FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

FEUP

Modelo de Replicação para a Preservação e Interrogação de Dados Científicos

Micael Ferreira Alves de Pinho

Mestrado Integrado em Engenharia Informática e Computação

Orientadora: Maria Cristina de Carvalho Alves Ribeiro (Eng^{a}.)

10 de Fevereiro de 2012

Modelo de Replicação para a Preservação e Interrogação de Dados Científicos

Micael Ferreira Alves de Pinho

Mestrado Integrado em Engenharia Informática e Computação

Orientadora: Maria Cristina de Carvalho Alves Ribeiro (Eng[a].)

10 de Fevereiro de 2012

Resumo

A entrada no mercado, das novas tecnologias digitais impulsionou o nosso mundo para uma era, em que a criação, manipulação e o armazenamento de informação de forma digital, cresceu exponencialmente. Por outro lado, surgiram alguns problemas relacionados com a preservação e interpretação dessa mesma informação.

No contexto de investigação, os conjuntos de dados (*datasets*) recolhidos, devido à existência de uma grande diversidade de áreas de investigação, têm estruturas e informações bastantes variadas. Os *datasets* podem ser conteúdos textuais, imagens ou audiovisuais.

Tendo em consideração esta diversidade existente é necessário uma descrição muito cuidadosa de cada um dos *datasets*, desde o tipo de dados envolvidos às condições de recolha e de utilização. Além disso, a preservação e acessibilidade destes conjuntos de dados é de extrema importância para a validação dos resultados obtidos em investigações e constituem uma importante fonte de evidência para trabalhos futuros.

Atualmente, já existem repositórios de conteúdos digitais, que permitem o armazenamento e preservação de *datasets*, contudo, existem alguns acontecimentos que podem colocar em riso o seu acesso. Como tal, pretende-se a criação e análise de um sistema que permita a replicação de *datasets* em diferentes localizações, precavendo assim as falhas de acesso. Para a implementação deste sistema, pretende-se testar duas tecnologias de replicação existentes, LOCKSS e DuraCloud.

Pretende-se também a implementação de um sistema de interrogação, que permita fazer consultas intuitivas sobre a listagem dos *datasets* existentes, bem como a extração de registos dos mesmos, segundo certas restrições.

Os dois sistemas desenvolvidos, o de interrogação e de replicação, serão testados num repositório de dados científicos, e como tal, será simulado um repositório, que está a ser desenvolvido no âmbito de um projecto da Reitoria da Universidade do Porto, denominado de UPData, que tem por objetivo, o armazenamento e preservação de dados científicos.

A preservação de dados, é um problema que todas as instituições de investigação enfrentam, tanto ao nível nacional como internacional e, como tal, pretende-se encontrar uma solução que contribua para a sua resolução, permitindo assim o acesso aos dados científicos a um conjunto diversificado de instituições de investigação.

Por fim, pretende-se fazer a avaliação dos sistemas implementados, tendo em vista os requisitos estabelecidos para o sistema de preservação e as funcionalidades do sistema de interrogação.

Abstract

The entry of the new digital technologies in the market, lunch our world into an era in which the creation, manipulation and storage of digital information, has grown exponentially. On the other hand, there were some problems with the preservation and interpretation of such information.

In research activities, duo to the existence of a wide range of areas, the datasets have many different's structures and kinds of information. The datasets can be textual content, images or audiovisual.

Because of the existence diversity, is required very thorough description of each of the datasets, for example: type of data involved, collection conditions and conditions to reuse. Behind that, the preservation and accessibility of these datasets is extremely important to validate the results obtained in investigations and for reuse in future researchs.

Nowadays, there are some digital repositorys that allow the storage and preservation of datasets, however, there are a few events that could put the access in risk. So the goal of this project, is the development and test of a system that allows the replication of datasets in different locations, avoiding the failures of access. Two technologies will be tested in this system: LOCKSS and DuraCloud.

Another goal, is the implementation of an interrogation system, which allows you to do intuitive consultations on the list of existing datasets, and the extraction of some records from a specific dataset, according to certain restrictions.

The two systems developed, will be tested in a repository of scientific data, and for that, a repository will be simulated, which is being developed under a project of the Oporto University, called UPData, which aims, storage and preservation of scientific data.

The data preservation is a problem that all research institutions face, both nationally and internationally and, we intend to find a solution that would contribute to their resolution, thus allowing access to scientific data to a set of various research institutions.

In the final, we intend to make the evaluation of the systems implemented, in view of the requirements for the replication system and the interrogation system marks.

Conteúdo

CONTEÚDO

Lista de Figuras

LISTA DE FIGURAS

Capítulo 1

Introdução

1.1 Contexto e Enquadramento

Desde que surgiram as tecnologias digitais, a criação, manipulação e armazenamento de informação digital cresceu exponencialmente.

No contexto de investigação, os conjuntos de dados recolhidos são denominados de *datasets*. Eles possuem estruturas e informações bastantes diversificadas, devido ao facto de existir uma grande diversidade de áreas de investigação, podendo assumir a forma de registos textuais, imagens ou vídeos. Devido à grande diversidade existente, é necessário a existência de uma boa descrição dos *datasets*, como por exemplo os tipos de dados envolvidos e as condições de recolha e utilização.

A preservação e acessibilidade futura destes *datasets* é importante para a validação de resultados de investigações e como fonte para investigações futuras.

Atualmente, já existem repositótios de conteúdos digitais, que permitem o armazenamento e preservação de dados científicos, sendo o *software open source* DSpace, um bom exemplo disso. Contudo, existem alguns pormenores sobre os quais se deve ter especial atenção, como por exemplo, como garantir que não ocorram falhas de acessibilidade à informação armazenada. As falhas podem acontecer por vários motivos, como por exemplo uma quebra de energia eléctrica, ou a destruição do edifício onde se encontra o servidor, devido à ocorrência de um desastre natural.

Tendo em vista a possibilidade de ocorrência de tais acontecimentos, ou outros de gravidade semelhante ou pior, pretende-se a implementação de um sistema de replicação dos conjuntos de dados em diferentes locais de maneira a precaver as falhas de acesso. É também pretendido, a implementação de um sistema de interrogação, para que se possa consultar e extrair informação dos *datasets*, o mais intuitivamente possível.

Os sistemas desenvolvidos, serão testados num repositório de dados científicos, e como tal, criar-se-á a simulação de um repositório, que está a ser desenvolvido no âmbito de um projecto da Reitoria da Universidade do Porto, denominado de UPData, que

tem por objetivo, o armazenamento e preservação de dados científicos.

Sendo este um problema existente a nível nacional e internacional, pretende-se que esta solução possa ser replicada por várias instituições de investigação.

1.2 Motivação e Objetivos

Apesar do aumento da popularidade e preocupação com a preservação e partilha de dados científicos, este tema ainda é muito novo, e como tal ainda existem muitas áreas a serem exploradas. Este projeto visa explorar os temas de replicação e interrogação de dados científicos, para que sejam precavidas as falhas de acesso e o próprio acesso seja intuitivo.

Pretende-se assim, o desenvolvimento do modelo de replicação para a preservação e interrogação de dados científicos, de modo a que ele possa futuramente ser adaptado num caso real de um repositório de dados da Universidade do Porto.

1.3 Estrutura da Dissertação

Além do capítulo de introdução, esta dissertação é composta por mais 4 capítulos.

No Capítulo 2, é apresentada a definição de curadoria de dados, e são abordadas várias temáticas relacionadas. No Capítulo 3, são apresentados os dois sistemas a serem desenvolvidos, sistema de replicação e o sistema de interrogação, e as tecnologias a serem utilizadas. No Capítulo 4 é traçado um plano de execução para a proposta. Por último, no Capítulo 5 apresentam-se as conclusões.

