Universidade do Minho

Hélder de Jesus Almeida da Silva

Serviço Centralizado de Estatísticas de Utilização de Repositórios

Tese de Mestrado
Mestrado em Engenharia Informática
Trabalho efetuado sob a orientação de
Doutor José Carlos Leite Ramalho

Outubro 2011

Agradecimentos

Ainda que esta dissertação tenha sido de forma principal um esforço pessoal e individual, são muitas as pessoas que de forma direta e indireta contribuíram para o seu sucesso, pelo apoio, pelo incentivo, pelas vezes que me "dispensaram" de outros afazeres, enfim, por estarem simplesmente comigo nesta demanda.

Começo assim pelos que me são mais chegados. Um agradecimento aos meus pais e avó, por terem sabido dar-me apoio e o "incentivo" necessário na altura correta da minha vida, contribuindo assim para que o caminho percorrido tenha sido correto, i.e., terminar a dissertação com sucesso. Até ofereceram a sua ajuda, ainda que em tom de brincadeira e sem grande conhecimento de informática, mas mesmo assim conseguiram ajudar bastante. Um agradecimento também à minha namorada Patrícia, que nestes seis anos de namoro sempre me apoiou e incentivou a conseguir o exito acadêmico, estando sempre presente nos bons e menos bons momentos desta caminhada.

Agradeço aos Serviços de Documentação da Universidade do Minho, na pessoa do Doutor Eloy Rodrigues e do José Carvalho, pelos conselhos e pela execução mais técnica do projeto. Agradeço também à Fundação para a Computação Científica Nacional, na pessoa do João Moreira, pelas longas conversas e discussões no contexto das temáticas abordadas nesta dissertação, de onde sairam sempre boas ideias e que foram contributo para o sucesso da dissertação e projeto desenvolvido.

Last but not least, quero agradecer à KEEP SOLUTIONS, por me ter sempre apoiado, mas de forma especial ao Professor Doutor José Carlos Ramalho e ao Doutor Miguel Ferreira, pelo seu conhecimento, por me terem orientado, "aturado" e por terem conseguido que este esforço se traduzisse num projeto e dissertação que são muito mais que isso, são a realização e o sucesso acadêmico e profissional.

Em suma, e usando palavras de Fernando Pessoa,

Um homem de génio é produzido por um conjunto complexo de circunstâncias, começando pelas hereditárias, passando pelas do ambiente e acabando em episódios mínimos de sorte.

Resumo

Nos dias de hoje presenciamos a uma mudança de paradigma no que diz respeito aos repositórios institucionais. Passamos do repositório isolado no seu contexto institucional para os consórcios onde conjuntos de repositórios partilham ideias, políticas e tecnologias, contribuindo para o crescimento do conhecimento das comunidades em que estão inseridos.

Podia falar-se da importância que os repositórios institucionais mantêm junto da sua comunidade, mas com a mudança de paradigma torna-se relevante explorar os desafios de gestão que o novo contexto apresenta. Para gerir um consórcio é necessário possuir indicadores que auxiliem na tomada de decisão e que atendam às necessidades de informação que as entidades de fomento possuem no que diz respeito ao impacto dos investimentos em cultura, investigação, inovação e desenvolvimento.

Esta dissertação apresenta o Serviço Centralizado de Estatísticas de Utilização de Repositórios (SCEUR). Trata-se de um projeto inserido no âmbito da iniciativa Repositório Científico de Acesso Aberto de Portugal (RCAAP) que visa a construção de uma arquitetura que permita recolher, processar e apresentar de uma forma intuitiva dados estatísticos de utilização em repositórios institucionais e também auxiliar a partilha de dados estatísticos quer pela disponibilização de *add-ons* que facilitem essa tarefa no *software* usado pelas instituições quer pela disponibilização de informação e recomendações variadas nesse contexto. São também apresentados projetos internacionais de referência neste contexto, normas existentes e tecnologias usadas para a implementação dos conceitos subjacentes.

Abstract

Nowadays we witness a paradigm shift regarding to institutional repositories. Repositories are no longer isolated in their institutional context. Instead they participate in consortia where ideas, policies and technologies can be shared, contributing to the growth of knowledge of the communities where they are included.

Much could be said about the importance that institutional repositories do maintain in their community, but with this new paradigm it becomes relevant to explore the management challenges that the new context presents. To manage a consortium is necessary to have indicators that help in decision making and meet the needs of information that the promoting entities have with regard to the impact of investments in culture, research, innovation and development.

This dissertation presents SCEUR. It's a project (developed under the RCAAP initiative) which aims to build an architecture that allows to collect, process and present in an intuitive manner statistics from institutional repositories usage data and also assist in the sharing of statistical data by developing add-ons that facilitate that task in software used by the institutions and by publishing information and various recommendations in this context. International projects are also presented, as well as existing standards and technologies used to implement the underlying concepts.

Conteúdo

Lista de Figuras

Listagens

Lista de Tabelas

Glossário

add-on
: Componente de software que visa adicionar novas funcionalidades a um determinado software.

dc
: Dublin Core. Apelido dado ao DCMES, também conhecido como DC Simples, desenvolvido pela DCMI [Baptista, 2010].

DCMES
: Dublin Core Metadata Element Set. Conjunto de elementos de metadados desenvolvido, mantido e recomendado pela DCMI [Baptista, 2010].

DCMI
: Dublin Core Metadata Initiative. Organização para o desenvolvimento de standards, a nível de metadados, para a interoperabilidade de sistemas [Baptista, 2010].

handle
: Identificador persistente de um *handle system*.

handle system
: Sistema que permite atribuir, gerir e resolver identificadores persistentes para objetos digitais [Kahn and Wilensky, 2006].

interoperabilidade
: Capacidade de tipos diferentes de computadores, redes, sistemas operativos e aplicações trabalharem em conjunto com eficácia, sem comunicação prévia, de forma a trocarem informação de uma maneira útil e com significado [Baptista, 2010].

log
: Resultado da tarefa de *logging*.

logging
Tarefa de registar informação útil da execução de um determinado programa ou processo, que pode ser usada para debug, *error checking* entre outros.

metadados
Dados sobre os dados [associação para a promoção da sociedade da informação, 2007] ou informação sobre recursos [Baptista, 2010].

repositório institucional
Software usado por uma instituição para armazenar, preservar e divulgar a sua produção intelectual.

software
Um ou mais programas de computador que permitem executar tarefas específicas.

workflow
Sequência de passos a executar com o objetivo de realizar uma tarefa e/ou missão. [Wikipedia, 2011].

XML Schema
Linguagem baseada em XML usada para definir a estrutura de documentos XML, que habitualmente também é usada para fazer a validação dos mesmos.

Acrónimos

API	Application Programming Interface.
COUNTER	Counting Online Usage of NetWorked Electronic Resources.
CSV	Comma Separated Values.
DEFF	Denmark's Electronic Research Library.
DRIVER	Digital Repository Infrastructure Vision for European Research.
FCCN	Fundação para a Ciência e Computação Nacional.
FECYT	Fundación Española para la Ciencia Y la Tecnología.
HTML	HyperText Markup Language.
HTTP	HyperText Transfer Protocol.
IP	Internet Protocol.
ISI	Institute of Scientific Information.
JISC	Joint Information Systems Committee.
JSON	JavaScript Object Notation.
KEV	Key Encoded Values.
MESUR	MEtrics from Scholarly Usage of Resources.
METS	Metadata Encoding and Transmission Standard.
MIT	Massachusetts Institute of Technology.
NEEO	Network of European Economists Online.
NISO	National Information Standards Organization.

OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting.
OpenDOAR	The Directory of Open Access Repositories.
RCAAP	Repositório Científico de Acesso Aberto de Portugal.
REST	REpresentational State Transfer.
SARI	Serviço de Alojamento de Repositórios Institucionais.
SCEUR	Serviço Centralizado de Estatísticas de Utilização de Repositórios.
SOAP	Simple Object Access Protocol.
SUSHI	Standardized Usage Statistics Harvesting Initiative.
UM	Universidade do Minho.
URL	Uniform Resource Locator.
XML	eXtensible Markup Language.

Capítulo 1

Introdução

As recentes políticas da Comissão Europeia [COMMUNITIES, 2007] no que diz respeito à adoção e promoção do acesso-livre à informação científica conduziram a um aumento significativo do número de repositórios institucionais em Portugal.

Se há 8 anos atrás, os promotores do movimento Acesso-Livre ao Conhecimento se resumiam a um conjunto limitado de pessoas e instituições, a realidade atual é bem distinta. Hoje, existem repositórios institucionais com elevada relevância em instituições tão distintas como hospitais, centros de investigação, institutos, laboratórios ou escolas secundárias. Segundo o diretório do site RCAAP[1], existem 33 repositórios nacionais indexados, sendo que 17 são de universidades (51,5%), 8 de institutos (24,2%) e 4 de hospitais (12,1%).

Na Figura 1.1 podemos ver o gráfico da evolução do número de repositórios portugueses, de acordo com o sítio *Web* The Directory of Open Access Repositories (OpenDOAR)[2], respeitante ao período de tempo de 2006 até ao presente. O OpenDOAR é um diretório de repositórios académicos em acesso-livre, onde a informação registada de cada repositório é verificada de forma manual por um membro do projeto, com o objetivo de a validar e assim garantir informação fidedigna.

Com a proliferação de repositórios, verificou-se também uma elevada dispersão de informação, com a inerente dificuldade na sua localização, bem como a adopção de diferentes modelos de gestão local. Este facto dificulta a implementação de quaisquer iniciativas de consolidação e análise da informação veiculada e dos indicadores de sucesso subjacentes à gestão do próprio repositório.

Com o objetivo de mitigar este problema, assiste-se atualmente à cria-

[1] http://www.rcaap.pt/directory.jsp
[2] http://www.opendoar.org/

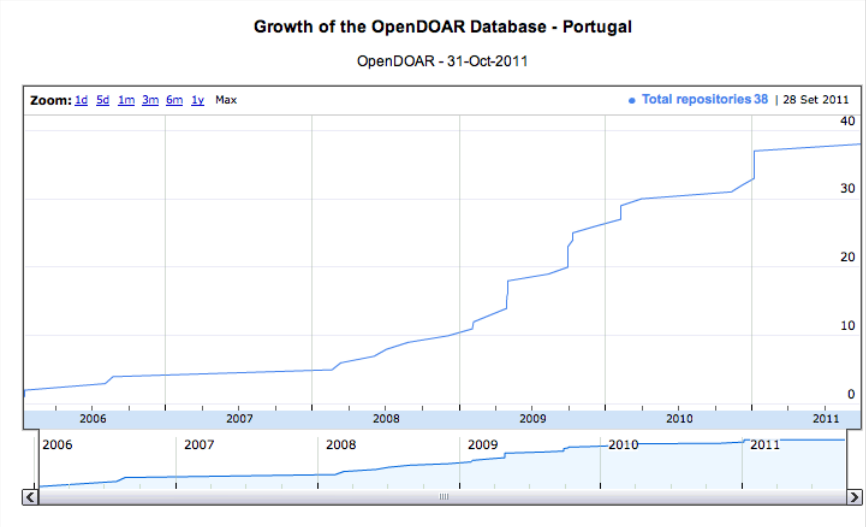


Figura 1.1: Evolução do número de repositórios portugueses no OpenDOAR

ção de consórcios de repositórios institucionais (e.g. RCAAP, RECOLECTA[3], DRIVER[4]). O RCAAP é uma iniciativa nacional de promoção do acesso-livre financiada pela UMIC - Agência para a Sociedade do Conhecimento e com a execução técnica da Fundação para a Ciência e Computação Nacional (FCCN) com o apoio da Universidade do Minho (UM). O RECOLECTA é o homologo espanhol promovido pela Fundación Española para la Ciencia Y la Tecnología (FECYT). O Digital Repository Infrastructure Vision for European Research (DRIVER) [Feijen et al., 2007] é uma iniciativa europeia, multi-fase, cujos objetivos iniciais eram ajudar e incentivar o desenvolvimento dos repositórios na Europa através da disponibilização de serviços e funcionalidades orientadas tanto para investigadores como para o público em geral assim como através da identificação, implementação e divulgação de normas e orientações para permitir o nível de interoperabilidade necessário para criar este serviço europeu. Estes consórcios têm como missão "aumentar a visibilidade, acessibilidade e difusão dos resultados da atividade académica e de investigação científica nacional, facilitar o acesso à informação sobre a produção científica nacional em regime de acesso aberto [...]" [Rodrigues et al., 2010] através de portais agregadores.

Entre outros serviços, estes oferecem ao consumidor final, a capacidade de pesquisar e consultar informação científica, sem necessidade de recorrer a diferentes interfaces de pesquisa, pois passam a poder fazê-lo

[3]http://www.recolecta.net/buscador/
[4]http://www.driver-repository.eu/

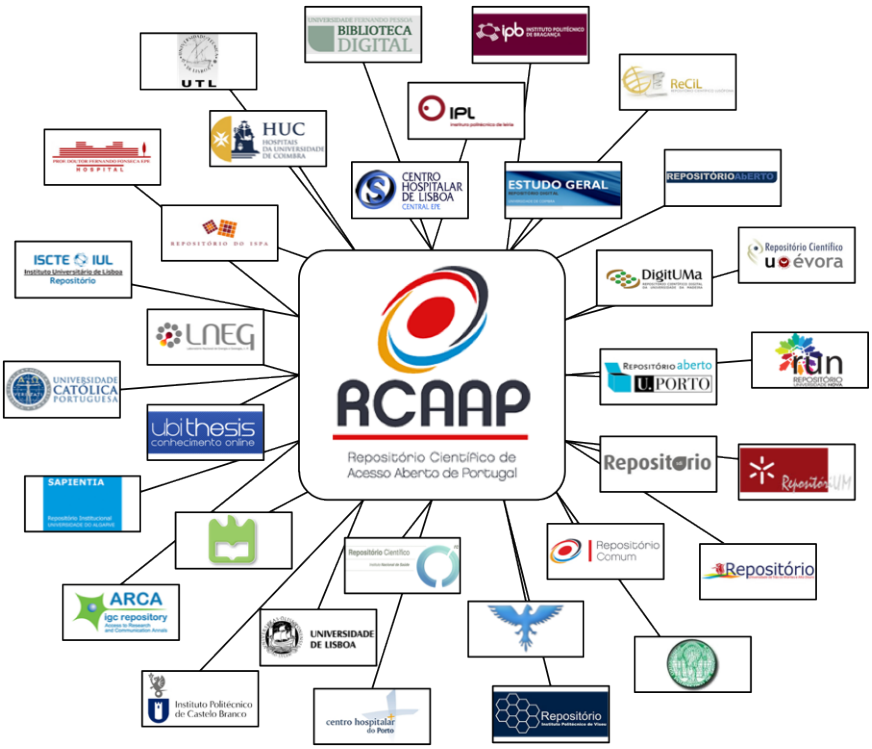


Figura 1.2: Portal RCAAP e repositórios associados

através de um único portal que congrega os registos de informação das várias instituições que participam no consórcio, assim como se pode ver na Figura 1.2. A salientar, no caso do RCAAP, o Serviço de Alojamento de Repositórios Institucionais (SARI). Este serviço permite, a instituições que não tenham capacidade financeira ou técnica (infraestrutura ou pessoal), ter um repositório institucional sem qualquer custo associado ou preocupação a nível de instalação e manutenção (tarefa ao cargo do projeto RCAAP).

Para as instituições participantes no consórcio, as vantagens também são assinaláveis. Passam a ter um modelo de gestão comum que emana orientações claras quanto à forma como um repositório deve ser gerido, tanto ao nível estratégico (definindo as políticas locais a implementar), como ao nível operacional, ditando, por exemplo, normas e vocabulários a adoptar na descrição dos seus materiais. Contudo, para realizar uma gestão de qualidade do consórcio, é fundamental que seja possível obter informação fiável e em tempo-real sobre o estado de cada repositório.

O SCEUR surge no contexto da 3ª edição do RCAAP como sendo um sistema central que permite agregar e apresentar dados estatísticos relativos à utilização de repositórios institucionais que participam no RCAAP, para assim concentrar e facilitar o acesso à informação necessária para a

gestão do consórcio, bem como para acompanhar iniciativas internacionais semelhantes.

