo servidor do Repositório seria virtual, assente na plataforma de virtualização *VMware*. Ficou ainda estabelecido que, aproveitando as alterações necessárias que o Repositório teria sofrer, seria feita a tentativa de incorporar todas as alterações feitas à medida para uma versão mais recente da plataforma DSpace, avançando assim da versão 1.3.2 para a versão 1.5.

A migração ficou planeada e realizada com as seguintes fases:

- instalação da plataforma DSpace e dependências no novo servidor;
- adaptação da versão 1.5 do DSpace com vista a nova versão das alterações previamente desenvolvidas para a versão 1.3.2;
- desenvolvimento de novas extensões, para potenciar a migração;
- testes de funcionamento;
- migração faseada de toda a informação para a nova plataforma;
    - estrutura hierárquica de organização (comunidades e coleções);
    - autores;
    - utilizadores;

- [definição de permissões de acesso]
- documentos e ficheiros:

- testes de funcionamento com grupos de teste;
- desativação do repositório DSpace no servidor *Andromeda*;

2.3 Arquitetura do servidor *Hecate*

Como referido na secção anterior, ficou estabelecido que o repositório passaria a estar alojado num servidor virtual, dedicado apenas a este serviço. Adicionalmente, foi determinado que alguns componentes do repositório ficariam alojados noutro servidor, nomeadamente a base de dados e sistema de ficheiros onde são guardados os documentos propriamente ditos (ficheiros). Esta configuração foi pensada com a finalidade de minimizar potenciais problemas a nível de escalabilidade e manutenção, permitindo que novas configurações e versões da plataforma sejam testadas paralelamente ao servidor virtual e sobre os dados reais, sem causar qualquer impacto na sua utilização. Do ponto de vista da manutenção, esta configuração permite gerir de forma isolada tanto a base de dados, como a localização dos documentos, e ainda as configurações do próprio servidor virtual.