Capítulo 2

Curadoria de Dados

O registo de diferentes observações, práticas e até mesmo experiências, tem ganho maior importância com o passar dos anos, sendo que em alguns casos particulares, como a ciência moderna, tornou-se um processo obrigatório e vital. Desde a origem das primeiras práticas de registo até à atualidade, verificou-se uma evolução dos mesmos, tanto na quantidade como na complexidade dos mesmos. Além disso, a própria forma de armazenar, preservar e partilhar os registos evoluiu com o passar dos anos, principalmente no âmbito de atividades científicas[RSRF10].

Nos meados do século XX, especialmente nas últimas duas décadas, verificou-se uma completa revolução na geração dos dados no âmbito de investigações científicas ao nível da sua dimensão, complexidade e importância. Pode afirmar-se que tal revolução foi fruto do crescimento tecnológico que se verificou nos ramos da informática e da comunicação. Atualmente, estas áreas disponibilizam meios que facilitam a criação, manipulação e armazenamento da informação digital, a uma escala nunca anteriormente vista. Apesar da preservação da informação ser de extrema importância, a preservação futura de toda a informação existente, é uma tarefa inviável, inútil e irrelevante. Mas, do mesmo modo que hoje podemos consultar informação oriunda de gerações passadas, principalmente dos últimos cinco séculos, é importante garantir que as gerações futuras tenham acesso a informação significativa e relevante da era atual[RSRF10, Fer06].

Em diversas áreas científicas, como por exemplo a genética, medicina, física ou meteorologia, as investigações estão dependentes do acesso a um grande volume de dados, que podem estar armazenados em bases de dados públicas ou privadas. Além disso, é fundamental que também seja possível recolher, recombinar e processar esses dados, sempre que surja essa necessidade, pois representam um investimento significativo e, em muitos casos, os dados existentes são insubstituíveis. Assim têm surgido várias pressões para a implementação de estratégias com o intuito de garantir a preservação e acesso

a longo termo. Este tipo de ciência já é designada de *data-intensive science*, que segundo os seus proponentes, tem três atividades essenciais: recolha, curadoria e análise de dados[RSRF10, HBH09, HTT09].

Por curadoria de dados, entenda-se a forma como os mesmos devem ser tratados, com o intuito de garantir a sua futura preservação. Além da importância do correto preenchimento da informação existente sobre os dados, denominados metadados, deve ter-se um cuidado especial com as ações que possam garantir a autenticidade, integridade e acessibilidade dos dados científicos. Resumidamente, a curadoria de dados envolve todas as atividades de preservação necessárias para garantir a possibilidade de os dados serem reutilizáveis no futuro[RSRF10].

2.1 Dados Científicos

Segundo a definição da *Organization for Economic Cooperation and Development* (OECD), dados científicos são "registos fatuais são usados como fontes primárias na investigação científica, e que são geralmente aceites na comunidade científica como necessários para validar os resultados de investigações"[1], que contêm estruturas e conteúdos bastante diversificados, devido à existência de uma grande diversidade de áreas de investigação, armazenados em conteúdos textuais, numéricos, imagens ou audiovisuais,

Além disso, os dados científicos também possuem uma dimensão bastante variável, que pode não ultrapassar algumas centenas de kilobytes, no caso de registos de observações individuais ou ensaios de pequenos laboratórios, ou noutro extremo, podem ser gerados várias dezenas de petabytes de dados científicos por dia, ocorrência normal no âmbito das atividades científicas no *Large Hadron Collider*[2] do CERN[RSRF10].

Não existe nenhuma perspetiva que pode ser considerada totalmente correta para a caracterização dos dados científicos. Por exemplo, a *National Science Foundation* (NSF)[3] dos Estados Unidos da América, baseia a caraterização dos dados científicos segundo a sua origem, obtendo-se os seguintes domínios:

- Dados de observação: compostos por registos históricos que não podem ser reproduzidos, e como tal, necessitam de preservação permanente. Neste contexto podemos referir os registos de atividades sísmicas;

- Dados computacionais: engloba os dados que são resultantes de simulações. Teoricamente estes dados podem ser reproduzidos se for preservada toda a informação sobre o seu modelo e a sua execução;

[1] Tradução da definição de "Research data"(pág.13) - *OECD Principles and Guidelines for Acess to Research Data from Public Funding*. Paris, 2007. Disponível em: http://www.oecd.org/dataoecd/9/61/38500813.pdf [consultado em: 5 de Janeiro de 2012]

[2] Mais informações sobre o *Large Hadron Collider* do CERN acessíveis em: http://public.web.cern.ch/public/en/LHC/LHC-en.html

[3] Site da *National Science Foundation* acessível em: http://www.nsf.gov/

- Dados experimentais: resultantes de experiências, e como tal, também não são facilmente reproduzíveis.

A comunidade científica, tem por norma, designar os conjuntos de dados científicos pelo nome de *Datasets*. Doravante, sempre que sejam referidos os dados científicos utilizar-se-á a designação científica, *dataset*.

Os *datasets* são de extrema importância para uso futuro. A validação de resultados obtidos em investigações baseadas nos mesmos e a reutilização dos datasets em investigações futuras, são os casos de utilização mais predominante, na atualidade. Desta forma, os *datasets* devem ser cuidadosamente preservados, para que, em casos de futura necessidade, possam ser corretamente acedidos e interpretados.

2.2 Preservação e Acessibilidade dos Dados

Nos últimos anos, generalizou-se o conceito e os movimentos a favor dos dados abertos ou *Open Data*, "um conteúdo ou informação é considerado aberto se, qualquer pessoa for livre de o utilizar, manipular e distribuir, sujeito unicamente à identificação da sua origem", definição segundo o projeto *Open Definition*[4] sobre a alçada da fundação *Open Knowledge Foundation*[5].

Os movimentos em favor de *datasets* abertos, defendem que os dados devem ser disponibilizados publicamente de forma gratuita, sem restrições de *copyright*, patentes ou outros mecanismos de controlo. Neste sentido, assemelham-se a outros movimentos de "abertura", tais como o *Open Source*[6] ou o *Open Access*[7], que, contudo, possuem dinâmicas e objetivos próprios.

Independentemente de qualquer movimento ou conceito existente, existe uma coisa sobre a qual todos os investigadores concordam, e perservação, acessibilidade e partilha de *datasets* é um processo que ainda tem que ser muito melhorado, para que se possa retirar o máximo proveito.

Quando se aborda o acesso a *datasets* preservados, é preciso ter em atenção alguns riscos e cuidados, gerados por questões de sigilo, confidencialidade, ou mesmo de direitos de autor. Apesar disso, a preservação e acessibilidade futura dos *datasets*, é de extrema importância, devido principalmente aos seguintes fatores[TR10]:

- Valor da informação: potencial valor dos dados em termos de reutilização, qualidade e importância nacional e/ou internacional, origem, tamanho, escala, custos associados com a sua geração ou o caráter inovador da investigação associada;

[4]Mais informações sobre o projeto *Open Definition* acessíveis em: http://opendefinition.org/
[5]Mais informações sobre a fundação *Open Knowledge Foundation* acessíveis em: http://okfn.org/
[6]Mais informações sobre o movimento *Open Source* acessíveis em: http://www.opensource.org/
[7]Mais informações sobre o movimento *Open Access* acessíveis em: http://www.eprints.org/openaccess/

- Informação única: os dados contêm informação de observações únicas, que não poderão ser reproduzidas novamente no futuro;

- Importância dos dados para a história, particularmente na história da ciência.

Como se pode verificar, a preservação de *datasets* é de extrema importância, contudo é preciso ter em atenção que nas duas últimas décadas, a quantidade de *datasets* existentes, deixou de ser uma raridade, passando a ter uma presença excessiva. Este crescimento explosivo na quantidade de *datasets* existentes, deveu-se principalmente à introdução de novas tecnologias no mercado informático e de comunicação, que facilitaram a criação, manipulação e o armazenamento de informação digitalmente. Sendo assim, tornou-se inevitável a existência de uma pré-seleção com o intuito de preservar unicamente os *datasets* relevantes.