1.1 Motivação

O aumento da quantidade de repositórios e o aparecimento de consórcios de repositórios, consórcios esses com papeis muito importantes no desenvolvimento e crescimento dos repositórios, comunidades e movimentos de acesso-livre à informação, só por si já é um bom motivo para construir uma ferramenta que ajude na gestão dos consórcios, sendo que para isso é necessário que essa ferramenta permita ter uma visão global do estado dos repositórios no contexto do consórcio, contrariamente às estatísticas locais que têm uma visão muito restrita e limitada. Mas por outro lado, temos os autores e as suas publicações. Essas habitualmente estão inseridas no contexto académico-científico , onde é necessário avaliar o impacto que as mesmas estão a ter, quer por questões de avaliação/financiamento, quer porque o autor pode querer acompanhar esse impacto e assim perceber a importância que o seu trabalho está a ter.

Até há alguns anos atrás, a única metodologia para o fazer era através da quantidade de citações que a publicação teve. Porem, este indicador tem alguns problemas, na medida em que por um lado fica limitado a trabalhos publicados, por outro fica limitado ao tipo de documento (o Institute of Scientific Information (ISI) apenas considera artigos em revista). Isto para não falar do intervalo de tempo entre a publicação e a obtenção destas métricas baseadas em citações.

Contudo, ao agregarmos informação de utilização, temos um factor de impacto não tão rígido e restrito mas um factor de impacto mais prático e genérico, na medida em que podemos medir/avaliar o impacto diretamente do uso/interesse que os utilizadores demonstram por um determinado documento. E esta metodologia tem várias vantagens. A mais imediata é que se aplica a todos os tipos de documentos (em contrapartida ao ISI que apenas considera artigos). Por outro, qualquer utilizador pode ver o impacto que os seus trabalhos estão a ter, e não apenas aqueles que publicam em revista. E ainda temos estas métricas em tempo-real, pois estas ficam disponíveis a partir do momento que o autor deposita o seu trabalho e começam a ser registadas consultas de metadados e *downloads*.

Esta metodologia, no imediato, não pretende ser uma substituta da que se baseia em citações, mas sim uma alternativa prática e fácil de implementar com melhores tempos de resposta e potenciais serviços associados de grande utilidade para o público em geral.

1.2 Objetivos

O SCEUR tem dois objetivos. Por um lado, disponibilizar uma arquitetura que permita recolher, processar e apresentar estatísticas associadas aos dados de utilização dos repositórios institucionais pertencentes ao consórcio RCAAP. Por outro lado, facilitar a disseminação de dados de utilização, quer pela disponibilização de add-ons para o software em uso, quer pela definição dos eventos de utilização a registar e informação extra associada. Com estes dois grandes objetivos, o SCEUR pretende ser a ferramenta ideal para auxiliar os vários intervenientes no processo de comunicação científica a avaliar/medir o impacto do seu trabalho.

1.3 Estrutura da dissertação

Nesta secção é apresentada a estrutura desta dissertação, descrevendo de forma sumária o conteúdo de cada capítulo.

No Capítulo 2 são apresentados os projetos mais relevantes para o SCEUR, assim como alguns conceitos-chave considerados importantes para a melhor compreensão da dissertação.

No Capítulo 3 são abordadas questões pertinentes relacionadas com estatísticas de utilização, tanto no contexto de um repositório como de um consórcio.

No Capítulo 4 explica-se as várias fases de desenvolvimento do SCEUR, passando tanto pelas decisões tomadas, detalhes da sua arquitetura ou mesmo pormenores mais técnicos relacionados com o seu desenvolvimento.

Por último apresentam-se algumas conclusões, abordando alguns assuntos que merecem discussão e também algum trabalho futuro.

Capítulo 2

Trabalho Relacionado

Neste capítulo serão apresentados os projetos mais relevantes para perceber as características e problemas associados à construção de uma arquitetura que permita agregar dados de utilização e prestar serviços associados, assim como alguns conceitos-chave considerados importantes para a melhor compreensão da dissertação. Por último, são tecidas algumas considerações acerca desses mesmos projetos.

2.1 Conceitos-chave

No decorrer desta secção apresentam-se alguns conceitos-chave necessários para uma melhor compreensão da dissertação.

2.1.1 Dados de utilização

Dados de utilização são toda a informação relativa às ações dos utilizadores nos repositórios institucionais, que fica registada, e permite aos repositórios apresentar estatísticas de utilização. Quando acontece uma consulta de metadados, fica registado a data e hora em que esse evento aconteceu, o endereço IP do utilizador que realizou a ação, o registo consultado (identificado pelo handle) e o identificador único do evento. Por outro lado, se houver um *download*, para além da informação registada associada a uma consulta de metadados, fica também registado o ficheiro descarregado (identificado pelo endereço através do qual se pode aceder diretamente ao mesmo), uma vez que um registo pode conter vários ficheiros.

2.1.2 OpenURL e OpenURL ContextObjects

O OpenURL [Van de Sompel and Hochstenbach, 1999a, Van de Sompel and Hochstenbach, 1999b, Van de Sompel and Hochstenbach, 1999c, Van de Sompel and Beit-Arie, 2001] foi desenvolvido por Herbert Van de Sompel, Patrick Hochstenbach e Oren Beit-Arie, tendo sido alguns anos mais tarde tornado uma norma pela National Information Standards Organization (NISO) [Organization, 2004]. Trata-se de um tipo de Uniform Resource Locator (URL), que permite aos utilizadores descobrir com maior facilidade uma cópia do recurso que procuram baseando-se no contexto em que os utilizadores se encontram, i.e., em diferentes contextos podemos aceder a diferentes sítios onde a cópia do recurso se encontra. Um exemplo prático da sua utilização é no contexto académico, onde habitualmente as universidades têm acordos (pagos, grande parte das vezes) com editoras, e como tal apenas se o utilizador estiver nesse contexto poderá aceder aos recursos disponibilizados pelas editoras.

No OpenURL há três entidades principais: *Source*, *Resolver* e *Target*.

Source Serviço consultado pelo utilizador e que contêm informação acerca do recurso a obter, informação com a qual se constrói um OpenURL que permite ao utilizador descobrir, por exemplo, uma cópia com o texto integral.

Resolver Serviço que sabe interpretar um OpenURL e que, dependendo do contexto, retorna uma lista de *links* onde o recurso referido pelo OpenURL pode ser encontrado. São comummente chamados de *linking servers*, ou seja, servidores/serviços que fazem a mediação do acesso a um recurso através de um *link* não estático. Um exemplo de *linking server* é o SFX [Van de Sompel and Hochstenbach, 1999b], que tem a informação guardada sobre que *Sources* podem aceder aos diferentes *Targets* no que eles chamam de "*knowledge base*".

Target Serviço onde está disponível o recurso, por exemplo, um texto integral.

Descreve-se agora um cenário onde o OpenURL é usado, cenário esse que podemos ver na Figura 2.1.

Um utilizador faz uma pesquisa num portal agregador de conteúdos (neste caso é a *Source*) e escolhe um registo da lista de resultados. Esse registo é composto por título, ano de publicação, nome do autor, o tipo de documento e outros tipos de informação. Com essa informação o portal constrói um OpenURL que pode ser usado para encontrar *links* para sítios onde se pode encontrar uma cópia do recurso. O utilizador carrega

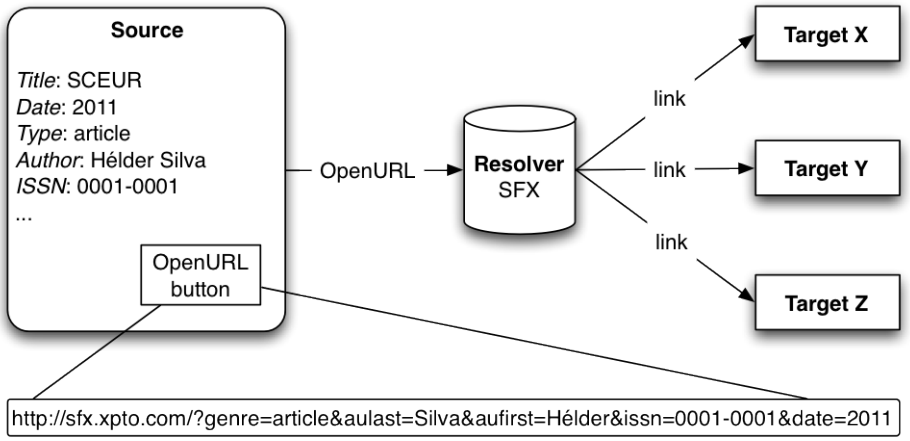


Figura 2.1: Cenário de utilização de OpenURL

nesse *link*, e são-lhe apresentados (pelo *Resolver*) zero ou mais alternativas. Se alguma alternativa for apresentada, o utilizador segue um desses *links* e chega ao sítio *Web* que contém a cópia (neste caso é o *Target*).

A informação contida no OpenURL é chamada ContextObject [Organization, 2004], sendo comummente representada em 2 formatos: Key Encoded Values (KEV) ou eXtensible Markup Language (XML). A representação KEV assemelha-se bastante à *query* contida num pedido HyperText Transfer Protocol (HTTP) GET. Um ContextObject têm 6 entidades principais, que são usadas para definir os vários aspetos de um pedido: *Referent*, *ReferringEntity*, *Requester*, *ServiceType*, *Resolver* e *Referrer*. Mas, essencialmente, um ContextObject pode ser resumido em três perguntas: **O quê?** - o item associado ao evento, **Quem?** - o utilizador que originou o evento e finalmente **Quando?** - informação sobre o instante de tempo em que o evento ocorreu.

2.1.3 COUNTER

Counting Online Usage of NetWorked Electronic Resources (COUNTER)[1] é uma iniciativa internacional com o objetivo de definir normas para facilitar o registo e partilha de estatísticas de utilização de recursos digitais de uma forma consistente, credível e interoperável. De forma simples, e para cada recurso eletrónico, define-se que informação deve ser guardada acerca da utilização e como a processar. Assim, os dados de utilização de diferentes de fontes que implementem as normas COUNTER, podem ser comparados e essa comparação fazer sentido.

[1] http://www.projectcounter.org

Para além desse objetivo, o projecto COUNTER também trabalha com algumas organizações na investigação e criação de serviços associados à utilização de recursos digitais. Em 2006 fizeram um trabalho de investigação, financiado pelo Joint Information Systems Committee (JISC), sobre o impacto das diferentes plataformas dos editores na utilização. Estão também a trabalhar em conjunto com o *UK Serials Group* sobre a possibilidade de criar uma nova métrica de utilização. Outro exemplo é o trabalho realizado com a NISO: o SUSHI.

2.1.4 SUSHI

O Standardized Usage Statistics Harvesting Initiative (SUSHI) [Needleman, 2006] é um protocolo descrito pela norma ANSI/NISO Z39.93-2007 que define um modelo pedido/resposta automatizado para recolher relatórios estatísticos de utilização de recursos digitais. Este protocolo usa *Web Services* baseados em Simple Object Access Protocol (SOAP) para fazer os pedidos e obtenção dos relatórios. O SUSHI tem uma extensão para COUNTER presente na sua especificação, uma vez que este representa grande parte dos relatórios estatísticos a serem recolhidos/partilhados.

2.1.5 OAI-PMH

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [Lagoze and de Sompel, 2001] é um protocolo usado para a interoperabilidade entre sistemas através da troca de metadados representados em XML. Desenvolvido pelo Open Archives Initiative[2], este protocolo permite, através de HTTP, que os repositórios (*Data Providers*) partilhem os seus metadados, para que outros repositórios (*Service Providers*) recolham esses mesmos metadados com facilidade.

2.2 Projetos relacionados

A seguir, apresentam-se alguns dos projetos mais relevantes para o contexto desta dissertação.

[2]http://www.openarchives.org

2.2.1 MEtrics from Scholarly Usage of Resources (MESUR)

Em 2006, Johan Bollen e Herbert Van de Sompel apresentaram um artigo intitulado de "*An Architecture for the Aggregation and Analysis of Scholarly Usage Data*" [Bollen and Van de Sompel, 2006]. Nesse artigo, justificam a criação de uma arquitetura que permita agregar e processar dados de utilização com a necessidade de se criar serviços de valor acrescentado construídos sobre dados de utilização. Esses serviços utilizam dados de utilização como entrada e permitem diferentes visões/análises desses mesmos dados. Nessa arquitetura, *linking servers* OpenURL registavam todos os pedidos dos utilizadores, que depois eram disponibilizados através de OAI-PMH.

Um dos problemas encontrados estava relacionado com o tamanho dos dados que estavam a analisar, que era significativo (3,5 milhões de pedidos, sendo que dados mais recentes apontam para mil milhões). Nesse contexto, escalabilidade, era o problema.

A arquitetura proposta é, segundo os autores, baseada em normas e é usada para "gravar, representar, partilhar e processar dados de utilização" [Bollen and Van de Sompel, 2006]. Tendo isso em consideração, eles explicam os vários componentes que integram a arquitetura proposta, dando grande importância aos *linking servers* OpenURL, como sendo as "entidades" que registam todos os pedidos dos utilizadores. OpenURL e a noção de ContextObjects é usado como a estrutura de dados, para encapsular pedidos "*context-sensitive*".

Segundo os autores, "uma nova geração de métricas de impacto científico e qualidade baseadas em dados de utilização podia ser criada, disponibilizada e usada para abalar o monopólio do *citation-based Impact Factor* do ISI[3]" [Bollen and Van de Sompel, 2006], o que é um bom motivo para a criação deste tipo de arquitetura e serviços de valor acrescentado.

Deste projeto, algumas conclusões foram tiradas pelos autores. Por um lado, que de facto era possível implementar a arquitetura descrita. Por outro, e de forma semelhante ao algoritmo *PageRank* da Google, algoritmo que determina que se uma página é referida em páginas com grande impacto, ela própria tem grande impacto, usando os dados de utilização eles criaram um *Usage PageRank*, onde reuniram as 10 principais revistas. Para tal, co-relacionaram os eventos de cada utilizador (identificado pelo endereço IP) e cada artigo, por exemplo, descarregado. Para co-relacionar, analisaram cada par de linhas subsequentes do *log* que continha essa informação. Se o artigo tivesse sido descarregado pelo mesmo

[3]http://www.isiknowledge.com

utilizador, num período de tempo inferior a uma hora, então os jornais onde foram publicados os artigos estão "relacionados" (para mais informação consultar [Bollen et al., 2005]). Outras relações foram feitas, tendo por base dados de utilização, dando assim origem a alguns serviços de recomendação [Bollen and Van de Sompel, 2006].

2.2.2 Open Access Statistics

Open Access Statistics é uma iniciativa alemã com os mesmos objetivos do MEtrics from Scholarly Usage of Resources (MESUR), i.e., criar uma arquitetura que permita agregar e analisar dados de utilização. No documento técnico [Metje and Hilse, 2009] disponibilizado no seu sítio *Web*, especificações sobre o formato e métodos de troca de informação são explicitadas. É também apresentado um novo facto: que o pedido pode não ser mediado por um *linking server* (e.g., se for uma infraestrutura mais simples e composta apenas por repositórios institucionais), o que desempenhava um papel fundamental no MESUR.

O XML é o formato escolhido para representação dos dados porque permite uma verificação formal usando *schemas* associados ao OpenURL ContextObject. OpenURL ContextObject é extensível (usando XML Schemas personalizados), mas essa extensibilidade tem um custo: perda de compatibilidade, na medida em que essa personalização pode não ser compreendida pelos outros sistemas.

Relativamente ao protocolo usado para partilhar a informação, OAI-PMH foi o escolhido. Uma vez que por definição ele é usado para trocar metadados de documentos, alguns ajustes tiveram que ser feitos, sendo que essas alterações são abordadas em profundidade no documento técnico.

O *Data Provider* desenvolvido é baseado no registo de dados de utilização (*logging*), quer usando serviços que geram diretamente OpenURL ContextObjects em XML quer convertendo dados de utilização guardados em ficheiros para esse mesmo formato.