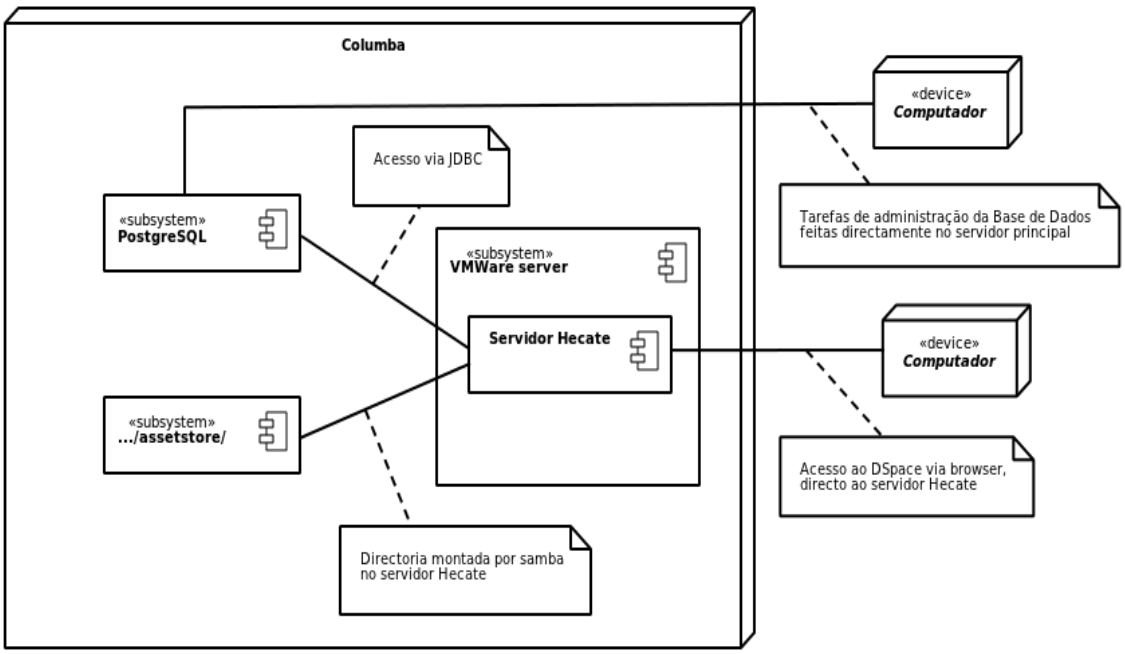

Figura 1: Arquitetura do sistema

A figura 1 mostra a arquitetura seguida. O servidor *Columba* age como anfitrião de várias máquinas virtuais, utilizando a plataforma *VMWare Server*, de entre as quais, o servidor *Hecate*. Apesar de ser uma máquina virtual, o servidor *Hecate* age como se fosse uma

máquina física, disponibilizando o repositório DSpace de forma a que seja acessível quando se acede por *browser* ao servidor *Hecate*. Como se pode ver ainda na figura 1, a base de dados está também alojada no servidor *Columba*, tal como o espaço em disco onde são guardados. O DSpace acede à base de dados por um *driver JDBC*, da mesma forma que faria caso a base de dados estivesse alojada no servidor *Hecate*. O local onde o DSpace irá guardar os ficheiros que compõem os documentos, embora fisicamente no servidor anfitrião, está montado via *Samba* no servidor *Hecate*, pelo que o acesso à diretoria de armazenamento (designada pelo DSpace como *assetstore*) é realizado de forma completamente transparente, tal como se estivesse no próprio servidor.

2.4 Migração da estrutura de comunidades/colecções

A versão 1.5 do DSpace possui uma ferramenta de auxílio da construção da estrutura hierárquica de comunidades e coleções, automatizando o seu processo de criação partindo da utilização de um ficheiro XML contendo a informação necessária. Pela quantidade considerável de comunidades e coleções existentes na configuração da estrutura do Repositório do LNEC, optou-se por aproveitar este mecanismo para fazer a migração desta mesma estrutura para o novo servidor. Para tal, foi necessário desenvolver, na plataforma de origem, o mecanismo de exportação da estrutura de comunidades e coleções, já que essa ferramenta não estava disponível, e o custo da sua implementação era inferior à replicação manual da mesma estrutura no novo servidor.

Na verdade, com a finalidade de homogeneizar as coleções de cada núcleo, o ficheiro resultante da exportação da hierarquia foi modificado manualmente de forma a que cada núcleo tivesse exatamente as mesmas coleções onde se depositariam os diferentes tipos de documentos. Depois desta alteração, o ficheiro foi alimentado ao novo servidor, para que a nova estrutura, completamente uniforme, pudesse ser criada.

A figura 2 exemplifica os componentes que efetuaram a exportação e importação da estrutura.

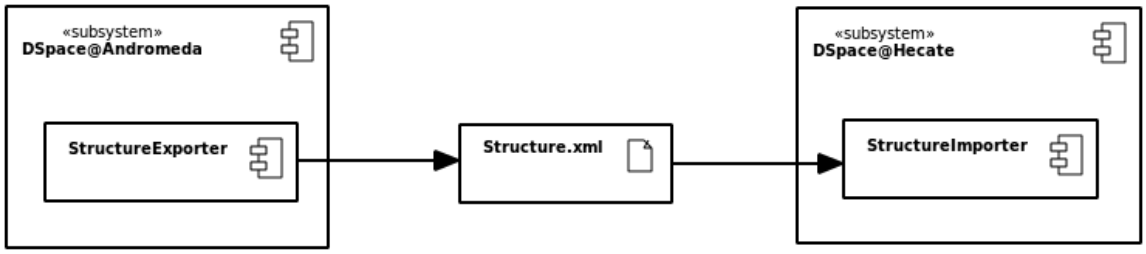

Figura 2: Migração da estrutura de comunidades e coleções

A migração da estrutura foi efetuada com sucesso, sendo posteriormente necessário ajustar manualmente as permissões manualmente.