Para que um determinado *dataset* seja um dos eleitos para o processo de preservação e acessibilidade futura, deve identificar-se com pelo menos um dos seguintes requisitos[TR10]:

- Reutilização: a reutilização dos *datasets* é o motivo mais comum e importante. Um caso de uso de reutilização, é a reanálise dos *datasets* sobre uma nova perspetiva de investigação, originada por avanços científicos. Outro caso de uso é a combinação ou recombinação de *datasets*, ou até mesmo comparação com antigos *datasets*, com o intuito de obter ou complementar informação. Esta reutilização, pode ser feita dentro ou fora do contexto de investigação no qual o *dataset* foi gerado, denominando-se como uso secundário;

- Verificação: este tipo de utilização, é quase sempre originado pela obrigação de cumprimento de códigos de conduta existentes para investigação, como por exemplo, o *Netherlands Code of Conduct for Scientific Practice*[8], que delibera que os *datasets* devem ser mantidos acessíveis para verificação, num determinado período de tempo. Pode-se afirmar que existe uma certa semelhança entre reutilização e verificação, visto que outros cientistas podem querer reanalisar *datasets* antigos, só com o intuito de verificar, ou até mesmo esclarecer dúvidas, sobre investigações baseadas nesses mesmos *datasets*;

- Património: para uso em investigação histórica, no caso particular da história da ciência, ou até mesmo, preservação com o único intuito de criar um património cultural.

Existem outros fatores, sobre os quais se deve também ter uma especial atenção quando se faz o processo de pré-seleção dos *datasets*: documentação, legalidade, infraestrutura e aspetos financeiros.

[8]Mais informações sobre *Netherlands Code of Conduct for Scientific Practice* acessíveis em: http://media.leidenuniv.nl/legacy/netherlands_code_of_conduct_for_scientific_practice.pdf

Relativamente a fatores documentais, entenda-se os metadados, que não são mais do que informações que permitem uma melhor preservação dos *datasets*. O principal intuito da existência de metadados, é o de descrever e documentar os dados, processos e atividades relacionadas com a geração dos *datasets*. Devem conter informação que descreva a sua origem (tempo ou espaço, métodos, instrumentos de recolha e transformações aplicadas), âmbito, autoria, propriedade e condições de recolha. Além disso, é importante que exista informação sobre os campos existentes nos registos, informação essa, que será composta por um descritivo de cada um dos campos e referência ao seu tipo de dado[TR10, Fer06, RSRF10].

De salientar, que quanto mais complexos forem os *datasets*, mais complexos terão que ser os detalhes das informações contidas nos metadados. A existência de metadados adequados e normalizados é um requisito essencial para o acesso e reutilização dos *datasets*, pois uma interpretação incorreta dos mesmos, poderá levar a enormes falhas nas investigações. Uma preservação de *datasets*, sem que exista a adequada documentação para cada um deles, é um completo desperdício de tempo e recursos [HBH09, RSRF10].

Os fatores legais e as limitações éticas também são de extrema importância, pois poderão existir alguns *datasets*, que por motivos de propriedade inteletual ou permissões, poderão estar restritos a um determinado grupo de pessoas, e como tal, não possam estar públicos para toda a comunidade [TR10, Fer06].

Por último, os fatores relacionados com as infraestruturas e os fatores financeiros, estão intimamente interligados. Neste caso, o problema de financiamento, mas além disso, outros requisitos são necessários: espaço físico, energia, especialistas e infraestruturas. Outro problema pertinente, é a mão de obra qualificada para fazer a gestão dos *datasets*, em estado de preservação atual ou futura. Além disso, devido à atual incerteza em volta das estratégias de preservação, ainda é impossível calcular os custos associados a longo prazo[TR10].

2.3 Prótotipo de Repositório de Dados Científicos

Atualmente, várias instituições de investigação espalhadas pelo mundo, começam a perceber o real valor da preservação digital, no contexto de dados científicos. Desde então têm surgido vários projetos neste âmbito: *Data Asset Framework*[9], *Edinburg DataShare*[10] ou *DANS Data Archive*[11] são bons exemplos desses esforços, com o intuito de assegurar futuramente uma melhor preservação digital de dados científicos.

A Universidade do Porto, tendo também notado este potencial, tem, neste momento, os seus serviços centrais em parceria com um grupo de investigação da Universidade

[9]Mais informações sobre o projeto *Data Asset Framework* acessíveis em: http://www.data-audit.eu/
[10]Mais informações sobre o projeto *Edinburg DataShare* acessíveis em: http://datashare.is.ed.ac.uk/
[11]Mais informações sobre o projeto *DANS Data Archive* acessíveis em: http://www.dans.knaw.nl/

do Porto, o desenvolvimento de um projeto denominado UPData, que tem por objetivo principal determinar as principais necessidades de curadoria de dados científicos, usando para caso de estudo os diferentes núcleos de investigação que a Universidade do Porto alberga[dSRL11, dSRL].

Neste momento, o UPData encontra-se a desenvolver um protótipo de um repositório de dados científicos, havendo uma grande proximidade entre a equipa de desenvolvimento e os investigadores dos diversos núcleos de investigação, de modo a que os resultados obtidos possam ser fiáveis e realistas[dSRL11, dSRL].

Com um variado leque de opções em repositórios *open-source*, a escolha da equipa de desenvolvimento responsável pelo projeto, acabou por recair sobre a plataforma de repositório DSpace. A preferência sobre o DSpace, entre outros motivos, deveu-se ao fato da proximidade que os investigadores da Universidade do Porto já têm com a utilização desta plataforma, devido à existência de dois repositórios operacionais nesta plataforma: o Repositório Aberto[12] e o Repositório Temático[dSRL11]. Para mais informações sobre a plataforma DSpace, consultar o Capítulo 3.3.1.

O formato digital escolhido, para a preservação dos dados científicos armazenados no repositório, foi o XML. A escolha do XML, deveu-se ao fato de o XML permitir bastante flexibilidade na representação das tabelas de dados, por mais diversificadas que elas sejam, e além disso, como os dados podem ser facilmente categorizados, permite a realização futura de consultas mais consistentes sobre os mesmos. No anexo A.1, encontra-se um exemplo de um XML gerado.[dSRL11].

2.4 Problemas

A preservação de informação, sempre foi um problema para a humanidade. Atualmente, o valor da informação tem um maior peso, mas desde à uns séculos atrás, que a humanidade tem vindo a ter o cuidado de tentar preservar o máximo de informação possível, para que futuras gerações possam ter acesso a esse conhecimento.

Apesar de todos os cuidados, acidentes e catástrofes acontecem, levando à perda de informação muito valiosa. Um dos momentos mais triste, na história da humanidade, no contexto de perda de informação, terá acontecido no ano de 646, quando a Biblioteca de Alexandria ardeu por completo, perdendo-se para sempre todo o conhecimento ali preservado[13].

Atualmente, a Universidade do Porto, como referido no capítulo anterior, Capítulo 2.3, encontra-se a desenvolver um protótipo de um repositório de dados científicos, com o

[12] Mais informações sobre o Repositório Aberto da Universidade do Porto acessíveis em: http://repositorio-aberto.up.pt/

[13] Mais informações sobre a Biblioteca de Alexandria acessíveis em: http://sdi.letras.up.pt/uploads/pdfs/alexandria3.pdf

intuito de garantir a preservação digital de *datasets*, mas isso, não impede que as ameaças externas não existam ou tenham um risco moderado, muito pelo contrário.

Já foram noticiadas, catástrofes que aconteceram um pouco por todo o mundo, desde inundações, terramotos, furacões, tornados, maremotos e até casos de acentuada queda de granizo, que provocaram sérios danos em universidades e institutos de investigação, levando à perda de grandes quantidades de informação. Basta recuar apenas anos e é possivel relembrar um vasto leque de exemplos: inundação na biblioteca da Universidade do Havaí em 2004, tsunami no Oceano Índico em 2004, furacão Katrina que em 2005 atingiu os Estados Unidos da América, inundações na biblioteca pública de IOWA nos Estados Unidos da América em 2008, terramoto no Haiti em 2010, chuva de granizo que atingiu a Universidade de Calgary no Canadá em 2011, entre outros casos.