2.2.3 NEEO Scholarly Works Usage Community Profile (SWUP)

Em 2009, o consórcio Network of European Economists Online (NEEO) definiu algumas orientações para a partilha de dados de utilização [Pauwels, 2009]. Neste projeto refere-se o facto de que grande parte dos repositórios institucionais mantêm os eventos registados em *logs*. É também comum que os repositórios institucionais apresentem dados de utilização

de alguma forma (quantas vezes uma página foi vista, quantas vezes um ficheiro foi descarregado, etc). Então, e baseados nos dados de utilização, serviços de valor acrescentado podem ser criados. Podem criar-se *clusters* de publicações relacionadas, podendo daí extrair-se informações de outro género, como por exemplo, relacionar revistas onde foram publicadas essas publicações ou mesmo serviços de recomendação onde o utilizador é informado de documentos relacionados com o que está a ver/descarregar.

Uma vez mais, se a informação usada para construir os serviços for proveniente de várias fontes, os serviços prestados serão de melhor qualidade. Por exemplo: uma publicação pode ter vários autores, de diferentes instituições, logo é provável que essa publicação se encontre em vários repositórios.

No que diz respeito ao uso do OAI-PMH, foi necessário fazer alguns ajustes, assim como aconteceu no projeto analisado anteriormente. A referir está também o facto de todas as respostas OAI-PMH serem validadas com XML Schemas associados ao OpenURL ContextObjects.

Finalmente alguns perfis são definidos dentro do consórcio para especificar a representação comum dos dados de utilização a serem partilhados.

2.2.4 KE Usage Statistics Guidelines

Em 2010, num esforço conjunto entre a JISC[4], SURF[5], Denmark's Electronic Research Library (DEFF)[6] e DFG[7], algumas orientações para agregar e partilhar dados de utilização foram publicadas [Verhaar and Vanderfeesten, 2010]. Alguns dos projetos desses consórcios, como o SURF SURE, PIRUS2, OA-Statistics, NEEO e COUNTER irão usar essas orientações para garantir a compatibilidade entre os diferentes fornecedores de dados.

É referido o facto de que até há alguns anos atrás a única metodologia para medir o impacto/qualidade dos documentos produzidos no contexto académico e científico era através do número de citações (habitualmente em revistas). Mas essa estratégia apresenta vários problemas, na medida em que por vezes apenas um tipo de publicação é considerado (no ISI apenas artigos em revistas), para não falar do intervalo de tempo entre a publicação e a disponibilização dessas métricas.

A nova estratégia tem várias vantagens. Todos os tipos de documentos

[4]http://www.jisc.ac.uk
[5]http://www.surf.nl
[6]http://www.deff.dk
[7]http://www.dfg.de

são considerados (não apenas artigos em revistas). E essas métricas estão disponíveis para todos os utilizadores (não apenas para utilizadores de revistas). Mais ainda, os dados de utilização começam logo a ser registados, e assim disponíveis para consulta.

Juntamente com essas vantagens, praticamente todos os servidores *Web* registam a atividade dos utilizadores, com algum nível de detalhe. O único problema é que diferentes servidores *Web* têm diferentes formatos. E se queremos comparar dados de diferentes entidades, os dados têm que ser comparáveis. A solução passa por criar guias de orientação e normalização dos dados de utilização. Vários projetos tiveram essa abordagem. O MESUR gera OpenURL ContextObjects em XML. O JISC faz o mesmo. Tendo isso em consideração, é possível implementar uma plataforma para agregar dados de utilização de várias fontes.

No que diz respeito ao protocolo para transferir os dados de utilização, e de forma diferente dos outros projetos analisados, tanto OAI-PMH como SUSHI são considerados. OAI-PMH, na sua versão 2.0, é detalhado, relativamente às alterações necessárias para "acomodar" OpenURL ContextObjects, sendo considerado um protocolo amplamente usado [Verhaar and Vanderfeesten, 2010]. Definem também o formato de metadados que a interface OAI-PMH dos projetos que se guiem por estas orientações deve responder: "CTXO". Relativamente ao SUSHI, o seu uso é justificado pela necessidade de um protocolo mais fiável no que diz ao tratamento/recuperação de erros. É também explicado como é feita a interação e que informação é necessária para efetuar o pedido de um relatório.

Finalmente, são tecidas algumas considerações no que diz respeito à normalização, assim como que tratamento dar aos duplos *clicks* e filtragem de *robots*.

2.3 Discussão

Tendo analisado os projetos mais relevantes para a construção de uma arquitetura que permita agregar dados de utilização e prestar serviços associados, explicitam-se agora os aspetos pertinentes que dessa análise se podem retirar e que serão úteis para o desenvolvimento de uma solução que dê respostas às necessidades estatísticas tanto dos utilizadores como dos administradores dos repositórios e/ou consórcios.

Por um lado, temos a quantidade de informação que é necessária armazenar e processar para conseguir gerar estatísticas de utilização, que pode ser na ordem dos milhões. Por outro, em alguns projetos a tarefa de

registar os eventos de utilização fica do lado dos *linking servers*, o que nem sempre é possível, principalmente se nos encontrarmos em arquiteturas mais simples onde não existem servidores a mediar os pedidos ou mesmo quando não temos controlo sobre os mesmos. Temos ainda a importância dada ao que podem ser novas métricas de impacto baseadas em dados de utilização e que se sucedidas, podem abalar o monopólio do *Impact Factor* do ISI. OpenURL e os ContextObjects são usados de forma generalizada para transportar a informação relativa aos eventos de utilização, na sua versão XML, pois facilita a verificação com XML Schemas e também porque o OpenURL é extensível e pode assim ser adaptado aos diferentes contextos. Relativamente aos protocolos a usar para troca de informação, tanto o OAI-PMH como o SUSHI são aceites, sendo que o OAI-PMH é considerado um protocolo estável e amplamente usado, e o SUSHI é descrito como ideal no que diz respeito a tratamento de erros. Por último, questões referentes à normalização dos dados de utilização são consideradas importantes, contudo não há concertação entre os projetos no que diz respeito ao tratamento dos "*double clicks*", ou seja, a quantidade de tempo que se deve considerar para assumir que dois *downloads* seguidos de um mesmo ficheiro por um mesmo utilizador devem ser considerados apenas um, assim como o tratamento dos *robots* que é feito nos diferentes projetos.

Podem-se considerar estas questões, tanto técnicas como administrativas, pertinentes e que merecem atenção uma vez que são fruto da experiência que saiu dos diferentes projetos analisados e que podem ser uma mais valia no desenvolvimento de projetos semelhantes.

Capítulo 3

Estatísticas de Utilização

Neste capítulo vai analisar-se a temática "estatísticas de utilização", quer no contexto de um repositório institucional quer no contexto de um consórcio. Existem diferentes motivações em cada contexto, com características muito próprias, assim como problemas e limitações.

3.1 Repositório Institucional

3.1.1 Para quem e para quê?

É comum os repositórios institucionais apresentarem estatísticas de utilização, quer sobre a forma de gráficos, quer sobre a forma de dados numéricos. Estas, para além auxiliarem os administradores do repositório a compreenderem a utilização que o repositório tem e suas características, permite aos autores perceber que impacto o seu trabalho está a ter. Estes dados, em alguns casos, foram disponibilizados com o objetivo de potenciar o auto-arquivo [Ferreira et al., 2008], sendo que em outros casos foram disponibilizados para convencer os autores que o repositório é uma excelente ferramenta para disseminar os seus resultados [Bell et al., 2005].

3.1.2 Tipo de estatísticas

Dependendo do software que suporta o repositório e/ou de add-ons que estejam a ser utilizados, diferentes estatísticas poderão ser apresentadas. No caso muito particular dos repositórios associados ao projeto RCAAP em regime de SARI, que usam DSpace[1] e que usam o add-on de estatísti-

[1] http://www.dspace.org/

Figura 3.1: Estatísticas disponíveis para o administrador (Rep. Comum)

cas da UM[2], existem diferentes níveis de estatísticas: geral, comunidade, coleção e documento. DSpace é um software para repositórios digitais, desenvolvido pelo Massachusetts Institute of Technology (MIT), que permite às instituições que o possuem recolher, preservar, gerir e disseminar a produção intelectual por eles produzida, seja ela fruto de investigação, de materiais educacionais e/ou outros [Rodrigues, 2005, Smith et al., 2003]. O add-on de estatísticas da UM, chamado de *MINHO STATS*, foi especialmente desenvolvido pois o sistema de estatísticas do DSpace era demasiado simplista/básico e não preenchia as necessidades estatísticas da instituição. Relativamente às estatísticas, no nível geral é possível ver estatísticas globais de acesso, de conteúdos e administrativas relativas ao conjunto de comunidades existentes no repositório. No nível comunidade e coleção é possível ver estatísticas relacionadas com as comunidades ou coleções às quais o utilizador pertence ou a todas caso seja administrador. Por último, no nível documento é possível ver estatísticas relacionadas com um documento em particular, identificado pelo seu handle.

Dependendo do tipo de utilizador, tem acesso às diferentes estatísticas. Se for um utilizador não autenticado, apenas tem acesso ao nível geral e documento. Já se o utilizador estiver autenticado e for administrador tem acesso a todos os níveis de estatísticas, como se pode ver na Figura 3.1, tendo como exemplo o Repositório Comum[3].

São agora listadas de forma mais detalhada as estatísticas disponibilizadas:

- Total de *downloads* efetuados (nos diferentes anos ou em períodos

[2]http://projecto.rcaap.pt/index.php/lang-pt/consultar-recursos-de-apoio/remository?func=fileinfo&id=327

[3]http://comum.rcaap.pt/

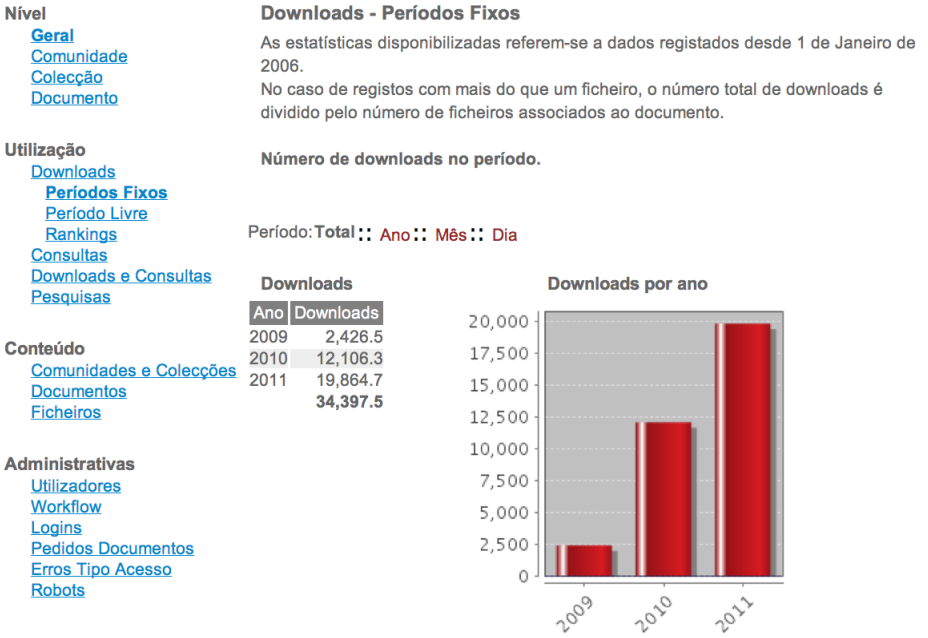


Figura 3.2: Estatísticas de download disponíveis para o administrador (Rep. Comum)

definidos) e valores associados a cada país de origem do *download* (ver exemplo na Figura 3.2);

- Total de consultas efetuadas (nos diferentes anos ou em períodos definidos) e valores associados a cada país de origem da consulta;
- Total de pesquisas efetuadas (nos diferentes anos ou em períodos definidos) e *top* de expressões pesquisadas;
- Quantidade de utilizadores registados (internos à instituição e externos);
- Quantidade de documentos processados em todos os passos do *workflow*;
- Total de *logins* efetuados (nos diferentes anos ou em períodos definidos);
- Quantidade de pedidos de documentos em acesso restrito (nos diferentes anos ou em períodos definidos) e *top* de documentos pedidos;
- Quantidade de comunidades;
- Quantidade de coleções;

- Quantidade de documentos total, por comunidade e por coleção;
- Quantidade média de documentos por comunidade e por coleção;
- Total de documentos (nos diferentes anos ou em períodos definidos);
- Total de documentos por tipo;
- Total de documentos por idioma;
- Total de documentos por tipo de acesso;
- Total de documentos por ano de publicação;
- Total de documentos por autor.

3.1.3 Limitações

No contexto dos repositórios, há um conjunto de funcionalidades/informações que não estão disponíveis, e que são úteis tanto para os administradores do sistema como para os outros utilizadores.

No que diz respeito aos utilizadores, são várias as vezes em que se produzem artigos e outro tipo de documentos envolvendo elementos de várias instituições. Nesses casos, é bem provável que cada instituição tenha o seu repositório digital. Estando um ou mais repositórios envolvidos, a tarefa de geração de estatísticas torna-se mais complicada. Essa tarefa pode tornar-se ainda mais complicada se estiverem envolvidos diferentes conjugações de software e/ou add-ons.

As estatísticas disponibilizadas podiam ser mais ricas, na medida em que no repositório tem-se acesso aos metadados descritivos dos registos associados aos quais aconteceram os eventos de utilização. Por exemplo, saber quais os tipos de documentos (dc.type) mais consultados num determinado período de tempo ou então quais os tipos de ficheiros (dc.format) mais descarregados num determinado período de tempo.

Há ainda duas questões pertinentes que tanto são úteis aos administradores como aos utilizadores. A primeira prende-se com o facto de não ser possível subscrever estatísticas de forma periódica. É algo útil na medida em que há utilizadores que podem querer acompanhar as estatísticas de alguns documentos relevantes, assim como os administradores têm que gerar de forma periódica relatórios onde as estatísticas fazem parte do conteúdo. A segunda está relacionada com o facto de não se puder exportar as estatísticas para formatos que permitam o processamento analítico das mesmas (Comma Separated Values (CSV), XML, etc), quer através

da interface gráfica dos repositórios quer através de interfaces normalizadas e específicas para expor dados (OAI-PMH, outras).

3.2 Consórcio

3.2.1 Para quem e para quê?

Com o surgimento de consórcios de repositórios institucionais (RCAAP, RECOLECTA, DRIVER, entre outros), surge também a necessidade de obter métricas para poder avaliar o comportamento dos mesmos. Esta avaliação, para além de justificar gastos inerentes ao funcionamento dos consórcios, serve também para determinar a eficácia das medidas e políticas implementadas. Se por um lado é necessário ser capaz de determinar o grau de aplicação das mesmas por parte dos repositórios, por outro é necessário determinar a eficácia dessas medidas na divulgação e disseminação do conteúdo científico produzido. Se considerarmos que um consórcio é composto por mais de 30 repositórios, conseguimos ter uma ideia do esforço necessário para que o gestor/administrador desse consórcio possa reunir algumas estatísticas. Para complicar ainda mais e aliado a essa quantidade de repositórios, os repositórios podem usar diferentes plataformas de software com diferentes interfaces, com diferentes metodologias para geração de estatísticas. E para complicar ainda mais um pouco, as estatísticas podem não estar em acesso-livre, o que implica ter credenciais para aceder ao sistema, o que ainda torna mais difícil de agilizar o processo de obtenção de estatísticas.

Por outro lado, ao dispor de um leque alargado de estatísticas com uma visão global e normalizada de todos os repositórios, a administração do consórcio pode conseguir delinear perfis de utilização e assim implementar novas medidas ou até mesmo serviços específicos para melhor servir as necessidades dos seus utilizadores.

Algo igualmente importante, e associado às políticas implementadas, é poder comparar os repositórios e perceber as suas evoluções, pois as estatísticas locais não conseguem dar uma visão global do consórcio, uma vez que estão restritas ao seu contexto. Assim, deve também ser possível determinar *rankings* relativamente a vários aspetos importantes na utilização de um repositório ou mais repositórios, como é o caso dos tipos de documentos mais consultados e/ou descarregados (e.g. artigo, tese de mestrado, outros), de que países foram acedidos, quais as suas políticas de acesso (para assim perceber o impacto do acesso-livre, por exemplo), entre outros.

Para além dos aspetos mais ligados à gestão/administração dos consórcios, estas estatísticas devem ser de acesso público, na medida em que podem incentivar a produção científica e o depósito nos repositórios institucionais.