2.5 Definição de permissões

Dado que existem vários papeis *(roles)* de utilizador no DSpace, foi necessário contemplar cada um deles com um grupo de permissões adequado. Desta forma, e seguindo a Nota de Serviço 15/2009, a atribuição de permissões efetuou-se nos seguintes moldes:

2.5.1 Submissão de documentos (artigos de revista e comunicações a congressos)

Para cada núcleo de cada departamento/centro foi criado um grupo que integraria todos os utilizadores autorizados a submeter documentos. Seguindo a Nota de Serviço 15/2009, os responsáveis pela introdução dos documentos seriam os próprios investigadores e bolseiros que os produzissem, desde que fossem os primeiros autores. Desta forma, os utilizadores a integrar estes grupos seriam os investigadores e bolseiros de cada núcleo, embora algumas exceções tenham surgido no que toca à delegação da responsabilidade de submissão de documentos para funcionários administrativos.

A formatação dos nomes dos grupos obedece à regra: *departamento_núcleo_SUBMIT*.

Outra exceção é referente ao DE/NEM, que é responsável pela impressão dos seus próprios relatórios, pareceres e notas técnicas, pelo que lhe foi atribuída permissão para depositar estes documentos, através da criação de um grupo dedicado *(de_nem_SUBMIT)*.

É de notar que os documentos introduzidos por estes grupos de utilizadores estão sujeitos a um processo de mediação, onde um determinado utilizador é responsável pela validação da submissão, podendo, em caso negativo, remetê-la novamente ao utilizador original para correções (vide secção 2.5.3).

2.5.2 Submissão de outros documentos

Seguindo a Nota de Serviço 15/2009, todos os outros documentos, cujo processo de impressão passasse pelas artes gráficas (DSLM) seriam introduzidos nesse mesmo sector. Como tal, um grupo foi definido de modo a integrar todos os responsáveis da DSLM pela introdução de documentos *(dslm_SUBMIT)*. A esse grupo foi atribuída permissão de submissão em todas as coleções de todos os núcleos, excetuando as contempladas no ponto anterior (vide secção 2.5.1).

2.5.3 *Workflow* de submissão

As permissões para o mecanismo de mediação de submissões, já existente no servidor Andromeda, foram também reformuladas e elaboradas de forma a ficarem ativas de forma mais simples e eficaz.

Para cada departamento foi selecionado um utilizador responsável por validar ou recusar os documentos. Desta forma, para contemplar a possível adição de mais utilizadores de

mediação, foram criados grupos de mediação, com a seguinte regra de nomenclatura: *MED_DEPARTAMENTO_internos*.

Cada grupo foi preenchido com o utilizador selecionado para desempenhar o papel de mediador, e posteriormente este grupo integrou o processo de "*depósito com workflow*" para cada coleção de cada núcleo do respetivo departamento.

O processo de submissão com *workflow* é descrito na figura 4.