É possivel concluir que não existe nenhum lugar no planeta, que possa garantir a 100% a proteção dos dados existentes. A Universidade do Porto não é uma exceção à regra, e também não consegue garantir que o servidor contendo o repositório de dados científicos, esteja totalmente isento de todo tipo de acidentes e catástrofes naturais que possam causar a perda de toda a informação armazenada e preservada, que se encontra nele.

Além disso, a acção humana também é susceptível a falhas e, involuntariamente, pode causar sérios danos no servidor, que contem o repositório de dados. Por outro lado, as componentes de *hardware* do servidor estão susceptíveis a uma enorme variedade de falhas, algumas delas que podem não permitir temporariamente o acesso aos dados preservados, mas noutros casos mais extremos, poderá significar a perda de todos os dados contidos no servidor.

Outro problema relativo à preservação de dados científicos, é relativo à perceção da informação contida nos *datasets*. A informação pretendida por um investigador poderá até existir num determinado *dataset*, mas não terá nenhum valor para o referido investigador, se o mesmo não tiver forma de questionar a sua existência. Por outro lado, existem casos onde o acesso a um determinado *dataset* é permitido, mas a morosidade do processo de interpretação e extração da informação contida no mesmo, torna o seu conteúdo menos valorizado.

Ambos os problemas são bastantes críticos para a correta preservação a longo prazo dos dados gerados em investigações científicas, e como tal devem ser abordados em qualquer processo de curadoria desta área.

Capítulo 3

Modelo de Replicação e Interrogação

Este projeto de dissertação, tem como principal intuito, a resolução dos problemas de acesso e perceção dos dados científicos preservados, anteriormente abordados na secção 2.4.

Pretende-se assim, a criação de um modelo de replicação, para a distribuição de réplicas de *datasets* em diferentes locais, com o objetivo de precaver as falhas de acesso. Para além disso, a criação de um modelo de interrogação, para facilitar a perceção dos *datasets* e da informação contida nos mesmos, por parte dos investigadores é outra das metas a atingir.

De realçar que o modelo de replicação é o objetivo prioritário deste projeto de dissertação, sendo o modelo de interrogação um objetivo secundário, mas não menos importante.

Nas secções seguintes, secção 3.1 e secção 3.2, existe uma descrição mais pormenorizada dos objetivos de cada um dos modelos, de replicação e de interrogação, respetivamente. Por último, na secção 3.3, é possível ver com mais detalhes as tecnologias que vão ser utilizadas na implementação do modelo de replicação.

Neste momento, ainda não se encontram definidas quais serão as tecnologias que vão ser utilizadas no modelo de interrogação, pois todos os esforços foram centrados no problema de replicação, que é o mais prioritário neste momento. As tecnologias a serem utilizadas para o modelo de interrogação serão definidas futuramente, após já se ter devidamente finalizado e testado o sistema de replicação.

3.1 Modelo de Replicação

O principal objetivo deste modelo é evitar que ocorram falhas de acesso aos *datasets*, falhas essas que podem ser temporárias ou definitivas, implementando-se para isso

um sistema de replicação de *datasets*, que faça a distribuição de réplicas por diferentes localizações geográficas.

A distribuição das réplicas de um determinado *dataset*, deve ser efetuada por diferentes localizações geográficas, o mais dispersas possível, havendo estudos, que aconselhável uma distância de 120 a 200 quilómetros entre as diferentes localizações. Deste modo, pretende-se assegurar que duas localizações geográficas contendo a mesma réplica, não sofram paralelamente danos provenientes do mesmo acontecimento, como por exemplo a ocorrência de uma catástrofe natural: furacão, terramoto, maremoto, entre outros cenários possíveis. Para além disso, com o intuito de redução do risco de ocorrência de danos devido a catástrofes naturais, a escolha das localizações, deve recair sobre zonas em que exista um baixo risco de ocorrência das mesmas[SS10, TW].

Outro fato a ter em atenção, é que o fornecimento de energia elétrica às diferentes localizações, deve ser proveniente de redes elétricas distintas, evitando-se assim falhas de acesso aos *datasets*, no caso de ocorrência de falhas no fornecimento da energia elétrica aos servidores. Além disso, o controlo e monitorização das réplicas, deve ser assegurado por uma administração própria, havendo uma administração por cada uma das localizações existentes[SS10, TW].

Por último deve ter-se em atenção, a integridade dos *datasets* preservados. Sendo que os *datasets* estão armazenados digitalmente, os mesmos estão sujeitos à ocorrência de eventos imprevistos, que podem provocar algum estrago nos *datasets* preservados digitalmente, e tal estrago pode não ser facilmente detetado externamente. Como tal, o sistema deverá constantemente comparar as réplicas preservadas nas diferentes localizações, de modo a detetar possíveis réplicas danificadas, e corrigi-las o mais rapidamente possível[SS10, TW].

Estudos realizados, aconselham que existam pelo menos 3 réplicas de cada *dataset*, de modo a que se possa assegurar a correta deteção e correção da réplica danificada. Por exemplo, se o resultado da comparação de duas réplicas supostamente idênticas não for positivo, existindo a terceira réplica, poder-se-á identificar qual das réplicas se encontra errada, e proceder-se à sua reparação. De realçar, que poderão ocorrer momentos em que três réplicas poderão não ser suficientes, e como tal, quanto maior for o número de réplicas existentes de um dado *dataset*, maior será a garantia de integridade[SS10, TW].

Para a implementação do sistema de replicação, vão ser testadas duas tecnologias de replicação existentes: LOCKSS e DuraCloud. A escolha da tecnologia final de replicação vai estar dependente da performance obtida em diferentes cenários de teste e nos custos associados à sua manutenção.

Pretende-se que o sistema de replicação possa ser adaptado a um repositório digital de dados científicos, e como tal, ambos os sistemas de replicação, irão ser testados numa simulação bastante básica do protótipo de repositório de dados científicos, que se encontra

atualmente a ser desenvolvido no âmbito de um projeto da Universidade do Porto, já referido anteriormente no Capítulo 2.3.

3.2 Modelo de Interrogação

Pretende-se que o modelo de interrogação resolva os problemas existentes na preceção da informação contida nos *datasets*, facilitando a consulta dos *datasets* existentes, e a extração da informação contida nos mesmos.

Com esse intuito, será desenvolvida uma interface, que permitirá aos investigadores a realização de consultas de forma intuitiva, que poderão ser realizadas sobre duas vertentes: sobre os metadados ou sobre os campos presentes num determinado *dataset*.

Pretende-se na consulta sobre os metadados, que os investigadores possam aceder de forma intuitiva aos *datasets* disponíveis, obtendo a listagem de todos, ou a listagem dos que cumprem as restrições impostas. As restrições impostas sobre os metados, podem ser sobre a individualidade ou a combinação dos seguintes campos: título do *dataset*; nome do investigador responsável pela geração do *dataset*; a data da última modificação ocorrida no *dataset*; direitos associados ao *dataset*; e sobre a descrição do *dataset*.

Um exemplo de uma consulta sobre os metadados, será por exemplo, perante um vasto conjunto de *datasets*, pedir a listagens dos *datasets*, gerados por um determinado investigador sobre o contexto de certo tema de investigação, ordenando a listagem de resultados pela data da última modificação.

A outra consulta, sobre a vertente dos campos presentes num determinado *dataset*, permitirá visualizar, e caso se pretenda, também extrair, os registos do dado *dataset*, que cumpram certas restrições impostas aos campos existentes. Por exemplo, supondo que se está a visualizar um *dataset* contendo os registos dos censos portugueses de toda a população, restringir a visualização dos registos, aos registos que correspondam aos residentes da cidade Porto e com menos de 25 anos de idade.