3.2.2 Tipo de estatísticas

Uma vez que um consórcio é composto por um conjunto de repositórios institucionais, a entidade elementar é então repositório. Nesse sentido todas (ou quase todas) as estatísticas disponibilizadas no repositório deverão estar disponíveis para os gestores/administradores do consórcio. Contudo, no consórcio o mais importante é ter uma visão sobre todos os repositórios. Assim sendo é importante:

- Poder comparar um ou mais repositórios;
- Gerar com facilidade uma ou mais estatísticas;
- Poder agendar a geração e envio de uma ou mais estatísticas;
- Obter estatísticas em diferentes formatos (gráfico, numérico, outros);
- Poder embeber gráficos em sítios *Web* externos (com dados sempre atualizados);
- Poder normalizar as estatísticas para que estas possam ser representativas e comparáveis.

3.2.3 Limitações

No contexto da gestão de um consórcio, há um conjunto de questões que devem ser colocadas, respeitantes tanto a pormenores técnicos como a questões administrativas, e que devem ser respondidas para que essa gestão seja feita com eficácia.

No que diz respeito a questões administrativas, pode falar-se sobre os eventos de utilização disponibilizados. Quais devem ser? Os que atualmente já se disponibilizam nos repositórios (consultas de metadados, *downloads* e depósitos)? Outros? E ainda associado aos eventos de utilização. Deve também disponibilizar-se informação extra, nomeadamente metadados dos registos associados aos quais aconteceram os eventos? Se sim, que informação extra é relevante disponibilizar, para questões estatísticas?

Há ainda outra questão administrativa que deve ser considerada, e que está relacionada com as comunidades, coleções e sub-coleções de um repositório. É prática corrente estruturar o repositório por comunidades, coleções e sub-coleções de forma a que este reflita a estrutura orgânica/organizacional da própria instituição. Nesse sentido a questão a colocar é, se esta informação é relevante para a administração de um consórcio. E em caso afirmativo, que tipo de estatísticas se devia gerar/apresentar.

Relativamente a questões técnicas, já muito relacionadas com o desenvolvimento de uma solução que permita a geração de estatísticas com o objetivo de auxiliar as tarefas de gestão de um consórcio, está a normalização dos dados estatísticos. Esta questão coloca-se pois o objetivo é recolher a informação de diferentes fontes. E essas mesmas fontes podem (ou não) expor dados que estejam normalizados, quer por incumprimento das orientações definidas pelo consórcio, quer pela inexistência de regras e/ou vocabulários controlados, levando deste modo a que possam existir diferentes valores nos diferentes repositórios associados a um mesmo campo. Nesse sentido, a questão a colocar é se se deve normalizar, e se sim se se deve guardar os valores antes da normalização e assim permitir estatísticas sobre valores não normalizados, assim como sobre valores normalizados.

Outra questão técnica é sobre a escalabilidade de um sistema de estatísticas. Depois de uma análise feita aos dados já armazenados de utilização dos diferentes repositórios associados ao projeto RCAAP em regime de SARI (ver anexo A), chegou-se à conclusão que se irá armazenar eventos de utilização na ordem dos milhões (sensivelmente 2 milhões de eventos de *download* e consulta de metadados registados entre 2008-04-14 e 2011-01-12 em 17 repositórios). O que nos leva a questionar se as estatísticas, tendo em consideração a ordem de grandeza dos valores encontrados, devem ser geradas em tempo-real ou de alguma forma devem ser pré-calculadas sobre a forma de valores agregados acumulados.

Capítulo 4

SCEUR: um Serviço de Estatísticas

Neste capítulo descrevem-se as várias fases do desenvolvimento do Serviço Centralizado de Estatísticas de Utilização de Repositórios. Numa primeira secção explicam-se as decisões tomadas, que dão resposta às várias questões técnicas e administrativas que existem relativamente às estatísticas no contexto de um repositório e/ou consórcio. Explicitam-se, ainda, as funcionalidades que serão disponibilizadas aos utilizadores. Na segunda secção explica-se a arquitetura do SCEUR, dando tanto uma visão geral do sistema, como um visão mais detalhada de cada um dos módulos que o compõe. Na terceira secção detalham-se questões técnicas relacionadas com o desenvolvimento, mostrando também alguns dos processos mais importantes, expondo assim algum do dinamismo do sistema de estatísticas. Por último, apresentam-se os resultados obtidos.

4.1 Decisões e Funcionalidades

Depois de analisadas as características das estatísticas dos sistemas que é necessário integrar, assim como as necessidades de um consórcio que gere um conjunto de repositórios, tomaram-se decisões relativamente às características que um serviço de estatísticas que tem por objetivo auxiliar nas tarefas de gestão deverá ter, assim como as suas principais funcionalidades.

Decisões:

- Tomar como referência as recomendações presentes no documento "KE Usage Statistics Guidelines" [Verhaar and Vanderfeesten, 2010] (ver

secção 2.2.4), com o objetivo de criar um sistema que seja compatível com outros sistemas semelhantes no contexto internacional;

- Publicar os eventos de utilização referentes a *downloads*, consulta de metadados (chamado de "consulta" no add-on *MINHO STATS*) e depósitos. Esse add-on também regista *logins*, pesquisas e informação relativa aos diferentes passos do *workflow* de inserção de um registo no repositório, porem estes não serão alvo de estatísticas a nível central;

- Enviar, juntamente com a informação de cada evento, metadados descritivos com o objetivo de obter estatísticas mais ricas. Será enviado, sempre que estejam disponíveis, informação sobre os autores (dc.author), o tipo de documento (dc.type), as políticas de acesso (dc.rights), o formato (dc.format) e a língua (dc.language). A notar está o facto de que a informação enviada é pública, na medida em que está disponível quer na interface *Web* dos repositórios, assim como noutras interfaces normalizadas de partilha de informação (e.g. OAI-PMH);

- Disponibilizar todas as estatísticas disponíveis num repositório, nas diferentes dimensões temporais, com a exceção de estatísticas relacionadas com: utilizadores, *workflow*, *logins* e pesquisas;

- Usar OAI-PMH como o protocolo para troca de informação relativa a eventos de utilização entre os repositórios e o serviço centralizado de estatísticas. Esta decisão é tomada tendo em consideração que é um protocolo estável, bastante usado (o próprio DSpace já o usa para partilhar metadados descritivos) e também porque existe um maior conhecimento sobre o protocolo e seu funcionamento em detrimento do SUSHI;

- Desenvolver um add-on que permita aos repositórios baseados na plataforma DSpace publicar eventos de utilização. Esta decisão prende-se principalmente com o facto de grande parte dos repositórios associados ao projeto RCAAP serem baseados em DSpace e porque é, segundo o *OpenDOAR* (ver Figura 4.1), o *software* de repositórios digitais mais usado no mundo no que diz respeito aos repositórios registados até 2011-10-31;

- Gerar estatísticas em tempo-real;

- Usar um sistema de indexação para dar resposta a questões relacionadas com escalabilidade e performance.

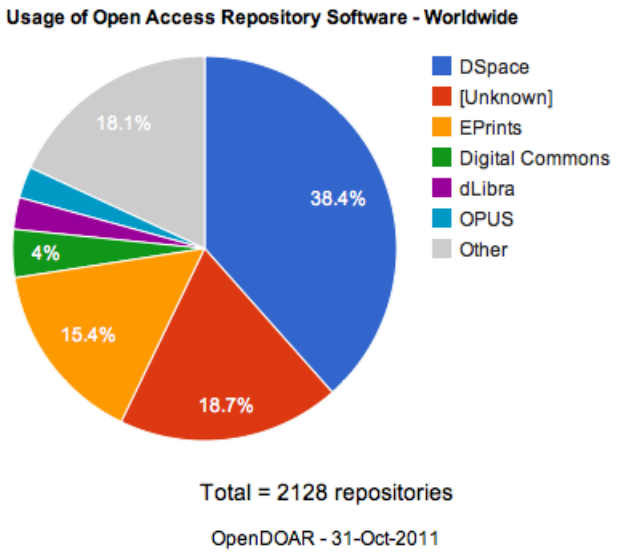


Figura 4.1: Ranking de software de repositórios digitais mais usados no mundo segundo o OpenDOAR

No que diz respeito às funcionalidades, o SCEUR irá disponibilizar as seguintes:

- Geração dinâmica de estatísticas, sobre a forma de gráficos e dados numéricos, assim como o código HTML que permite incluir essas mesmas estatísticas em sítios *Web*;
- Exportação de estatísticas para CSV;
- Subscrição do envio de estatísticas (gráfico e dados numéricos) de forma periódica por e-mail;
- Disponibilizar um conjunto de estatísticas pré-definidas, relativas ao consórcio RCAAP e repositórios que o compõe.

4.2 Arquitetura

Nesta secção vai explicar-se a arquitetura do SCEUR. Começa-se por dar uma visão geral do sistema, identificando os diferentes intervenientes. Depois, apresenta-se de forma mais detalhada o OAI-PMH como protocolo a usar para troca de eventos de utilização. Por último, apresenta-se cada um dos módulos funcionais, explicando como funcionam e qual o seu papel no sistema de estatísticas.

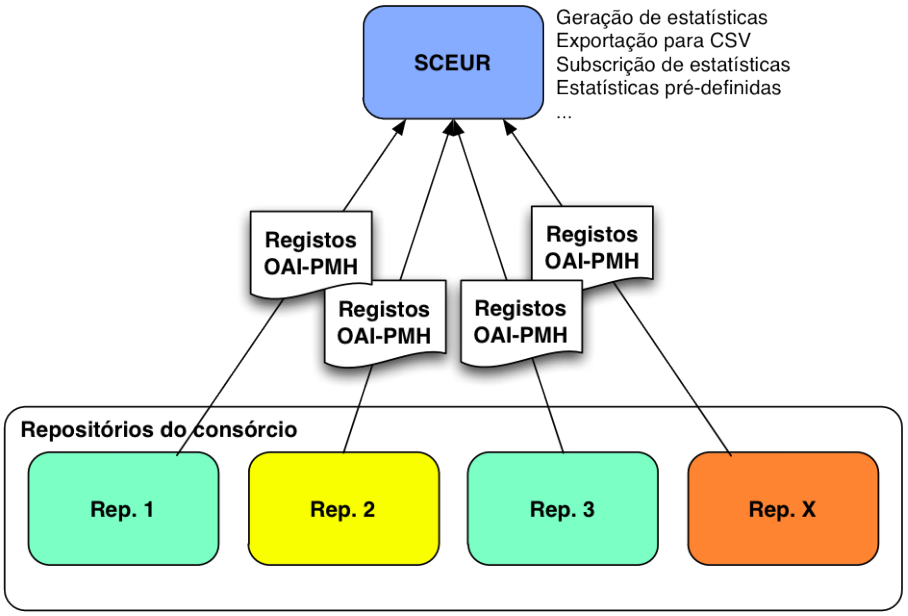


Figura 4.2: Arquitetura geral do SCEUR

4.2.1 Visão Geral

O SCEUR, como sistema centralizado de estatísticas, tem dois intervenientes principais: o próprio SCEUR e os repositórios institucionais dos quais será possível gerar estatísticas. Assim como se pode ver na Figura 4.2, um ou mais repositórios institucionais partilharão os seus dados de utilização através de OAI-PMH, dados esses que depois serão processados, armazenados e disponibilizados sobre a forma de estatísticas (gráficos, dados numéricos, outros). Desta forma, os repositórios institucionais são fornecedores de dados (*Data Provider* na terminologia OAI-PMH) e o SCEUR é fornecedor de serviços (*Service Provider* na terminologia OAI-PMH).

4.2.2 OAI-PMH

O OAI-PMH desempenha um papel fundamental no processo de partilha/obtenção dos dados de utilização, a partir dos quais se disponibilizarão estatísticas. Nesse sentido, explica-se agora o que é, para que serve, algumas das suas entidades principais e o seu funcionamento, para melhor compreender este protocolo.

O OAI-PMH [Lagoze and de Sompel, 2001] é um protocolo usado para a interoperabilidade entre sistemas através da troca de metadados repre-

sentados em XML. Desenvolvido pelo Open Archives Initiative[1], este protocolo permite, através de HTTP, que os repositórios (*Data Providers*) partilhem os seus metadados, para que outros repositórios (*Service Providers*) recolham esses mesmos metadados com facilidade.

Descrevem-se agora algumas das entidades/conceitos principais no OAI-PMH.

Harvester *Harvester* ou agregador é uma aplicação cliente que emite pedidos OAI-PMH com o objetivo de recolher metadados. Este é habitualmente executado por um *Service Provider*.

Repository *Repository* ou repositório é um servidor que está acessível na rede e que sabe processar os seis pedidos definidos no OAI-PMH. Habitualmente é gerido por um *Data Provider* que expõe os metadados aos *harvesters*. Há três entidades distintas relacionadas com os metadados acessíveis através de OAI-PMH:

Resource Objeto descrito por metadados, de natureza física ou digital, armazenado no repositório ou noutro sistema, que está fora do âmbito OAI-PMH;

Item Componente de um repositório acerca de um *resource*, associado ao qual se pode disseminar metadados;

Record Registo de metadados num formato específico. É retornado, codificado em XML, como resposta a um pedido de recolha de um *item* num determinado formato de metadados. É identificado de forma unívoca combinando o identificador único, o *metadataPrefix* que identifica o formato de metadados e um *datestamp* do registo.

O registo é composto por três partes: *header*, *metadata* e *about*. No *header* está informação sobre o registo, nomeadamente o seu identificador, a data de criação/última modificação, a que *Sets* pertence (não sendo obrigatório pertencer a um) e opcionalmente um elemento chamado *status* com o valor *deleted* caso este registo tenha sido eliminado. Em *metadata* aparece então os metadados, que se quer partilhar, no formato escolhido. No *about*, de forma opcional, pode aparecer informação relacionada com a origem do registo, entre outras.

Unique Identifier Identificador que identifica de forma unívoca um *item* dentro de um repositório. É usado nos pedidos OAI-PMH para obter os metadados de um *item* em específico.

[1] http://www.openarchives.org

Set Elemento opcional de um *Data Provider* que permite agrupar/organizar *items* e que possibilita a sua recolha de forma seletiva. A organização dos *sets* pode ser plana (uma simples lista) ou hierárquica. Quando um repositório tiver uma organização orientada ao *set*, essa informação deve ser adicionada ao *header* dos registos.

Selective Harvesting A recolha seletiva permite aos agregadores limitar a recolha de metadados a porções bem definidas disponíveis no repositório. O OAI-PMH suporta dois tipos de agregação seletiva baseando-se em dois tipos de critérios, que podem ser combinados: *datestamps* e pertença a *set*.

- **datestamps** A agregação seletiva permite recolher registos que foram criados, alterados ou eliminados dentro de um intervalo de datas específicas. Para tal, e de forma opcional, pode usar-se dois argumentos, *from* e *until* para especificar a data a partir do qual o registo foi criado, alterado ou eliminado (*from*) ou a data até à qual o registo foi criado, alterado ou eliminado (*until*) para que faça parte da lista de resultados.
- **set** A agregação seletiva permite recolher registos que pertençam a um determinado *set*, especificado através do argumento *setSpec*, sendo que se não for especificado nenhum serão retornados todos os registos.

Os *Data Provider* respondem a seis pedidos, chamados *verbs*, que são descritos a seguir:

Identify Este pedido retorna informação útil sobre o *Data Provider*, nomeadamente o nome, o endereço eletrónico do administrador, entre outras informações;

ListMetadataFormats Este pedido retorna a lista dos diferentes formatos de metadados que o *Data Provider* disponibiliza, i.e., as diferentes maneiras de fazer *output* da informação armazenada segundo diferentes XML Schemas. Exemplos de formatos para informação bibliográfica são: Dublin Core, METS, entre outros;

ListSets Pedido que retorna o conjunto de *Sets* disponíveis;

ListIdentifiers Este pedido permite recolher apenas a parte do *header* dos registos de um repositório. Este pedido é parametrizável, para efeitos de recolha seletiva, permitindo também definir o formato de metadados (*metadataPrefix*);

ListRecords Este pedido permite recolher os registos de um repositório. Este pedido é parametrizável, para efeitos de recolha seletiva, permitindo também definir o formato de metadados (*metadataPrefix*);

GetRecord Este pedido permite recolher os metadados de um registo em particular. Para tal deve especificar-se o identificador do *item* (identifier) e o formato de metadados que deve ser incluído no registo (*metadataPrefix*).