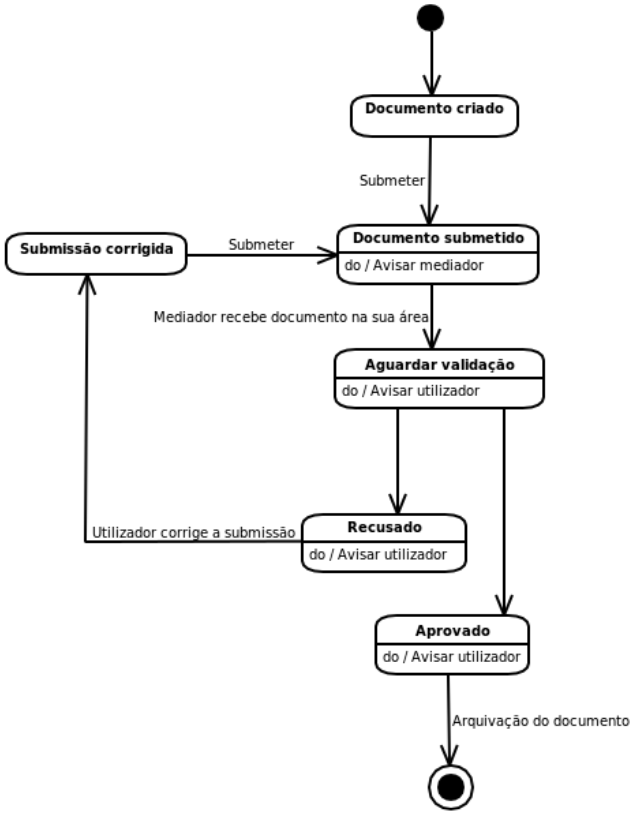

Figura 3: Diagrama de estados do processo de mediação

2.5.4 Acesso a documentos e ficheiros

Para a atribuição de permissões de consulta de documentos e ficheiros, foram criados grupos aos quais foram afetados os utilizadores consoante os núcleos a que pertencem – a nomenclatura seguida foi a seguinte: *departamento_núcleo*

As regras seguidas para a atribuição das permissões aos diferentes núcleos foram:

- um utilizador anónimo (não autenticado) tem acesso a todas as coleções não confidências presentes no sistema;
- um utilizador autenticado, não pertencente a nenhum grupo de núcleo apenas tem acesso a todas as coleções não confidenciais presentes no sistema;

- um utilizador autenticado, pertencente ao núcleo X tem acesso a todas as coleções não confidenciais, bem como às coleções confidenciais do núcleo X, mas não às coleções confidenciais do núcleo Y;

A política de acesso para consulta da informação é exemplificada pelo diagrama na figura 4.

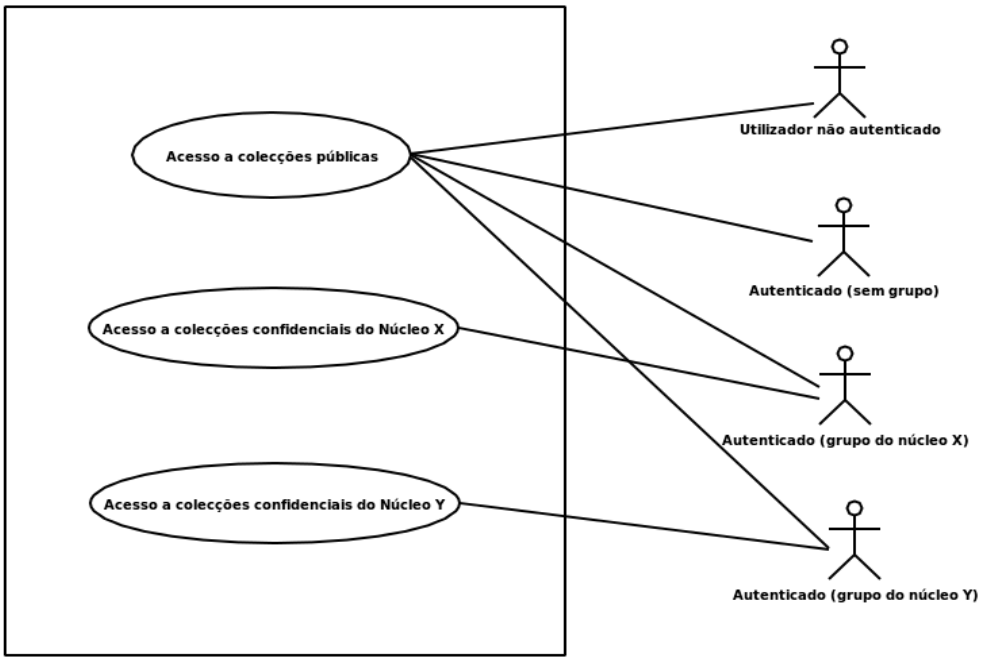

Figura 4: Diagrama de caso de uso de acesso às coleções

2.6 Migração dos dados

Findas todas as configurações de permissões, iniciou-se o processo de migração dos documentos para o novo servidor. Para realizar esta migração foi necessário alterar o procedimento de exportação de itens, de modo a organizar de forma mais legível a informação. Na prática a organização final replicou a organização interna do repositório, com comunidades e coleções. Assim, uma pasta com o nome do departamento continha pastas com o nome dos núcleos, que por sua vez continham pastas das coleções que nele estavam representadas. Finalmente, as pastas das coleções continham todos os documentos que nelas constavam no repositório.