Também é pretendido que se possa exportar os registos presentes num determinado *dataset*, na sua totalidade ou parcialmente, para diferentes formatos digitais: CSV, EXCEL ou XML. De salientar, que o investigador, também terá a possibilidade de selecionar os campos sobre os quais quer obter a informação, tanto para a visualização como para a extração.

A grande vantagem da consulta ao nível dos campos de um *dataset*, é que permite que o investigador possa analisar um dado *dataset* de forma interativa, podendo assim já tirar conclusões, sobre a relevância do mesmo para sua investigação. Por outro lado, o investigador terá a possibilidade de extrair com bastante facilidade, os registos e os campos do *dataset* que considere relevantes, para a investigação em curso.

De modo a avaliar o sistema de interrogação implementado, será utilizado um método não empírico, denominado de avaliação heurística. Resumidamente, uma avaliação heurística, é um método de diagnóstico, no qual especialistas assumem o papel de utilizadores menos inexperientes, e examinam a interface de um dado sistema à procura de problemas, que caso encontrados, são classificados consoante o seu nível de gravidade. Os níveis de gravidade, estão classificados numa escala numérica de 0 a 4, inclusive, sendo que a cada nível, estarão associadas algumas tomadas de posição. A escolha deste método, recaiu sobre o fato de ser um método que pode ser utilizado tanto ao nível de estágio, protótipo ou mesmo após a implementação da interface, além de que, este método é classificado como sendo um método fiável, fácil, rápido e barato[Bra06].

3.3 Tecnologias e Ferramentas

Para que sejam atingidos os resultados esperados no desenvolvimento deste projeto de dissertação, vão ser utilizadas diversas tecnologias e ferramentas.

O sistema de interrogação e de replicação, já referidos no capítulos 3.2 e 3.1, respectivamente, serão implementados e testados num cenário de um repositório de conteúdos digitais. Sendo que se encontra em desenvolvimento um protótipo de um repositário de dados científicos, no âmbito de um projeto denominado UPData[dSRL], já anteriormente referido no Capítulo 2.3, vai-se fazer uma simulação o mais abstrata possível deste mesmo protótipo, para que deste modo, se possa implementar e testar nele, os diferentes sistemas desenvolvidos. A simulação do repositório de conteúdos digitais, será implementada na mesma tecnologia que o protótipo do UPData, ou seja, num repositório digital DSpace. Para mais informações sobre a tecnologia DSpace, consultar o Capítulo 3.3.1.

Além disso, um dos principais fatores que proporcionou a escolha da plataforma DSpace, e não uma das outras opções existentes no mercado, é que os investigadores da Universidade do Porto já estão familiarizados com a interface desta plataforma, pois já existe um repositório em funcionamento implementado em DSpace, denominado de Repositório Aberto da Universidade do Porto[1].

Para a implementação do sistema de replicação de *datasets*, serão testadas duas tecnologias de replicação existentes: LOCKSS e DuraCloud. Para mais informações sobre a tecnologia LOCKSS e DuraCloud, consultar o Capítulo 3.3.2 e 3.3.3, respetivamente.

Ainda não foram definidas as tecnologias a serem utilizadas no sistema de interrogação, sendo que as mesmas só serão definidas após o sistema de replicação se encontrar totalmente desenvolvido e testado. Isto deve-se ao fato de que o sistema de replicação é, por agora, a prioridade máxima deste projeto de dissertação.

[1] Mais informações sobre Repositório Aberto da Universidade do Porto acessíveis em: http://repositorio-aberto.up.pt/

3.3.1 DSpace

DSpace é um software *open-source*, que suporta o *Open Archive Initiative Protocol for Metadata Harvesting*[2] (OAI-PMH), e foi projetado de modo a suportar trocas de informações com outros repositórios de conteúdos digitais, implementados neste ou noutro repositório *open-source*. O DSpace também usa a norma *Dublin Core*[3], como formato de preservação dos metadados de todos os conteúdos existentes.

Foi desenvolvido numa parceria entre o MIT Libraries (MIT) e o Hewlett-Packard (HP), tendo sido lançado em 2002, com o intuito de disponibilizar sistemas de repositórios, para o armazenamento de documentos digitais, destinados à educação ou provenientes de investigações científicas[RSRF10, Sin07].

Em 2007, o MIT e a HP criaram a DSpace Foundation, uma organização sem fins lucrativos para promover a plataforma e suportar os seus utilizadores, sendo que em 2009, o suporte aos utilizadores, ficou encarregue de outra organização sem fins lucrativos, a DuraSpace Foundation[4][RSRF10].

Atualmente, estima-se que a plataforma DSpace contenha cerca de 1000 utilizadores, de todo o mundo, sendo alguns deles, instituições portuguesas de renome, como por exemplo: Universidade do Porto; Universidade de Lisboa; Universidade de Coimbra; e Universidade de Trás-os-Montes e Alto Douro[Orgb].

Uma das grandes vantagens do DSpace, é que este torna fácil o processo de criação de repositórios institucionais, que permitem a recolha, partilha e preservação digital de conteúdos intelectuais. A plataforma permite também o armazenamento de uma grande variedade de formatos digitais, como por exemplo: artigos, *datasets*, imagens, ficheiros de áudio, ficheiros de vídeo, programas de computador, entre outros. O DSpace também fornece um vasto conjunto de ferramentas, para ajudar as instituições na gestão dos seus conteúdos digitais. Além disso, em caso de um formato digital preservado tornar-se obsoleto, é possível que esse formato seja migrado para um novo formato existente[WBT+05, Pru05].

Outra grande vantagem da utilização de um repositório DSpace, é a enorme capacidade de personalização do mesmo às necessidades das organizações, principalmente nos casos de organizações grandes e complexas, que fazem a submissão de uma grande quantidade de conteúdos digitais, provenientes dos mais diversificados departamentos[Orgb].

Segundo a figura 3.1, retirada de [Orgb], a hierarquia dos conteúdos digitais num repositório DSpace, encontra-se retratada da seguinte maneira[dSRL, Orgb]:

[2]Mais informações sobre *Open Archive Initiative Protocol for Metadata Harvesting* acessíveis em: http://www.openarchives.org/

[3]Mais informações sobre *Dublin Core* acessíveis em: http://dublincore.org/

[4]Mais informações sobre a organização DuraSpace Foundation acessíveis em: http://www.duraspace.org/

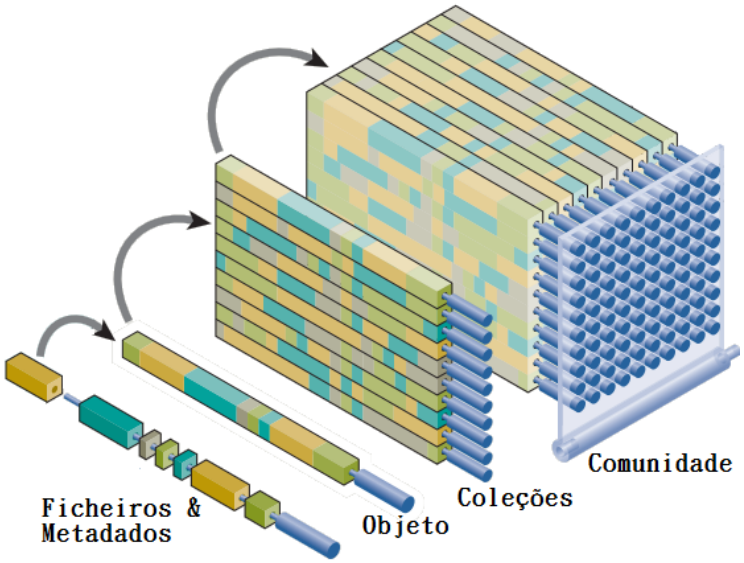

Figura 3.1: Hierarquia dos conteúdos digitais em DSpace

- Comunidade: é o nível mais alto da hierarquia. Representa as entidades de uma organização, como por exemplo: departamentos, laboratórios, centros de investigação ou instituições de ensino;
- Coleção: faz parte de uma determinada comunidade, sendo que cada uma delas é composta por um ou mais objetos. Cada colecção, pode ser por exemplo a representação de um tema, atividade ou projeto;
- Objeto: faz parte de uma determinada coleção. Cada objeto tem alguma informação associada, para uma melhor identificação do mesmo e do contexto onde se encontra inserido. Essa informação associada, denominada de metadados, encontra-se indexada, para que seja possível fazer-se consultas e pesquisas sobre o objeto. Cada um dos objetos existentes é composto por um ou mais ficheiros;
- Ficheiros & Metadados: são os arquivos digitais, dos mais diversificados formatos, que se encontram preservados na plataforma, e os seus respetivos metadados.