A notar está o facto de haver, no OAI-PMH, três pedidos de *list*, pedidos esses que podem retornar listas com um tamanho pouco prático de trabalhar. Nesse sentido, torna-se necessário e prático puder partir essas listas em vários pedidos/respostas. Assim, quando uma resposta retornar uma lista incompleta (tamanho definido pelo próprio repositório), o repositório deve também adicionar um elmento chamado *resumptionToken*, que deve ser usado no pedido seguinte para dar continuidade ao processo, sendo esta metodologia usada de forma repetida até que o repositório emita uma lista com o elemento *resumptionToken* vazio, querendo com isso dizer que essa lista completa a lista total.

4.2.3 Módulos

Seguindo uma filosofia de conceção e desenvolvimento de software modular, o SCEUR é composto por sete módulos funcionais: **core**, **workbench**, **dashboard**, **subscription service**, **data generation service**, **SolrClient** e **OAI-Stats**:

core Módulo principal do SCEUR. É o responsável por recolher (através do *harvester* OAI-PMH), processar e armazenar os dados de utilização dos diferentes repositórios. Disponibiliza também a API para gerar as diferentes estatísticas nas diferentes formas (gráficos, dados numéricos, outros), assim como para aceder a um conjunto variado de informações sobre o sistema, os repositórios, entre outros.

Num modelo de software de três camadas [Fowler, 2002], assim como se pode ver na Figura 4.3, este módulo situa-se na camada de lógica de negócio;

workbench Módulo que permite ao utilizador configurar, gerar e subscrever estatísticas. Fazendo alusão ao nome, este pretende ser uma "bancada de trabalho", que permite ao utilizador configurar as estatísticas que pretende obter. Através de uma página *Web*, é disponibilizado ao utilizador um conjunto variado de opções, quer

relativas às diferentes estatísticas que podem ser produzidas quer relativas às diferentes configurações do gráfico. Relativamente às estatísticas, eis as opções disponibilizadas:

- O tipo de evento de utilização que se quer analisar, i.e. *downloads*, depósitos ou consulta de metadados;
- Quais os repositórios que se pretendem consultar. O sistema permite agrupar vários repositórios e obter um gráfico comparativo;
- O período de tempo que pretendemos analisar, e.g., entre datas, último més, último ano, etc;
- O tipo de estatística que se pretende obter, i.e., evolução ou *ranking*.

O utilizador pode ainda configurar parâmetros relacionados com a apresentação gráfica das estatísticas, como é o caso das dimensões do gráfico, cores, título, entre outros. É ainda disponibilizada também a capacidade de exportação de dados em CSV para tratamento estatístico através de ferramentas especializadas e a subscrição do envio periódico, por correio-eletrônico, da estatística produzida.

Assim sendo e fazendo alusão ao modelo de software de três camadas, este módulo faz parte da camada de apresentação, assim como se pode ver na Figura 4.3;

dashboard Módulo que permite ao utilizador consultar um conjunto de estatísticas pré-definidas sobre os diferentes repositórios agregados pelo RCAAP. Este é especialmente útil para os administradores dos repositórios e/ou consórcio RCAAP na medida em que tem estatísticas já definidas, que habitualmente são necessárias para efeitos de *reporting*, sendo que podem ser imediatamente consultadas e/ou os seus dados exportadas para CSV.

Este módulo, e de forma análoga ao **workbench**, faz parte da camada de apresentação, num modelo de software de três camadas;

data generation service Módulo que disponibiliza um serviço *Web*, baseado em REST, que permite a geração de estatísticas em diferentes formatos (gráfico, CSV, entre outros). Este módulo faz parte da camada de serviços;

subscription service Módulo que disponibiliza um serviço *Web*, baseado em REST, que suporta todo o sistema de subscrições, i.e., criação, ativação, eliminação e a geração de estatísticas subscritas de forma periódica (sendo que a última não está acessível ao publico,

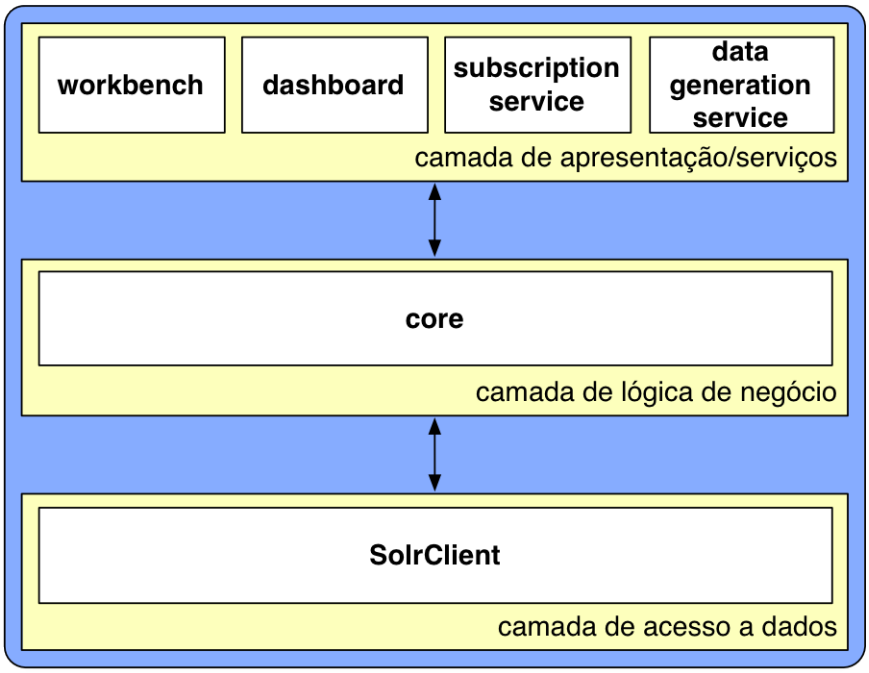


Figura 4.3: SCEUR como um modelo de software de três camadas

uma vez que é para uso interno). Este módulo, e de forma análoga ao **data generation service**, faz parte da camada de serviços;

SolrClient Módulo que permite ler e escrever do sistema de indexação Apache Solr [2]. Este módulo faz parte da camada de acesso a dados;

OAI-Stats Módulo necessário para que o DSpace consiga expor os seus dados de utilização através de OAI-PMH. Juntamente com esses dados, são enviados os metadados descritivos associados ao registo sobre o qual o evento ocorreu.

4.3 Desenvolvimento

Nesta secção pretende-se mostrar os aspetos mais práticos do desenvolvimento do SCEUR. Para tal começa-se por explicar algumas das tecnologias e ferramentas usadas. Depois abordam-se os diferentes módulos, como estes funcionam, fazendo também alusão às tecnologias e ferramentas usadas. Por último referem-se alguns dos problemas encontrados durante a fase de desenvolvimento.

[2] http://lucene.apache.org/solr/

4.3.1 Tecnologias

Com o objetivo de desenvolver um sistema centralizado de estatísticas, e tendo já analisado as suas características funcionais, surge agora a necessidade de decidir quais a tecnologias que foram usadas para desenvolver esse mesmo sistema. Trata-se de uma aplicação *Web*, desenvolvida em Java, recorrendo a um conjunto de ferramentas/tecnologias para assim conseguir disponibilizar as funcionalidades definidas.

Armazenamento/Pesquisa Apache Solr. Para armazenar os dados de utilização escolhemos o Apache Solr. O Solr é uma framework de pesquisa cujas principais funcionalidades são: pesquisa em texto integral, *faceted search*, pesquisa distribuída, entre outros. Este foi o sistema escolhido pois ele próprio é escrito em JAVA e também pelas inúmeras empresas/instituições que atualmente usam[3] o Solr como sistema de pesquisa e indexação.

A nível de organização, o Solr é composto por *indexes*, tendo cada *index* um conjunto de *documents* (podemos ver e comparando com uma base de dados relacional um *index* como uma tabela e um *document* como a linha de uma tabela). Um *document* é definido como um conjunto de *fields* (tendo um *field* um nome e um valor, assim como um conjunto de propriedades definidas no *schema* como são o caso do tipo (e.g., string, date, int, etc) e de propriedades mais técnicas relacionadas com a indexação, e.g., *indexed*, *stored*, *multiValued*, etc), sendo identificáveis de forma unívoca pelo seu identificador.

Geração de gráficos Google Chart API[4]. Este serviço da Google permite, através de uma interface REST, gerar gráficos em tempo-real, dispondo de um número variado de tipos de gráficos (linha, tarte, barras, etc) com um número bastante grande de configurações possíveis (largura, altura, cores, legendas, etc). Assim, e para cada gráfico de estatísticas a gerar, basta então gerar o URL para o serviço da Google que gera esse mesmo gráfico. Na Figura 4.4 podemos ver um exemplo de um gráfico de linha (cht=lc), com 600 píxeis de largura por 330 píxeis de altura (chs=600x330), com os valores 20 no primeiro ponto, 60 no segundo, 40 no terceiro e 50 no quarto (chd=t:20,60,40,50).

[3] http://wiki.apache.org/solr/PublicServers
[4] http://code.google.com/intl/pt-PT/apis/chart/image/

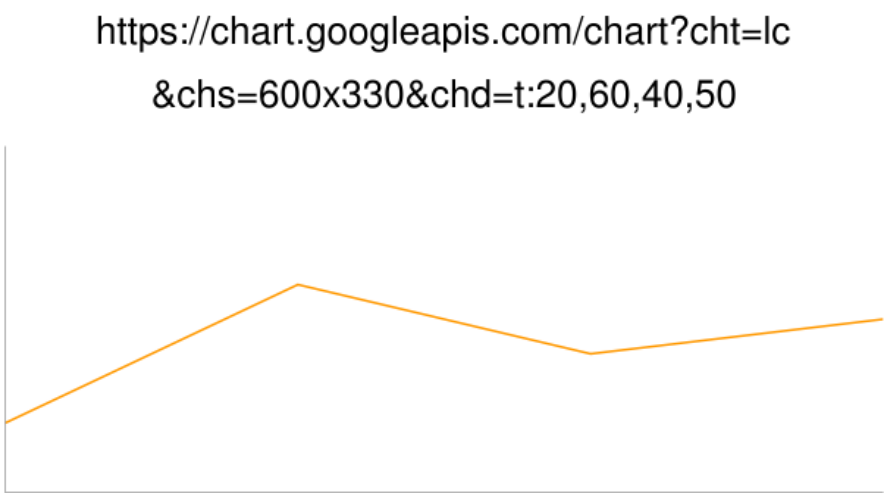


Figura 4.4: Exemplo de gráfico gerado pelo Google Chart API

4.3.2 Módulos

Nesta secção dá-se igual atenção a cada módulo no que diz respeito ao seu desenvolvimento e funcionamento, assim como aos processos mais relevantes do funcionamento do serviço centralizado de estatísticas.

core

Fazendo alusão ao nome, este módulo desempenha um papel "nuclear" no funcionamento do SCEUR. Para ser possível disponibilizar estatísticas, não basta saber "fazer contas", é necessário dados sobre os quais se vão gerar essas estatísticas. Nesse sentido, a ingestão de dados de utilização é fundamental. E por ingestão de dados de utilização entenda-se a recolha de dados de utilização dos diferentes repositórios, através do *harvester* OAI-PMH, dados que depois são analisados e processados para então serem indexados e assim permitir posterior consulta.

Na Figura 4.5 podemos ver o diagrama de comportamento desse processo, ainda que simplificado. De forma periódica este processo é executado com o objetivo de recolher os dados de utilização, tando dos repositórios que já foram agregados pelo menos uma vez (denominando esse processo de "*harvest* incremental") ou dos novos repositórios (denominando esse outro processo de "*harvest* total"), diferindo a nível prático na definição ou não do parâmetro *from* no pedido OAI-PMH *ListRecords*, caso seja incremental ou não. Assim, para cada repositório, enquanto o pedido *ListRecords* retornar uma lista não vazia, e para cada elemento dessa lista, faz-se o *unmarshall* (processo de transformar um elemento

XML em objetos Java) e insere-se no Solr para ser indexado, guardando-se também o evento num ficheiro de histórico, para consulta textual uma vez que no Solr a informação fica sobre a forma de *indexes*. Quando a lista retornada for vazia guarda-se a informação de *harvest* desse repositório, isto é, a data de termino do *harvest* e que será a data a partir da qual se irá recolher da próxima vez, e associado a isso muda-se o tipo de *harvest* de total para incremental.

workbench

Este módulo pretende ser a "bancada de trabalho", para os utilizadores que quiserem obter estatísticas. Assim, este modulo disponibiliza uma interface *Web* que "guia" o utilizador através das diferentes opções relacionadas com a geração de estatísticas, seja sobre a forma de gráfico seja sobre a forma de CSV. Assim como se pode ver na Figura 4.6, o **workbench** tem 5 secções: *Tipos de evento*, *Repositórios*, *Períodos de tempo*, *Tipo de estatísticas* e *Personalização do gráfico*.

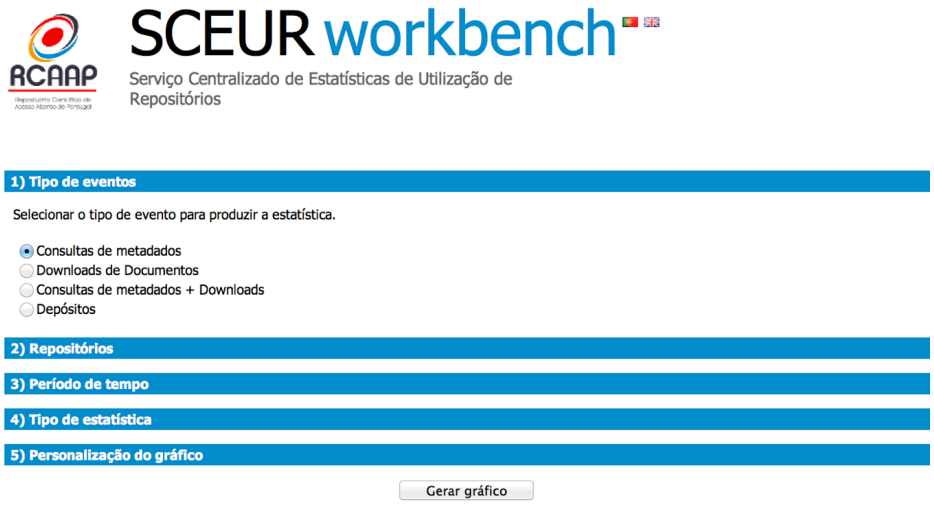


Figura 4.6: SCEUR workbench

Na primeira secção, o utilizador pode escolher o tipo de evento de utilização sobre o qual pretende obter a estatísticas. Na segunda secção pode escolher um ou mais repositórios sobre os quais quer obter a estatísticas, sendo que se o utilizador escolher mais que um repositório pode também escolher a cor relativa a cada repositório escolhido, assim como se pode ver na Figura 4.7.