Como descrito na secção 2.4, a organização hierárquica do repositório foi uniformizada, de forma a que todos os núcleos tivessem as mesmas coleções, tornando mais fácil a futura introdução de documentos dos diversos tipos. No entanto, como no servidor *Andromeda* as coleções não estavam uniformes, foi necessário um esforço de mapeamento para as novas coleções, mapeamento este que ocorreu manualmente, coleção a coleção.

3 Migração HECATE – COLUMBA

Na secção anterior foi descrita a justificação e o processo de migração do servidor *Andromeda* para o servidor *Hecate*. Embora os problemas tenham ficado resolvidos com esta migração, passados alguns meses de funcionamento sem problemas a assinalar, os problemas de acesso voltaram-se a manifestar. Seguidamente descreve-se o recente processo de passagem da plataforma para um servidor físico, abandonando assim a aproximação do servidor virtual.

3.1 Problemas com o servidor *Hecate*

Os problemas sentidos apenas se tenham manifestado a nível de acesso (sentido através de sucessivos avisos de "Erro interno de sistema", o erro mostrado por omissão quando alguma ação falha, enquanto os utilizadores realizavam as suas atividades no Repositório), e não a nível de tempo de acesso ou capacidade de resposta do servidor, como era sentido no servidor anterior. Coerente com o sentido anteriormente, ao reiniciar o servidor aplicacional *Tomcat*, verificou-se que o problema passava, embora com carácter temporário, e não linear, isto é, ocasionalmente o problema não se manifestava durante alguns dias, de outras vezes não demorava alguns minutos a repetir-se. Uma das formas de tentar contornar a situação, foi a implementação de um *cronjob* que reiniciava o servidor aplicacional, regularmente (de hora a hora) bem como limpava as ligações que indicavam o estado de "*idle in transaction"*. Verificou-se que esta abordagem reduz o número de vezes que é necessário intervir no servidor, embora não tenha resolvido completamente o problema.

Com a finalidade de averiguar se o problema poderia ser atribuído a alguma atualização não controlada dos componentes onde assentam o Repositório, ou a outro fator, considerou-se a passagem temporária do Repositório para o servidor físico *Columba*. Esta passagem aconteceu durante o mês de Dezembro de 2011.

3.2 Migração para o servidor Columba

Em contraste com a anterior migração, o procedimento para a passagem do servidor *Hecate* para o servidor *Columba* substancialmente mais simples, fruto da arquitetura levada a cabo na migração anterior. Desta forma, e visto que tanto a base de dados e sistema de ficheiros se encontravam já alojados no servidor *Columba*, apenas foi necessário instalar e configurar a plataforma DSpace, e suas dependências novo servidor, ligando posteriormente esta instância à base de dados e sistema de ficheiros. Adicionalmente, de forma a tornar esta transferência de sistemas completamente transparente ao utilizador, todos os acessos ao servidor *Hecate* foram redirecionados para o novo servidor.

Após a sua ativação sem incidentes, detetou-se a ocorrência dos mesmos erros, embora de forma menos numerosa. Como tal, optou-se por deixar que a plataforma se mantivesse a funcionar no servidor *Columba*, tendo-se posteriormente desativado o servidor *Hecate.*

4 Outras linhas de trabalho

Esta secção detalha algumas linhas de trabalho paralelas ao Repositório Digital do LNEC.