Os utilizadores só podem ter acesso aos objetos de determinada coleção, se estiverem nas lista de membros dessa mesma coleção. Um utilizador pode estar envolvido em uma ou mais coleções, sendo que as permissões podem variar nas várias coleções onde está inserido. Pode-se afirmar que existem três classes gerais de utilizadores[Orgb]:

- Utilizador final: pode consultar e aceder aos arquivos digitais contidos nas coleções, caso tenha permissões para tal;

- Curador: responsável pelos processos de curadoria dos arquivos digitais armazenados, de modo a garantir a sua preservação e acessibilidade futura;
- Submissor: utilizador que tem permissões para a inserção de conteúdos digitais numa determinada coleção.

Resumidamente, a plataforma DSpace, é uma excelente ferramenta *open-source* para a submissão de conteúdos digitais, pois pode ser facilmente adaptada às necessidades existentes, e dispõe de um vasto conjunto de ferramentas para a gestão da comunidade e dos conteúdos digitais.

3.3.2 LOCKSS

LOCKSS é uma sigla que é derivada de *Lots of Copies Keep Stuff Safe*, ou seja, quantas mais cópias existirem mais segura a informação estará.

O programa LOCKSS, foi fundado em 1998, pela Universidade de Stanford. Os primeiros testes foram iniciados em 1999, sendo que entre 2000 e 2002 foi lançada uma versão beta. Uma nova versão foi entretanto desenvolvida em 2002, com o intuito de refazer a arquitectura do *software*, tendo os testes sido iniciados nos finais de 2002. No período decorrido entre 2002 até meio de 2004, foram levantadas e debatidas várias questões sobre a manipulação de coleções. Por fim, com base nos resultados expostos numa publicação de uma pesquisa premiada da ACM[MRR+03], o sistema foi relançado para produção[Pro].

A equipa de engenharia responsável pela criação do LOCKSS, aquando do seu desenvolvimento, teve em atenção a criação de um sistema que prevenisse a ocorrência de uma vasta gama de falhas que poderiam ocorrer a três níveis: *hardware*, económicas ou sociais. Na totalidade foram consideradas 13 ameaças existentes, sendo algumas delas, por exemplo: falhas de *hardware*, falhas de serviços de rede, suporte e *hardware* obsoleto ou desastre natural. É importante que, para cada umas das ameaças contabilizadas seja realizada uma análise custo/benefício[RR09a].

A tecnologia LOCKSS disponibiliza livrarias com ferramentas e suporte para preservação digital, de modo a que se possa facilmente e sem grandes custos recolher e preservar conteúdos digitais disponíveis *online*. LOCKSS é uma tecnologia *open-source*, *peer-to-peer*, com uma infra-estrutura descentralizada[RR09a, Ros10, Pro].

Inicialmente, a tecnologia de replicação LOCKSS, foi desenvolvida para a replicação de jornais electrónicos, sendo que neste momento, é responsável pela replicação global de cerca de 8600 jornais electrónicos. Porém, a comunidade LOCKSS cresceu bastante, tendo atualmente uma vasta comunidade internacional, trabalhando conjuntamente com o objetivo comum de preservação a longo prazo de conteúdo *web*[Ros10, Pro].

Um diagrama geral do funcionamento do LOCKSS, pode ser observado na figura 3.2, figura essa proveniente de [Pro]. Na imagem, temos menção a um objeto denominado

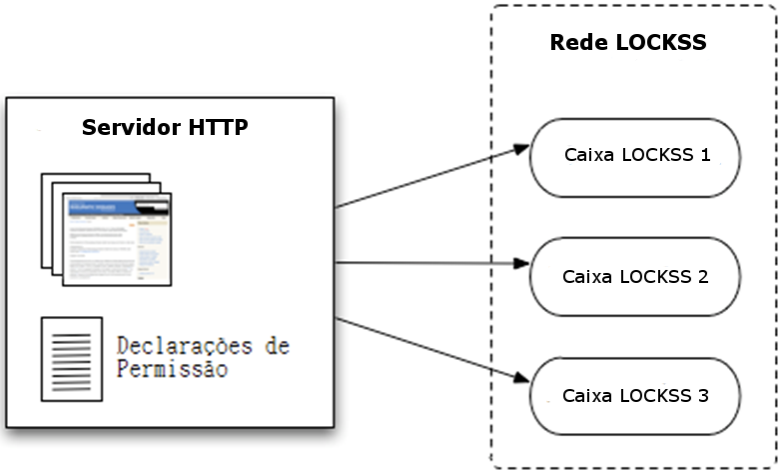

Figura 3.2: Diagrama de funcionamento LOCKSS

caixa LOCKSS, que é basicamente um computador normal, onde foi instalado o *software* LOCKSS, e que inserido numa rede partilhada, permite a partilha de réplicas de conteúdos digitais.

As caixas LOCKSS recolhem conteúdos digitais de páginas *web*, usando para isso um *web crawler*, similar com os que são usados nos motores de busca. De modo a que uma caixa LOCKSS saiba qual o conteúdo digital que deve preservar, o administrador da rede LOCKSS, deverá indicar onde se encontram os conteúdos digitais a serem preservados. Por outro lado, os detentores desses conteúdos, deverão fazer a inserção de algumas declarações de permissão, nesses mesmos conteúdos, de modo a que as caixas LOCKSS, os identifiquem como válidos para preservação. Os detentores dos conteúdos listados para preservação, podem também inserir e manipular alguns parâmetros nas declarações de permissão, com o intuito de identificar a que nível de profundidade a preservação deve ser efetuada[RR09a, Pro].

Uma das grandes vantagens de utilização da tecnologia LOCKSS para a preservação de conteúdos digitais, é devido ao fato de ela conter um excelente sistema de verificação da integridade das réplicas preservadas. As caixas LOCKSS que compõem a rede, comunicam constantemente entre si, com o intuito de localizar réplicas comuns, para que assim possam compara-las, usando para isso um protocolo *anti-entropy peer-to-peer*, baseado num sistema de voto. Deste modo, as caixas LOCKSS através de comparações, conseguem detetar réplicas danificadas, sendo que a caixa LOCKSS contendo a réplica danificada, fará um pedido de reparação a outra caixa LOCKSS, para que a réplica correta possa ser enviada e assim o erro possa ser reparado[RR09a, Ros10, Pro].