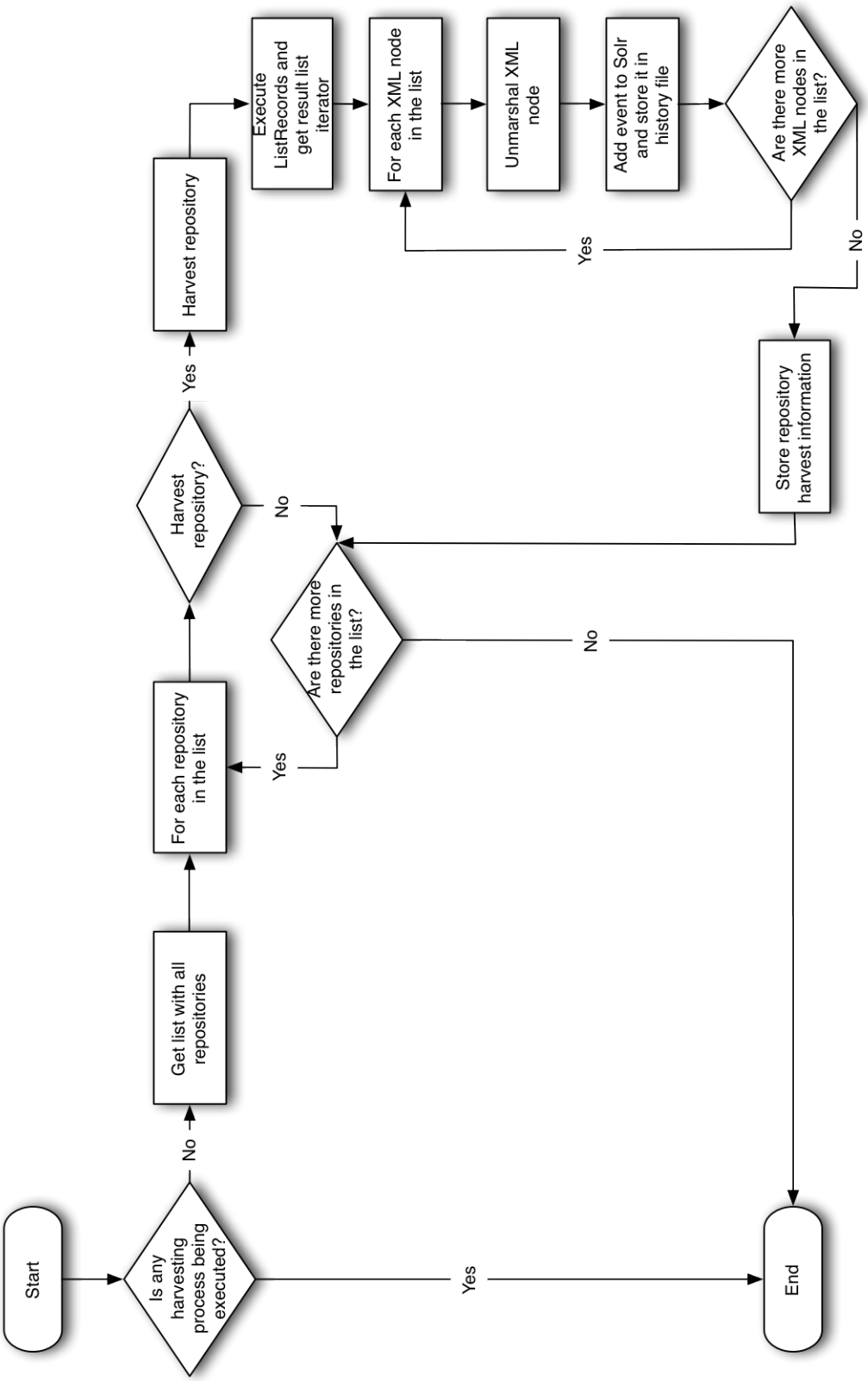


Figura 4.5: Diagrama de comportamento do processo de *harvest*

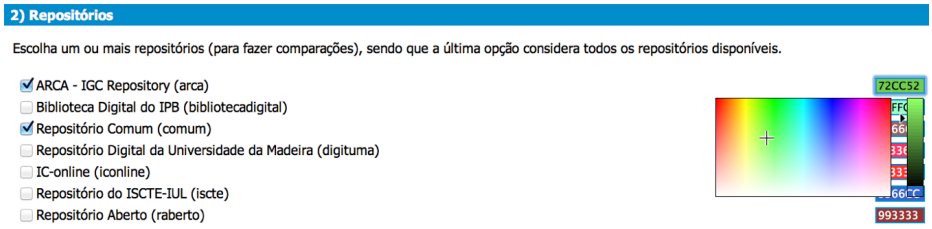


Figura 4.7: Seleção de repositórios e suas cores no SCEUR workbench

Na terceira secção pode escolher o período de tempo sobre o qual quer obter a estatística. Pode definir a data a partir do qual quer a estatística, a data até quando quer a estatística, as duas ou nenhum para obter todas a estatística de todos os anos que estão disponíveis. Pode ainda optar por escolher períodos de tempo pré-definidos, como é o caso de "Este ano", "Este mês", "Último mês", entre outros, assim como se pode ver na Figura 4.8.

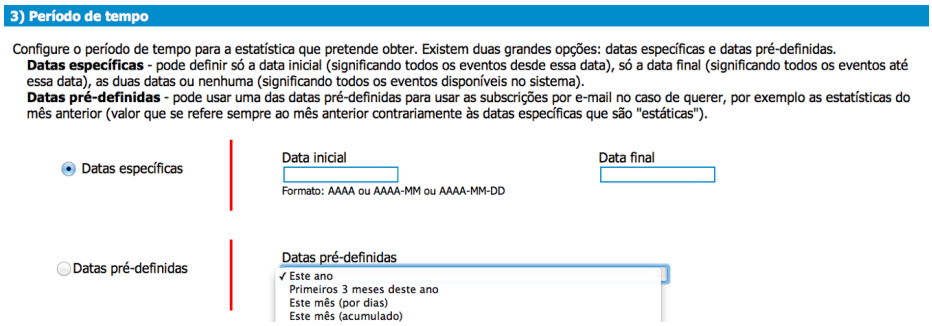


Figura 4.8: Seleção de período de tempo no SCEUR workbench

Na quarta secção pode escolher-se o tipo de estatísticas, separando em dois grupos: evolução e *ranking*. No primeiro grupo estão estatísticas de evolução, seja de forma genérica seja por documento (referenciado pelo seu identificador persistente) ou por autor (referenciado pelo nome). No segundo grupo estão as estatísticas de *ranking*, onde são apresentadas estatísticas *top* 10 referentes a um conjunto de características dos eventos de utilização como é o caso do *top* 10 documentos, autores, formatos, etc, assim como se pode constatar na Figura 4.9.

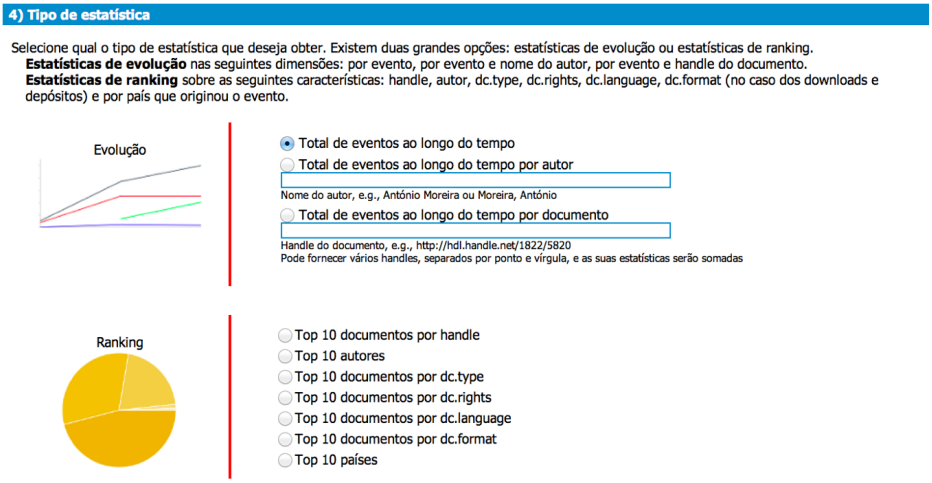


Figura 4.9: Seleção de tipo de estatística no SCEUR workbench

Na última secção pode configurar-se um conjunto de características mais técnicas no que diz respeito ao aspeto, como é o título, o tamanho (largura e altura), se o fundo deve ter um gradiente, entre outras.

dashboard

Este módulo disponibiliza um conjunto pré-definido de estatísticas, para consulta imediata. Pode-se ver na Figura 4.10 o seu aspeto.

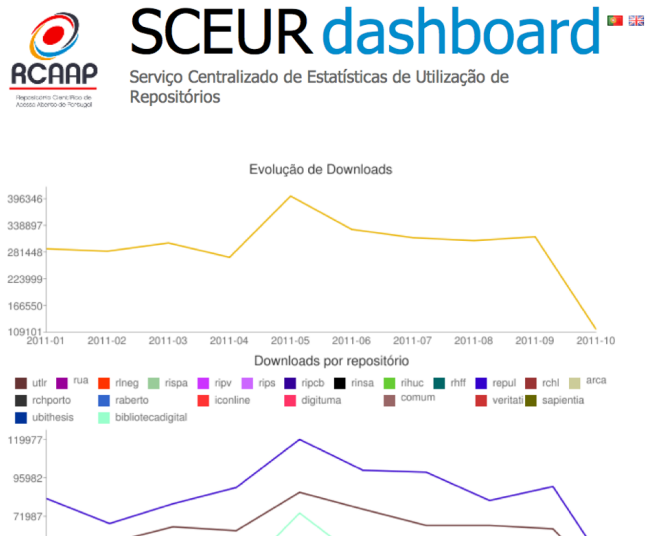


Figura 4.10: SCEUR dashboard

data generation service

Este módulo disponibiliza um serviço *Web* para a geração de estatísticas sobre a forma de gráficos e sobre a forma de CSV. Por ser o módulo responsável por exteriorizar as estatísticas sobre formas mais visíveis e fáceis de processar, explica-se agora o seu funcionamento, recorrendo ao diagrama de comportamento que se pode ver na Figura 4.11.

Sempre que o serviço é invocado, analisa os parâmetros e verifica se é para gerar um gráfico e esse gráfico já está em cache. Se isso for verdade, então gera o gráfico a partir da cache e retorna-o. Esta cache foi desenvolvida para otimizar a geração de gráficos previamente gerados e assim não adicionar carga desnecessária ao serviço. Caso contrário, valida os parâmetros (pois na primeira análise apenas se quer determinar se é para gerar o gráfico e se o mesmo está em cache). Se não estiverem válidos, mostra-se uma mensagem de erro, caso contrário gera-se a estatística pedida, isto é, os dados estatísticos. Se houver um erro ao gerar esses mesmos dados, mostra-se um erro, caso contrário gera-se o gráfico ou o ficheiro CSV, dependendo do caso.

subscription service

Este serviço faz o processamento de toda a informação relacionada com a subscrição de estatísticas. O funcionamento desse módulo está descrito no diagrama de comportamento que se pode observar na Figura 4.12. Como se pode ver, quando o serviço é invocado, este analisa os parâmetros para determinar de que tipo de ação se trata uma vez que este serviço também processa da geração de estatísticas periódicas por e-mail. Este serviço é invocado de forma periódica através de um *cronjob*, usando um endereço (*path*) ligeiramente diferente por questões de segurança, estando também protegido por nome de utilizador e palavra-chave. Se for o caso então gera as estatísticas e envia-as por e-mail, caso contrário o serviço determina se é uma das outras três ações possíveis, i.e., pedido de subscrição, ativação de subscrição ou então remoção de subscrição. Caso isso se verifique, os parâmetros são validados (e-mail, token, etc). Se forem válidos o serviço realiza a ação escolhida. Se acontecer algum erro, o utilizador é redirecionado para uma página de erro, caso contrário para uma página onde é confirmado o sucesso da ação.

SolrClient

Este módulo permite ler e escrever do Solr, servindo assim de interface entre o SCEUR e o Solr. Depois de perceber como o mesmo está es-