4.1 Instância externa e RCAAP

Dado que o Repositório Digital do LNEC é apenas acessível através da rede interna do Laboratório, facto justificado pela diversidade de documentos confidenciais que nele constam, considerou-se em 2009 a possibilidade de ativar uma instância paralela, acessível do exterior, para que se pudesse disseminar para a comunidade algum do conhecimento produzido. A escolha de uma plataforma paralela, fisicamente separada, tencionava assegurar que não haveria qualquer perigo de acesso a documentação confidencial, já que na instância exterior apenas seriam colocados os documentos selecionados. Em finais de 2009, começou a ser estudado o mecanismo de passagem semi-automático de documentos da instância interna para a futura instância externa, para que não houvesse esforço duplicado de introdução de documentos em ambas as instâncias.

Com o inicio do projeto RCAAP[1] (Repositório Científico de Acesso Aberto de Portugal) pela FCCN, abandonou-se a ideia do repositório externo, dada a perspetiva de colaboração com esta plataforma. Durante os anos de 2010 e 2011, foram realizadas algumas atividades, pelo Engº Miguel Rosado, para prosseguir com esta linha de trabalho, embora esteja de momento parada.

4.2 Documentos normativos do CQC/NNR

Com a finalidade de estender a utilização do DSpace a outro domínio, foi realizada uma tentativa de especificar os requisitos para que se pudesse ter, paralelamente à instância normal do Repositório, uma outra instância que contivesse os documentos normativos do CQC/NNR. Embora no final de 2009 tudo estivesse acertado para que se prosseguisse esta linha, nomeadamente as necessidades a nível estrutural a realizar na plataforma, sendo que o passo seguinte seria a concretização do servidor e instalação da plataforma, esse passo nunca se deu, tendo-se, aparentemente, abandonado este objetivo.

1 http://www.rcaap.pt/

5 Estatísticas

Desde que entrou em funcionamento, o Repositório do LNEC já acumulou cerca de 10400 publicações, o que denota uma considerável metodologia de uso, apesar das dificuldades de usabilidade que se fizeram sentir.

5.1 Depósitos no DSpace

O gráfico da figura 5 ilustra o número de publicações inseridas no repositório, mensalmente, desde Maio de 2010, data da entrada em funcionamento do servidor *Hecate*.