Na figura 3.3, encontra-se um exemplo do processo de verificação da integridade das réplicas de um dado conteúdo digital, numa rede composta por 4 caixas LOCKSS. Como se pode verificar, no lado esquerdo da imagem, as caixas LOCKSS estão a comunicar

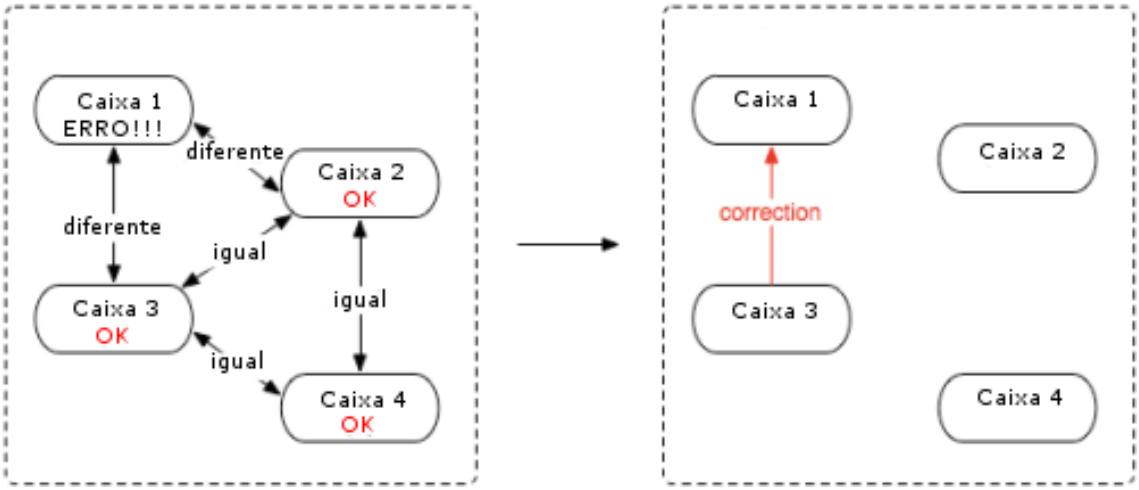

Figura 3.3: Verificação de Integridade - LOCKSS

entre si para verificação da integridade das replicas preservadas. As caixas LOCKSS 2, 3 e 4 afirmam que as replicas preservadas coincidem. Porém, a caixa LOCKSS 1, ao comparar a sua réplica com as caixas LOCKSS 2 e 3, denota que sua réplica não coincide, e visto que o consenso da votação, afirma que as replicas presentes nas caixas LOCKSS 2 e 3, são as que se encontram corretas, então pode-se afirmar que a réplica presente na caixa LOCKSS 1, encontra-se danificada. Como tal, sendo que a caixa LOCKSS 3 contém a réplica correta, a caixa LOCKSS 1 faz um pedido de reparação à mesma, e a caixa LOCKSS 3 atende positivamente ao pedido, fazendo o envio da réplica correta, como se pode verificar pelo lado direito da imagem.

Apesar de o exemplo acima descrito, só conter na sua totalidade 4 caixas de LOCKSS, é aconselhável que uma rede LOCKSS tenha pelo menos 7 caixas LOCKSS, com o intuito de assegurar a correta preservação digital[RR09b].

Outra vantagem do uso da tecnologia LOCKSS, é nos casos em que o conteúdo digital original não está disponível para o utilizador final, e como tal, havendo uma réplica desse conteúdo noutro lugar, é possível carregá-la e mostrá-la ao utilizador, sem que o mesmo se aperceba da diferença[Pro].

De realçar que todo o conteúdo digital é preservado e replicado no seu formato original, mas caso o formato se torne obsoleto, o conteúdo pode ser migrado diretamente do formato original para o corrente, minimizando assim os efeitos de conversão entre formatos. Além disso, todo o conteúdo preservado é migrado para a mais recente, e provavelmente melhor tecnologia disponível no momento em que o pedido é efetuado[Pro].

Por último, toda a gestão dos conteúdos replicados é feita por através de uma interface, acessível apenas aos administradores da rede LOCKSS, ou a utilizadores a quem lhes foi atríbuido acesso.

3.3.2.1 Private LOCKSS Network (PLN)

Atualmente a tecnologia LOCKSS está a ser utilizado na preservação de conteúdo digital em dois tipos distintos de ambientes: redes públicas LOCKSS e algumas redes privadas LOCKSS. As redes privadas LOCKSS, são as denominadas PLN, *Private LOCKSS Network*, ou também conhecidas pelo nome de *CLOCKSS*[RR09a].

Enquanto as redes públicas LOCKSS preservam conteúdos digitais de geral interesse para uma larga comunidade, as PLN são mais direcionadas na preservação de conteúdos digitais específicos no contexto de certas comunidades. Através deste esforço de preservação mais concentrado em determinados conteúdos, as PLN oferecem às organizações uma melhor cooperação para garantir a preservação dos conteúdos que lhes interessam verdadeiramente[RR09a, RR09b].

As PLN normalmente são compostas por 7 a 15 organizações, que possuem algum ponto de interesse em comum. Cada uma das PLN é responsável pela administração técnica da infra-estrutura. Além disso, cada uma das PLN estabelece as suas próprias políticas e práticas: governo; financiamento; desenvolvimento da coleção de objetos; e acessos[RR09a, RR09b].

3.3.3 DuraCloud

A organização responsável pelo serviço DuraCloud, é a DuraSpace Foundation[5], a mesma organização responsável pela plataforma DSpace, descrita anteriormente no Capítulo 3.3.1.

O serviço DuraCloud facilita o processo de armazenamento de conteúdos digitais em serviços de *cloud*, fazendo assim a replicação dos conteúdos por diferentes locais. Usando este serviço, é facil mover as réplicas dos conteúdos digitais entre os diferentes serviços de *cloud* existentes. Além disso, disponibiliza ferramentas intuitivas, para uma melhor controlo dos conteúdos preservados na *cloud*[Orga].

Este serviço tem como principal vantagem, a capacidade de abstrair os detalhes das diferentes APIs dos serviços *cloud*, através da disponibilização de uma API, que padroniza as chamadas aos diferentes serviços de *cloud*, não sendo assim preciso ter conhecimento de todas as API's dos serviços de *cloud* existentes[Orga]. Seguem alguns exemplos de serviços *cloud*: Amazon S3[6]; Rackspace[7]; e Windows Azure[8].

Sendo o serviço DuraCloud ainda bastante recente no mercado, ainda não existe nenhuma lista de clientes publicada online. Por outro lado, este serviço aposta numa continua evolução, com o intuito de melhorar os serviços de preservação[Orga].

[5]Mais informações sobre a organização DuraSpace Foundation acessíveis em: http://www.duraspace.org/
[6]Mais informações sobre Amazon S3 acessíveis em: http://aws.amazon.com/pt/s3/
[7]Mais informações sobre Rackspace acessíveis em: http://www.rackspace.com/
[8]Mais informações sobre Windows Azure acessíveis em: http://www.windowsazure.com/pt-br/

Capítulo 4

Plano de Trabalho

Neste capítulo expõe-se o planeamento previsto para o desenvolvimento deste projeto de dissertação.

O plano de trabalho será realizado de uma forma iterativa e incremental, implementando um componente funcional de cada vez antes de passar para o seguinte. Em caso de ocorrência de imprevistos, dar-se-á prioridade a funcionalidades que permitam maximizar a obtenção de resultados para posterior.