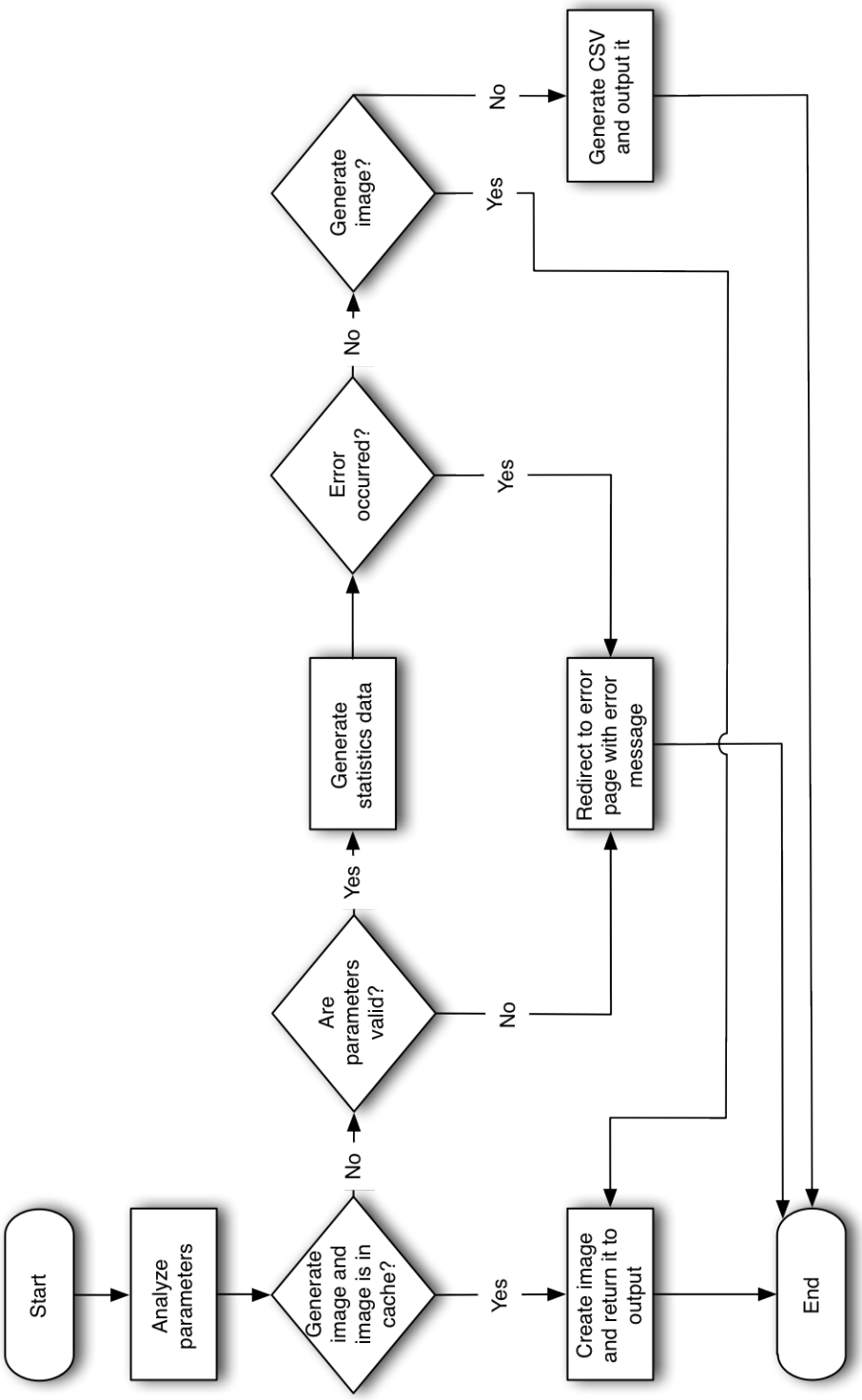


Figura 4.11: Diagrama de comportamento do *data generation service*

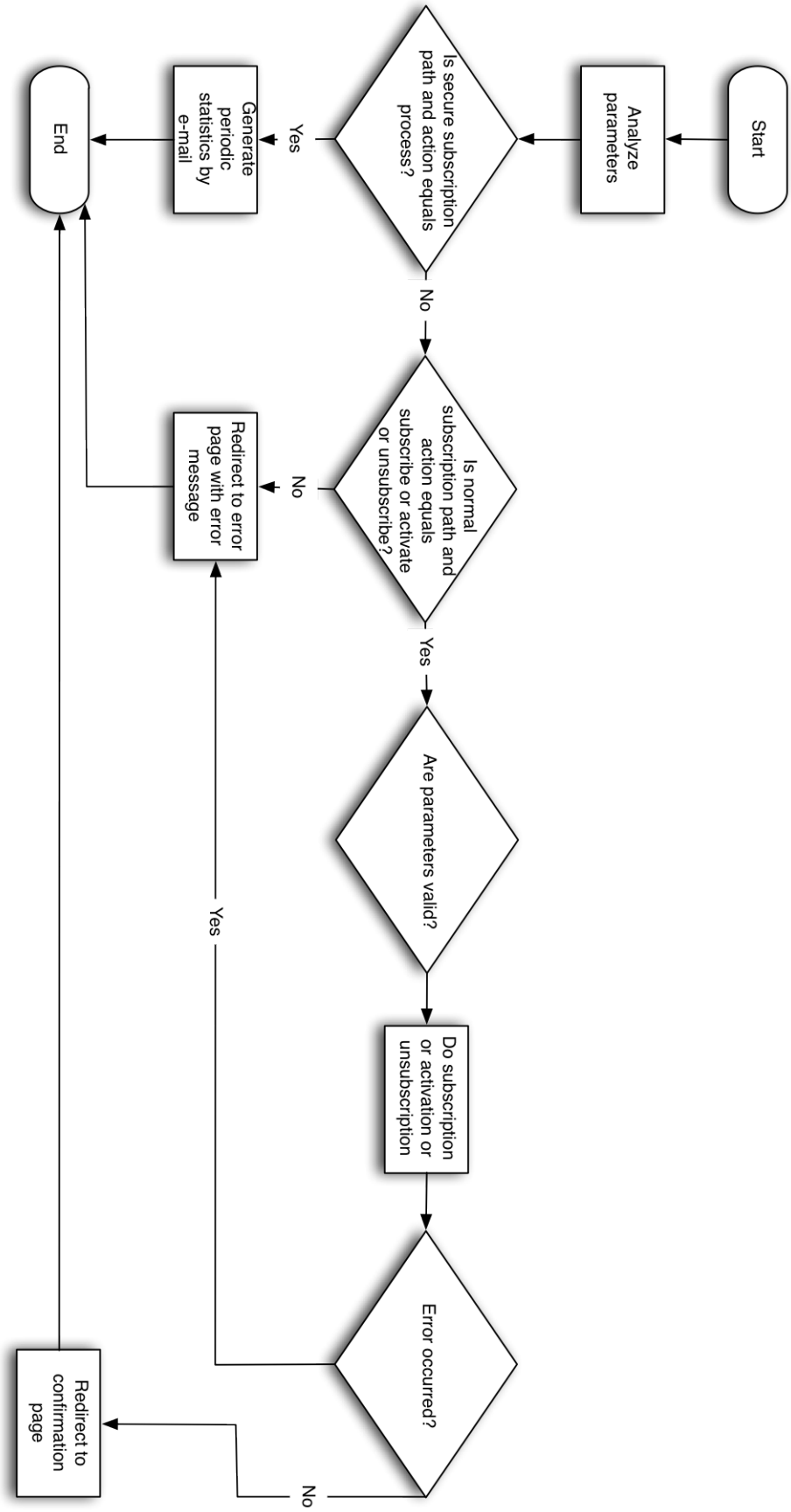


Figura 4.12: Diagrama de comportamento do *subscription service*

truturado (ver secção 4.3.1), e depois de se ter analisado e decidido os tipos de eventos de utilização que se iria partilhar, assim como os diferentes metadados que seriam também partilhados juntamente com cada evento, definiu-se então o *schema* do Solr, onde se explicita os *fields* que são necessários para acomodar a informação de cada evento, assim como se pode ver na Listagem 4.1 (versão simplificada para facilitar leitura). Nessa mesma listagem, pode ver-se os *fields* onde fica armazenada a informação que dá resposta às três questões referidas na secção 2.1.2, i.e., **O quê?**, **Quem?** e **Quando?**.

Relativamente ao **O quê?**, guarda-se o identificador do evento (event_id), o identificador persistente (handle) e o tipo de evento (event_type). Guarda-se ainda, no que diz respeito a metadados, o tipo de documento (dc_type), as condições de acesso (dc_rights), autor/autores (dc_authors), formato do documento (dc_format) e língua no qual foi escrito (dc_language).

No que diz respeito ao **Quem?**, guarda-se o endereço IP sobre a forma de uma *hash* por questões de privacidade (ip_hash) e o país de onde foi originado esse evento (country).

Por último e relativamente ao **Quando?** guarda-se a data em que o evento ocorreu (event_date). Existem ainda alguns campos (*fields*) para guardar informação interna do sistema centralizado de estatísticas, como é o caso da data em que foi recolhido (harvest_date) e o identificador do repositório ao qual pertence (repository_id).

```
<fields>
  <field name="event_id" type="string" />
  <field name="event_date" type="tdate" />
  <field name="harvest_date" type="tdate" />
  <field name="repository_id" type="string" />
  <field name="country" type="string" />
  <field name="handle" type="string" />
  <field name="ip_hash" type="string" />
  <field name="event_type" type="string" />
  <field name="dc_type" type="string" />
  <field name="dc_rights" type="string" />
  <field name="dc_authors" type="text_rev" />
  <field name="dc_format" type="string" />
  <field name="dc_language" type="string" />
</fields>
```

Listagem 4.1: Definição de *fields* no *schema* do Solr para acomodar informação de um evento de utilização

OAI-Stats

Este é o único módulo que não faz parte do serviço centralizado de estatísticas, uma vez que fica do lado dos repositórios, mas sem ele não

haveria dados de utilização. Este disponibiliza, através de OAI-PMH, os dados de utilização relativos a eventos de consulta de metadados, *downloads* e depósitos. Para tal, encapsula a informação a partilhar usando OpenURL ContextObjects no formato de metadados "CTXO", dando assim resposta às três questões referidas na secção 2.1.2, i.e., **O quê?**, **Quem?** e **Quando?**.

No que diz respeito ao **O quê?**, assim como se pode ver na Listagem 4.2, temos informação sobre o recurso associado ao qual foi registado o evento (http://comum.rcaap.pt/handle/123456789/12), o tipo de documento (artigo), o tipo de acesso (acesso-livre), a língua no qual foi escrito (inglês) e os seus autores (Joao Afonso e Pedro Veiga).

```
<ctx:identifier>http://comum.rcaap.pt/handle/123456789/12</
    ctx:identifier>
<ctx:metadata-by-val>
  <ctx:format>http://dublincore.org/documents/dcmi-terms/</ctx:format>
  <ctx:metadata>
    <dcterms:type>article</dcterms:type>
    <dcterms:rights>open access</dcterms:rights>
    <dcterms:language>eng</dcterms:language>
    <dcterms:author>Afonso, Joao</dcterms:author>
    <dcterms:author>Veiga, Pedro</dcterms:author>
  </ctx:metadata>
</ctx:metadata-by-val>
```

Listagem 4.2: Informação do recurso associado ao qual ocorreu o evento de utilização, presente num registo OAI-PMH no formato de metadados CTXO

Relativamente ao **Quem?** podemos ver um exemplo dessa informação na Listagem 4.3. Aí, podemos ver informação do endereço IP do utilizador que originou o evento sobre duas formas, i.e., sobre a forma de uma *hash* e sobre a forma de um endereço de rede onde apenas são apresentados os primeiros 24 *bits*, assim como é sugerido nas recomendações da KE [Verhaar and Vanderfeesten, 2010]. Podemos ver ainda a informação sobre o país de origem do evento de utilização (PT, i.e., Portugal).

```
<ctx:identifier>data:0a28242b48dea9247c25c1438ad4a571</ctx:identifier>
<ctx:identifier>data:193.136.44.0</ctx:identifier>
<ctx:metadata-by-val>
  <ctx:format>http://dublincore.org/documents/dcmi-terms/</ctx:format>
  <ctx:metadata>
    <dcterms:spatial>PT</dcterms:spatial>
  </ctx:metadata>
</ctx:metadata-by-val>
```

Listagem 4.3: Informação sobre o utilizador associado ao evento de utilização, presente num registo OAI-PMH no formato de metadados CTXO

Por último, e no que diz respeito a **Quando?**, está presente a informação sobre o instante de tempo em que o evento ocorreu, informação essa presente no atributo *timestamp* do elemento XML do OpenURL ContextObject, assim como se pode ver na Listagem 4.4.

```
<ctx:context-object identifier="oai:comum.rcaap.pt:download_64"
    timestamp="2009-11-02T16:59:02Z">
...
</ctx:context-object>
```

Listagem 4.4: Informação sobre o instante de tempo em que o evento de utilização ocorreu, presente num registo OAI-PMH no formato de metadados CTXO

No anexo B pode consultar-se o registo completo usado como exemplo nas listagens 4.2, 4.3 e 4.4, onde se pode ver a estrutura e toda a informação presente num registo OAI-PMH no formato de metadados "CTXO". No que diz respeito aos tipos de eventos de utilização partilhados, i.e., consulta de metadados, *downloads* e depósito, este módulo disponibiliza primeiro os de consulta e os de *download*, pois estão relacionados e armazenados em duas tabelas usadas pelo add-on *MINHO STATS*, e depois os de depósito, que são obtidos diretamente dos registos analisando para isso a sua data de depósito (dc.date.available).

4.3.3 Problemas Encontrados

Nesta secção são referidos alguns problemas e complicações que surgiram no desenvolvimento do Serviço Centralizado de Estatísticas de Utilização de Repositórios, sendo que alguns tiveram resolução durante a execução da dissertação e outros serão resolvidos em trabalho futuro.

Google Chart API

O Google Chart API disponibiliza um serviço REST para geração de gráficos, podendo esse serviço ser invocado através de um pedido HTTP GET ou POST. Inicialmente começou-se por usar pedidos baseados em GET, pois assim o SCEUR apenas precisava formar o URL para o serviço e usar esse mesmo URL no **data generation service** para disponibilizar o gráfico para embeber em sítios *Web*, sendo que seria então a Google a "servir" a imagem. Contudo esse tipo de pedido tem um problema: o URL não pode ter mais de 2000 carateres. E isso constituía um problema, na medida em que alguns gráficos por causa dos valores, cores, legendas e título ultrapassavam com facilidade esse valor.

A **solução** encontrada foi usar pedidos POST, que permite pedidos com um máximo de 16000 carateres. Mas esta solução, apesar de resolver o problema inicial, tem uma desvantagem. Para embeber em sítios externos não se pode disponibilizar código HTML para submeter o POST, pois este teria que se basear também em Javascript (fazer o *submit* do POST para obter a imagem), algo que pode estar inativo por vontade do utilizador. Nesse sentido, tem que ser o SCEUR a fazer o POST e assim "servir" a imagem.

Dados de utilização - Perda de eventos

O add-on desenvolvido para o DSpace partilha eventos de utilização relativos a consultas de metadados, *downloads* e depósitos, disponibilizando primeiro os de consulta e *download*, pois estão relacionados e armazenados em duas tabelas usadas pelo add-on MINHO STATS ou no Solr caso use o sistema de estatísticas do DSpace, e depois os de depósito, que são obtidos diretamente dos registos analisando para isso a sua data de depósito (dc.date.available).

Habitualmente, os metadados disponibilizados através de OAI-PMH estão sempre relacionados ou armazenados num mesmo sítio, e como tal é fácil iterar sobre eles e assim garantir que se disponibilizam todos. Contudo, devido às característica da informação que se quer partilhar neste caso (três eventos de utilização), podem perder-se alguns eventos de consulta e *download* pois a data do último *harvest* é definida no fim de recolher, neste caso, todos os depósitos, e como tal os eventos de consulta e *download* que possam ter ocorrido enquanto se obtinham os de depósito podem perder-se.

Uma **possível solução** pode passar por associar a cada evento de utilização um *set*, e agregar de forma independente por *set*, associando a cada *set* uma data de último *harvest*.

Dados de utilização - Registos eliminados

O add-on desenvolvido para o DSpace, quando conjugado com o add-on MINHO STATS, origina um problema no que diz respeito a registos eliminados que têm dados de utilização associados, problema esse que tanto tem de filosófico como de prático. O que acontece é que o DSpace regista os dados de utilização, i.e., consulta de metadados, *downloads*, etc, porem quando os registos são eliminados, esses dados não são apagados. Contudo, o DSpace com o seu sistema próprio de estatísticas elimina esses dados de utilização, o que origina vários problemas.

A nível filosófico, não parece que haja um certo ou errado no facto de eliminar dados de utilização de registos eliminados pois por um lado pode alegar-se que se o registo foi eliminado não faz sentido ter estatísticas dele, se bem que por outro lado pode alegar-se que apesar do registo ter sido eliminado, esse utilização aconteceu e como tal faz parte do repositório. A nível prático isto origina problemas de coerência entre os valores estatísticos disponibilizados pelos sistema centralizado e os valores disponibilizados pelos repositórios, pois do lado do repositório os dados de utilização existem mas como o registo está eliminado não se tem acesso à informação mínima necessária, uma vez que habitualmente apenas se guarda o identificador único do registo (usando-o para aceder à sua informação), e como tal não pode ser partilhado através de OAI-PMH.

ContextObject

Seguindo as recomendações da KE [Verhaar and Vanderfeesten, 2010], usa-se os OpenURL ContextObjects como "contentores" para transportar os dados de utilização em OAI-PMH, associando a isso o formato de metadados "CTXO", com o objetivo também de acompanhar os projetos europeus. Contudo, no XML as anotações dos elementos causam um grande *overhead* na quantidade de informação transmitida, informação essa que não é útil a nível estatístico. Como tal, há um problema de eficiência.

Uma **possível solução** seria criar um novo formato de metadados que encapsulasse a informação, por exemplo, em JSON, que tem bastante menos *overhead* ao nível da informação transmitida, se bem que por outro lado como não tem XML Schemas associados teria que se arranjar outra forma de fazer a validação.

4.4 Resultados

Nesta secção são apresentados alguns resultados obtidos, referentes à execução do projeto. Nesses resultados engloba-se tanto informação prática e técnica acerca da agregação de dados de utilização dos diferentes repositórios, assim como algumas das estatísticas que se podem obter e das quais se consegue tirar algumas ilações no contexto do acesso-livre e da produção científica.

4.4.1 Agregações

Tendo o projeto RCAAP começado em 2008, e estando alguns dos repositórios já em funcionamento antes desse período, seria expectável que a agregação inicial seria mais demorada, na medida em que envolvia recolher todos os dados de utilização armazenados nos repositórios. Assim, apresentam-se na Tabela 4.1 os dados referentes às agregações iniciais dos 23 repositórios agregados pelo SCEUR (repositórios identificados pelos seus acrônimos).

Repositório	Qtd. de eventos	Tempo demorado	Eventos/Segundo
arca	72.548	814 segundos	89
bibliotecadigital	1.391.197	28.980 segundos	48
comum	114.457	2.077 segundos	55
digituma	98.401	1.315 segundos	74
iconline	241.594	4.691 segundos	51
iscte	1.479.097	37.244 segundos	39
raberto	1.473.395	47.466 segundos	31
rchl	18.394	354 segundos	51
rchporto	73.117	1.208 segundos	60
repul	1.351.096	31.367 segundos	43
rhff	77.084	1.220 segundos	63
rinsa	4.601	96 segundos	47
ripcb	187.076	4.006 segundos	46
rips	124.681	2.504 segundos	49
ripv	123.172	2.811 segundos	43
rispa	153.385	3.078 segundos	49
rlneg	207.837	4.575 segundos	45
rua	476.606	9.468 segundos	50
sapientia	139.976	2.160 segundos	64
ubithesis	42.410	955 segundos	44
utlr	1.762.335	43.492 segundos	40
veritati	379.670	7.924 segundos	47
Total	9.992.129	237.805 segundos	42

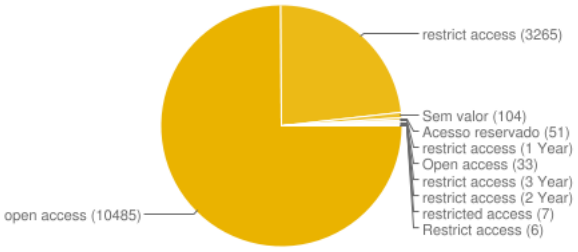
Tabela 4.1: Agregação inicial do SCEUR

No que diz respeito às agregações que são feitas periodicamente (numa base diária), e considerando os dados dos últimos cinco dias (à data da escrita), consegue-se determinar que em média são agregados 31.150 eventos demorando em média 10 minutos e 30 segundos, o que representa uma média de 49 eventos por segundo.

4.4.2 Estatísticas

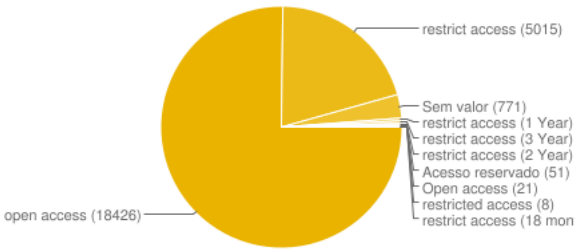
Estando o SCEUR inserido na iniciativa RCAAP, e sendo o RCAAP uma iniciativa em que um dos seus principais objetivos é promover o livre acesso à informação, usa-se essa temática para exemplificar algumas das estatísticas disponibilizadas pelo SCEUR e que ilações se podem tirar nesse contexto.