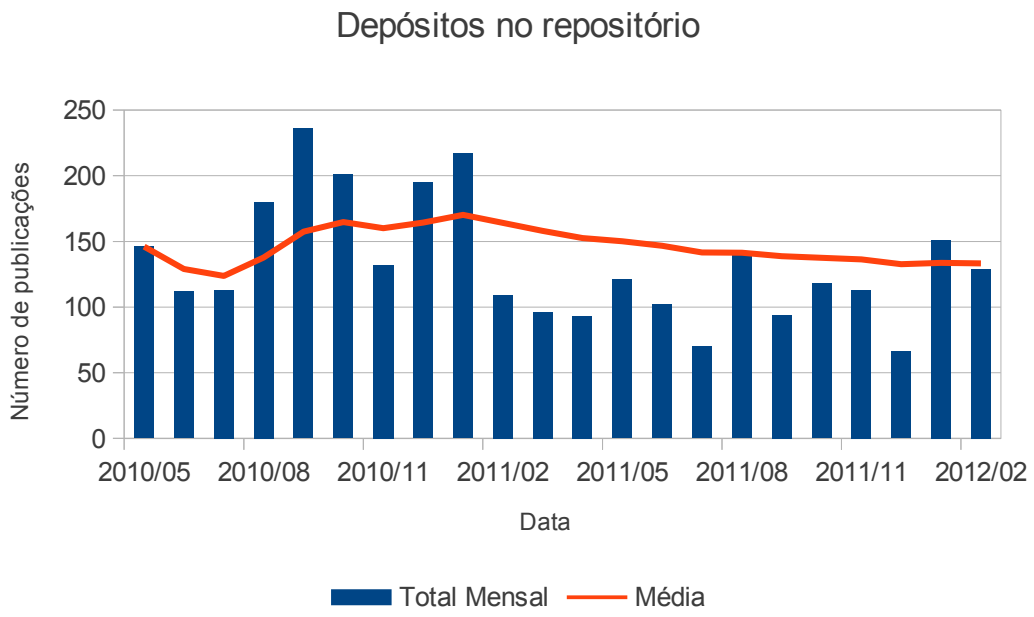

Figura 5: Depósitos mensais no DSpace e respetiva média

5.2 Outras estatísticas

O gráfico da figura 6 ilustra o número de vezes, por mês, que algumas ações foram efetuadas. Estes valores, provenientes do módulo de estatísticas do repositório, permitem ter uma ideia da intensidade da utilização do repositório. Pode-se constatar claramente um elevado uso de pesquisas e documentos consultados, contrastando com um menor número, mas ainda assim significativo, de ficheiros consultados e autenticações.

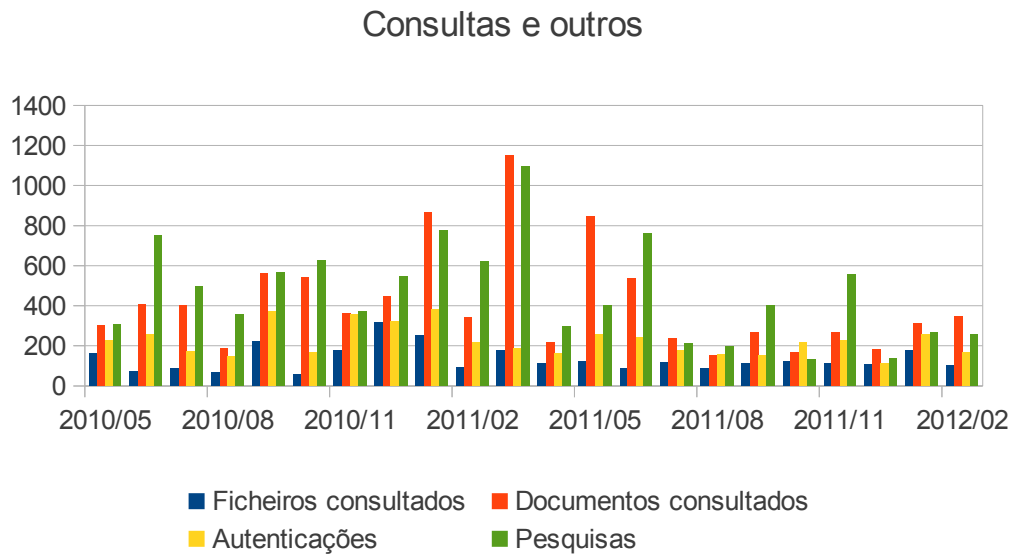

Figura 6: Consultas, pesquisas e autenticações efetuadas no repositório

O gráfico da figura 7 mostra algumas correlações entre as ações referidas na figura anterior. Como se pode ver, existe uma forte correlação entre a autenticação e a consulta de ficheiros, ao contrário da autenticação e a consulta de documentos. Isto pode indicar que, no caso das consultas de ficheiros, se está a falar de documentos confidenciais. No caso dos documentos consultados, a sua fraca correlação com a autenticação e forte com as pesquisas, clarifica o caso de uso mais habitual no repositório, ou seja, a pesquisa resultando numa consulta rápida da informação do documento, sem autenticação, e sem consultar os ficheiros do documento. Este cenário é também apoiado pela fraca correlação entre a autenticação e as pesquisas.