4.1 Tarefas

As tarefas definidas para o desenvolvimento deste projeto de dissertaçao, foram:

- Instalação das tecnologias e testes: inclui inicialmente a instalação do DSpace, LOCKSS e DuraCloud. É também previsto que nesta fase sejam feitos alguns pequenos testes, para tentar perceber as restrições impostas. Por último, fazer os pedidos de recuros ao CICA, caso seja necessário;
- Implementação de sistema de replicação: implementação do sistema de replicação de *datasets*;
- Estudo das tecnologias para o sistema de interrogação: levantamento do estado da arte para as tecnologias necessárias para o desenvolvimento do sistema de interrogação;
- Implementação de sistema de interrogação;
- Avaliação dos sistemas: avaliação dos dois sistemas implementados;
- Escrita de artigo científico;

- Escrita da dissertação;

4.2 Diagrama de Gantt

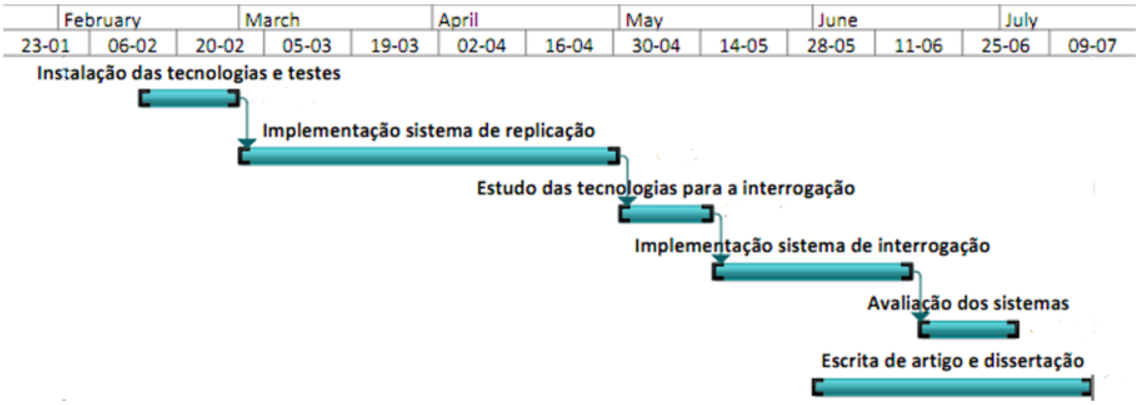

Figura 4.1: Diagrama de gantt do plano de trabalho

Na figura 4.1, pode-se observar o digrama de gantt do plano de trabalho.

Como se pode observar a tese será iniciada a 14 de Fevereiro, com a instalação das tecnologias e execução de alguns testes sobre as mesmas, sendo que é previsto que este tarefa esteja finaliza no final de Fevereiro.

Do início de Março até ao inicio de Maio, será implementado o sistema de replicação.

Posteriormente, haverão duas semanas para o levantamento do estado da arte para o desenvolvimento do sistema de interrogação, iniciando-se logo de seguida o desenvolvimento do sistema de interrogação, durante um período planeado de 1 mês.

No fim do sistema de interrogação estiver terminado, haverão duas semanas para que se possa fazer a avaliação dos sistemas desenvolvidos.

Pretende-se que a escrita para o artigo científico e para a dissertação, sejam iniciados no início de Junho e se prolonguem até ao prazo final de entrega.

Capítulo 5

Expetativas Futuras

O tema deste projeto de dissertação é bastante atual e aborda as necessidades de diversos grupos de investigação, nacionais e internacionais, tanto a nível da preservação de *datasets* como da própria acessibilidade aos mesmos.

Pretende-se que o sistema de replicação desenvolvido, consiga fazer a replicação dos *datasets* por diferentes localizações, evitando-se que aconteçam falhas de acesso aos dados. Por outro lado, é pretendido que este sistema de replicação, não traga custos extra muito avultados às organizações envolvidas.

Outro objetivo é que o sistema de interrogação consiga cumprir as necessidades dos investigadores, facilitando assim a sua incorporação e utilização no contexto de projetos de investigação.

Quem sabe, se num futuro próximo, não é criada uma rede nacional para a preservação e acessibilidade de *datasets*, com a participação de todos os institutos de investigação a nível nacional.

Referências

[Bra06] Eduardo Rangel Brandão. Publicidade on-line, ergonomia e usabilidade: o efeito de seis tipos de banner no processo humano de visualização do formato do único na tela do computador e de lembrança da sua mensagem. Master's thesis, Pontifícia Universidade Católica do Rio de Janeiro, Março 2006.

[dSRL] João Rocha da Silva, Cristina Ribeiro e João Correia Lopes. Updata - scientific data curation at u.porto. http://sciencedata.up.pt/doc/.

[dSRL11] João Rocha da Silva, Cristina Ribeiro e João Correia Lopes. Updata - a data curation experiment at u.porto using dspace. Technical report, Faculdade de Engenharia da Universidade do Porto, Novembro 2011.

[Fer06] Miguel Ferreira. *Introdução à preservação digital : conceitos, estratégias e actuais consensos*. Universidade do Minho, Escola de Engenharia, 2006.

[HBH09] Mark Hedges, Tobias Blanke e Adil Hasan. Rule-based curation and preservation of data: A data grid approach using irods. *Future Generation Comp. Syst.*, 25(4):446–452, 2009.

[HTT09] Tony Hey, Stewart Tansley e Kristin Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington, 2009.

[MRR+03] Petros Maniatis, David S. H. Rosenthal, Mema Roussopoulos, Mary Baker, TJ Giuli e Yanto Muliadi. Preserving peer replicas by rate-limited sampled voting. *SIGOPS Oper. Syst. Rev.*, 37:44–59, October 2003.

[Orga] DuraCloud Organization. Duracloud - official web site. http://www.duracloud.org/.

[Orgb] DuraSpace Organization. Dspace - official web site. http://www.dspace.org/.

[Pro] LOCKSS Program. Lockss - official web site. http://www.lockss.org/lockss/Home.

[Pru05] Marion Prudlo. E-Archiving: An Overview of Some Repository Management Software Tools. *Ariadne*, (43), April 2005.

[Ros10] David S. H. Rosenthal. Lockss: Lots of copies keep stuff safe, March 2010.

[RR09a] Victoria Reich e David S. H. Rosenthal. Distributed digital preservation: Lots of copies keep stuff safe. In *Indo-US Workshop on International Trends in Digital Preservation*, pages 51–55, March 2009.

[RR09b] Victoria Reich e David S.H. Rosenthal. Distributed digital preservation: Private lockss networks as business, social, and technical frameworks. *Library Trends*, 57(3), 2009.

[RSRF10] Eloy Rodrigues, Ricardo Saraiva, Cristina Ribeiro e Eugénia Matos Fernandes. Os repositórios de dados científicos: Estado da arte. Technical report, Universidade do Minho e Universidade do Porto, Julho 2010.

[Sin07] Neha Singh. Solution for e-journal archiving and framework for evaluation of archiving software. February 2007.

[SS10] Katherine Skinner e Matt Schultz. *A Guide to Distributed Digital Preservation*. Educopia Institute, Atlanta, 2010.

[TR10] Heiko Tjalsma e Jeroen Rombouts. Selection of Research Data. Guidelines for appraising and selecting research data. A report by DANS and 3TU.Datacentrum, July 2010.

[TW] Aaron Trehub e Andrew Waller. Private lockss networks: overview and working examples.

[WBT+05] I. Witten, D. Bainbridge, R. Tansley, C. Y. Huang e K. Don. Stoned: bridge between greenstone and dspace. *Library Hi Tech*, 11(9), September 2005.

Anexo A

Anexos do DSpace

A.1 Exemplo de um Dataset convertido em XML

```
<?xml version="1.0"?>
<record>
  <metadata>
    <dc.creator>Bastos, Lu&amp;#237;sa; Deurloo, Richard</dc.creator>
    <dc.title>Aereal Gravimetry Run (GPS Processed Data for Terceira Island –
        Beach) Sensor – tail of airplane</dc.title>
    <dc.type>Numerical Data</dc.type>
    <dc.rights>open access</dc.rights>
    <dc.date.issued>1992.0</dc.date.issued>
    <dc.description>Processed GPS coordinates for the airplane, for the
        Terceira Island (Beach)</dc.description>
  </metadata>
  <headers>
    <header>grav.gpstime</header>
    <header>grav.latitude</header>
    <header>grav.longitude</header>
    <header>grav.height</header>
  </headers>
  <data>
    <rows>
      <row>
        <grav.gpstime>488496.999194</grav.gpstime>
        <grav.latitude>38.760267507</grav.latitude>
        <grav.longitude>-27.08411373</grav.longitude>
        <grav.height>112.989</grav.height>
      </row>
      <row>
        <grav.gpstime>488497.999193</grav.gpstime>
        <grav.latitude>38.760267485</grav.latitude>
        <grav.longitude>-27.084113744</grav.longitude>
        <grav.height>112.995</grav.height>
      </row>
    </rows>
  </data>
</record>
```