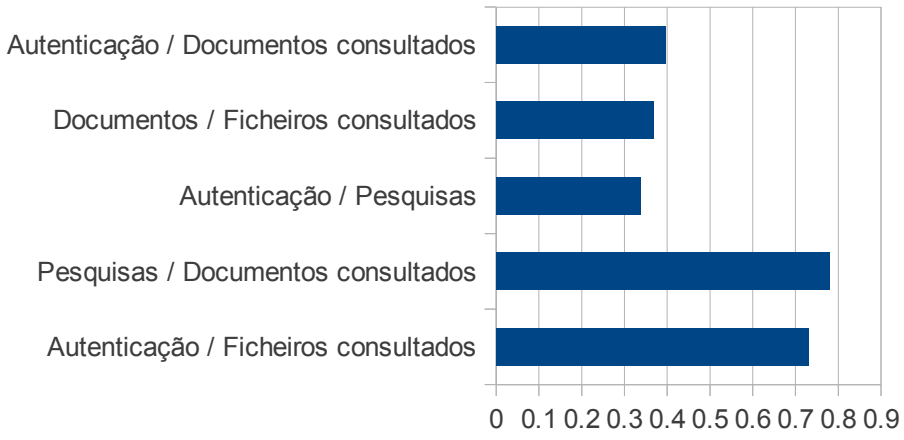

Figura 7: Correlações entre ações

6 Conclusões e trabalho futuro

6.1 Atualização da plataforma

Embora não seja uma situação de extrema importância, dever-se-ia ponderar fazer uma atualização da plataforma para uma versão mais recente, nomeadamente a 1.8.1, à altura da elaboração deste relatório. Isto envolverá um novo esforço de replicação das alterações realizadas, embora se considere que seria benéfico uma alteração/otimização de alguns procedimentos. Concomitante com a atualização seria a possibilidade da identificação e eliminação na sua totalidade dos erros de acesso que tanto incomodam os utilizadores, embora, na versão atual, sejam consideravelmente mais reduzidos.

6.2 Aumento da qualidade da informação

Na sua fase inicial, a possibilidade da introdução descentralizada e pouco regulada de publicações originou uma redução da qualidade de dados. A utilização de maiúsculas para preencher a totalidade dos campos dos formulários de introdução, incluindo títulos e autores, bem como a possibilidade de escrever manualmente os nomes dos autores, foram os dois principais mecanismos responsáveis pela redução da qualidade da informação constante no repositório. Embora essas duas práticas tenham sido desaconselhadas e controladas, ainda existe uma quantidade considerável de ruído que convém ser corrigido afim de melhorar principalmente a usabilidade do repositório, embora também o seu aspeto em termos de organização de resultados de pesquisas.

6.3 Alteração do visual

O visual do repositório tem-se mantido inalterado desde a sua entrada em vigor no laboratório. Este visual corresponde também ao aspeto por omissão que a plataforma DSpace possui. Embora se trate apenas de uma questão de visual, o autor considera relevante realizar um *facelift* ao repositório, que talvez se possa coadunar com a atualização da plataforma, transmitindo um carácter de renovação e preocupação com a plataforma por parte do Laboratório.

6.4 Implementação de mecanismos de controlo

Criação de eventos para enviar e-mails afim de haver registo sobre quem altera/remove o quê e quando, em vez de o mesmo se realizar apenas quando há depósito/aceitação de documentos. Adicionalmente, ou em substituição, poder-se-ia implementar a persistência desta informação na base de dados, posteriormente consultável de forma manual, ou mesmo integrada na plataforma.

6.5 Estatísticas

Embora a plataforma DSpace possua um módulo de estatísticas, estas são baseadas nos

seus *logs*, e a informação que se pode nele consultar é mais genérica. De facto, não é preocupação do DSpace fornecer estatísticas muito complexas, já que o seu "negócio" é apenas de repositório de dados, sendo que as estatísticas que disponibiliza são mais direcionadas à gestão da própria plataforma. No entanto, a Direção do LNEC, com a necessidade de extrair informação relevante e necessária para o desempenhar das suas funções no que toca à contagem do número de publicações, já manifestou por diversas vezes a vontade de possuir uma forma de obter essa informação sem passar pelo tratamento manual da informação. Desta forma, o autor considera relevante que, alinhado com a atualização da plataforma, seja analisada a possibilidade de desenvolvimento de uma forma de persistência e consulta desta informação, embora não considere taxativo que esta deva figurar integrada no Repositório.

Lisboa, em 29 de Fevereiro de 2012

VISTOS	AUTORIA
Coordenador do CTI	
Luís Arriaga da Cunha	Rui Gamito
Investigador Coordenador	Bolseiro de Doutoramento