Depósitos em Open Access O SCEUR permite determinar, para um determinado instante de tempo, quais as *top* 10 políticas de acesso dos documentos depositados nos diferentes repositórios. Assim, nas Figuras 4.13, 4.14 e 4.15 podemos ver com alguma clareza e apesar de o ano de 2011 ainda não ter terminado que ao longo destes três últimos anos a quantidade e consequente percentagem de documentos depositados em acesso-livre (*open access*) aumentou em detrimento de políticas de acesso com restrições (*restrict access* para acesso restrito sem termo ou acesso restrito com termo de um, dois ou três anos).



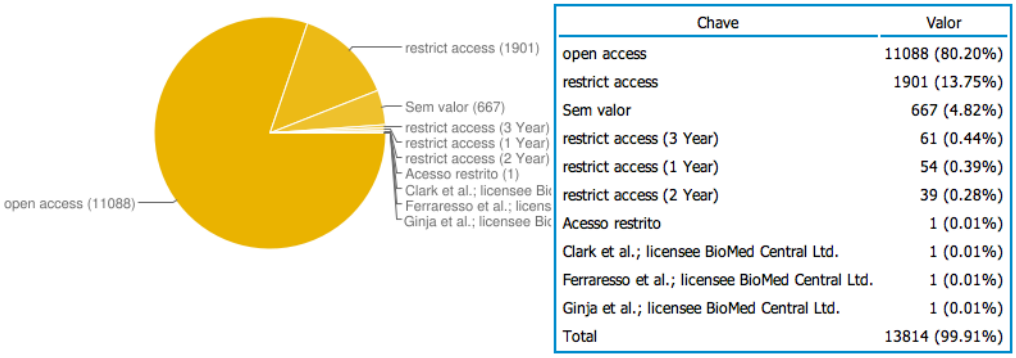
Chave	Valor
open access	10485 (74.61%)
restrict access	3265 (23.23%)
Sem valor	104 (0.74%)
Acesso reservado	51 (0.36%)
restrict access (1 Year)	36 (0.26%)
Open access	33 (0.23%)
restrict access (3 Year)	31 (0.22%)
restrict access (2 Year)	21 (0.15%)
restricted access	7 (0.05%)
Restrict access	6 (0.04%)
Total	14039 (99.90%)

Figura 4.13: Top 10 políticas de acesso relativas a depósitos (2009)



Chave	Valor
open access	18426 (75.05%)
restrict access	5015 (20.43%)
Sem valor	771 (3.14%)
restrict access (1 Year)	89 (0.36%)
restrict access (3 Year)	88 (0.36%)
restrict access (2 Year)	56 (0.23%)
Acesso reservado	51 (0.21%)
Open access	21 (0.09%)
restricted access	8 (0.03%)
restrict access (18 months)	4 (0.02%)
Total	24529 (99.91%)

Figura 4.14: Top 10 políticas de acesso relativas a depósitos (2010)



Chave	Valor
open access	11088 (80.20%)
restrict access	1901 (13.75%)
Sem valor	667 (4.82%)
restrict access (3 Year)	61 (0.44%)
restrict access (1 Year)	54 (0.39%)
restrict access (2 Year)	39 (0.28%)
Acesso restrito	1 (0.01%)
Clark et al.; licensee BioMed Central Ltd.	1 (0.01%)
Ferraresso et al.; licensee BioMed Central Ltd.	1 (0.01%)
Ginja et al.; licensee BioMed Central Ltd.	1 (0.01%)
Total	13814 (99.91%)

Figura 4.15: Top 10 políticas de acesso relativas a depósitos (2011)

Estes valores não surgem com grande espanto precisamente por causa do esforço que o projeto RCAAP tem feito junto das instituições de ensino e de investigação, no sentido de criar políticas de auto-arquivo e arquivo em acesso-livre.

Evolução dos *downloads* e consultas de metadados Nas diferentes alturas do ano, há diferentes utilização dos repositórios, no que diz respeito por exemplo a *downloads* e consulta de metadados. Mas quais serão essas alturas? Será transversal a todos os repositórios? É precisamente isso que podemos ver na Figura 4.16. Podemos ver que no mês de Julho a utilização é menor e apesar de ter diferentes dimensões nos diferentes repositórios pode ver-se que essa menor utilização é transversal a todos os repositórios.

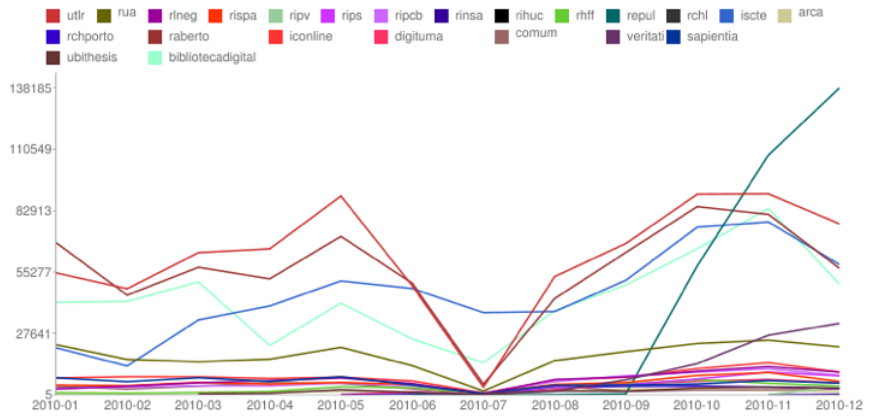
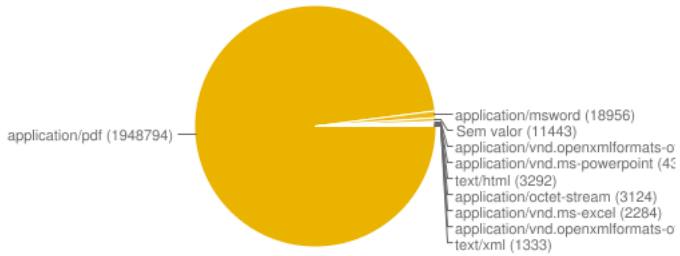


Figura 4.16: Evolução dos downloads e consultas de metadados (2010)

Formato de ficheiro mais descarregado Esta estatística permite saber quais são, para assim se tomar medidas no que diz respeito às

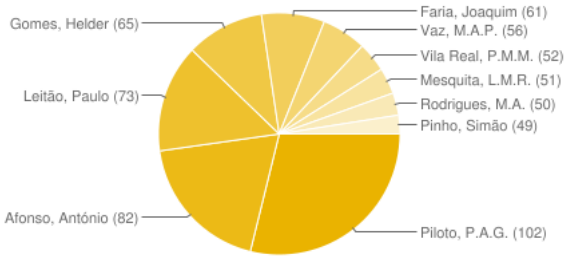


Chave	Valor
application/pdf	1948794 (97.31%)
application/msword	18956 (0.95%)
Sem valor	11443 (0.57%)
application/vnd.openxmlformats-officedocument.wordprocessingml.document	4548 (0.23%)
application/vnd.ms-powerpoint	4347 (0.22%)
text/html	3292 (0.16%)
application/octet-stream	3124 (0.16%)
application/vnd.ms-excel	2284 (0.11%)
application/vnd.openxmlformats-officedocument.presentationml.presentation	1648 (0.08%)
text/xml	1333 (0.07%)
Total	1999769 (99.86%)

Figura 4.17: Top 10 formatos de ficheiro mais descarregados (2010)

funcionalidades oferecidas sobre esses tipos de ficheiros no contexto do repositório. Habitualmente os repositórios disponibilizam pouca informação/funcionalidades, quando podiam por exemplo e associado aos PDFs disponibilizar visualizadores, assim como informação extra para além do básico que é nome e tamanho do ficheiro como é o caso da quantidade de páginas, a ferramenta usada para criar o ficheiro e outros metadados que normalmente se adiciona a esses ficheiros. Na Figura 4.17 pode constatar-se que os formatos de ficheiros mais descarregados em 2010 foram PDF e Microsoft Word.

Autor com mais depósitos No que diz respeito à produção científica, a quantidade de documentos produzidos influencia a avaliação de um investigador no contexto de uma instituição de ensino, consoante este produza mais ou menos. No contexto de um repositório é simples obter essa métrica, contudo num consórcio de 30 repositório a tarefa é um pouco mais complicada. Assim como se pode ver na Figura 4.18, em 2010 os dez autores com mais depósitos correspondem a seis porcento de todos os depósito efetuados, com um total de 641 depósitos.



Chave	Valor
Piloto, P.A.G.	102 (0.95%)
Afonso, António	82 (0.76%)
Leitão, Paulo	73 (0.68%)
Gomes, Helder	65 (0.61%)
Faria, Joaquim	61 (0.57%)
Vaz, M.A.P.	56 (0.52%)
Vila Real, P.M.M.	52 (0.48%)
Mesquita, L.M.R.	51 (0.48%)
Rodrigues, M.A.	50 (0.47%)
Pinho, Simão	49 (0.46%)
Total	641 (5.98%)

Figura 4.18: Top 10 autores com mais depósitos (2010)

Estas são apenas algumas das estatísticas disponibilizadas pelo SCEUR, que podem ser simples estatísticas assim como auxiliares no processo de decisão e avaliação no contexto acadêmico-científico.

Capítulo 5

Conclusões e Trabalho Futuro

Neste capítulo apresentam-se algumas conclusões, abordando também alguns assuntos que merecem alguma discussão e deixam ideias para trabalho futuro.

5.1 Conclusões e Discussão

O SCEUR é uma ferramenta ideal para auxiliar os vários intervenientes no processo de comunicação científica a avaliar/medir o impacto do seu trabalho, sendo uma alternativa mais prática e imediata às metodologias até então praticadas que se baseiam em citações. Baseado em experiências internacionais, este foi desenhado para responder às necessidades dos utilizadores nacionais, tendo atingido um estado de maturidade que lhe permite estar já em funcionamento[1]. Assim, este serviço passa a integrar o rol de ofertas disponibilizadas à comunidade científica nacional. Tendo noção de que as necessidades dos utilizadores irão evoluir, o SCEUR está preparado para evoluir com eles e ser adaptado a novos contextos de utilização.

Nesse sentido, e estando ciente da base que se construiu, i.e., uma arquitetura que permite agregar, processar e disponibilizar estatísticas de utilização, há ainda alguns assuntos que merecem alguma discussão:

Normalização Por normalização entende-se todo o trabalho que é feito aos dados de utilização com o objetivo de obter estatísticas mais fidedignas e equiparáveis. Se por um lado um dos objetivos do SCEUR é disponibilizar add-ons que ajudem a partilhar os dados de utilização, nem sempre se consegue controlar a informação partilhada pelos repositórios e muito menos a sua "qualidade". Nesse

[1] http://sceur.rcaap.pt

contexto, pode falar-se em normalização no serviço centralizado no que diz respeito a *robots*, *double clicks* e metadados.

Relativamente aos *robots*, a ideia é garantir que os dados de utilização sobre os quais se irão disponibilizar estatísticas são de origem humana, sendo já várias as fontes onde se pode obter de forma atualizada informação sobre os mesmos. No que diz respeito aos *double clicks*, i.e., a quantidade de tempo entre duas ações iguais por parte de um mesmo utilizador para que se considere uma apenas, pretende-se precisamente determinar a quantidade de tempo a considerar para efeitos de normalização. Com esta técnica de normalização tenta-se prevenir que possíveis adulterações dos dados de utilização tenham resultados, i.e., que através de ações repetidas se consiga inflacionar a quantidade de consultas de metadados ou de *downloads* de um determinado registo. Neste aspeto, o problema é que quantidade de tempo se deve considerar, pois segundo o documento da KE [Verhaar and Vanderfeesten, 2010] diferentes projetos/consórcios usam diferentes quantidades de tempo.

Por último, e no que diz respeito a metadados, falta decidir que tipo de processamento se deve dar. Com esse processamento pretende-se garantir que todos os metadados recolhidos dos diferentes repositórios se podem comparar. Assim como se pode ver na Figura 4.13 existem políticas de acesso iguais identificadas por termos diferentes, como é o caso de "open access" e "Open access", resultado da não conformidade/inadequação de um determinado repositório no que diz respeito ao conjunto de termos a usar para identificar as políticas de acesso.

Nome dos autores Um dos metadados recolhidos contém o nome dos autores do registo associado ao qual ocorreu o evento, sendo que depois se disponibiliza estatísticas sobre esses mesmos nomes. Contudo, a informação que recolhemos dos repositórios é texto apenas, não permitindo de uma qualquer forma identificar de forma unívoca um autor uma vez que pode haver dois ou mais autores cujos nomes, na forma abreviada de primeiro e último nome, sejam iguais. No contexto nacional poderia usar-se a informação já disponibilizada pelo DeGois[2], que é um sistema que organiza o curriculum dos investigadores portugueses concentrando informação do próprio investigador e da sua produção intelectual, porem este só teria validade/utilidade no contexto nacional, não resolvendo assim um problema já sentido por todo o mundo. Há já algumas iniciativas internacionais que de forma simples pretendem criar iden-

[2]http://www.degois.pt/

tificadores persistentes a nível dos utilizadores, com o objetivo de esses identificadores terem abrangência e validade global.

De SCEUR para SCER Com o aumento das necessidades estatísticas no contexto nacional, quer a nível de tipos de estatísticas que se pretende quer a nível de formatos em que se disponibilizam essas mesmas estatísticas, o "Serviço Centralizado de Estatísticas de Utilização de Repositórios" pode passar a "SCER - Serviço Centralizado de Estatísticas de Repositórios", através da generalização do conceito de dado de utilização para uma qualquer coisa que se pretende usar para fins estatísticos. Com isto se quer dizer que o serviço centralizado de estatísticas pode ser usado em outros contextos através da redefinição de alguns conceitos, nomeadamente o de evento e repositório, e do processamento que se dá aos dados que chegam ao sistema para efeitos estatísticos, tornando-se assim um sistema versátil e adaptável.

5.2 Trabalho Futuro

O *roadmap* do SCEUR inclui os seguintes itens:

Agregação de outros tipos de eventos de utilização ou informação associada

Com a evolução das plataformas de suporte à implementação de repositórios institucionais e com a necessidade sempre crescente de informação por parte dos utilizadores, o SCEUR terá de agregar mais eventos de utilização e mais informação associada. Um dos "eventos" que se irá disponibilizar será "publicação", sendo este e de forma análoga ao depósito informação estatística extraída dos registos pela análise de uma propriedade que é a data mas em vez de usar a data de depósito (dc.data.available) irá usar-se a data de publicação (dc.date.issued). Outro "evento" será definido para partilhar a quantidade de registos que o repositório tem no instante de tempo que é agregado através de OAI-PMH, permitindo assim no sistema centralizado ter a percepção da evolução da quantidade de registos de um ou mais repositórios.

Otimizações ao add-on e ao serviço centralizado

Há sempre espaço a melhorias, quer a nível de processamento, quer a nível de uso dos recursos de rede. Terá de se fazer uma análise para

determinar as otimizações necessárias e a melhor maneira de as fazer.

Desenvolvimento de add-ons para outras plataformas

Está previsto o desenvolvimento de um novo add-on, desta vez, para a plataforma de gestão de arquivos definitivos DigitArq[3] implementada em toda a rede de Arquivos Distritais da Direção-Geral de Arquivos[4].

SCEUR como *Data Provider* OAI-PMH e *Data Provider* SUSHI

Uma vez que existem no contexto internacional iniciativas semelhantes, será fundamental que o SCEUR seja capaz de partilhar a informação estatística de utilização dos repositórios a si associados, quer via OAI-PMH quer via SUSHI.

[3] https://www.keep.pt/node/151
[4] http://dgarq.gov.pt/

Bibliografia

[associação para a promoção da sociedade da informação, 2007] associação para a promoção da sociedade da informação (2007). Glossário da sociedade da informação. Technical report, associação para a promoção da sociedade da informação.

[Baptista, 2010] Baptista, A. A. (2010). A falar nos entendemos : a interoperabilidade entre repositórios digitais. In *Repositórios institucionais : democratizando o acesso ao conhecimento*, pages 71–90. EDUFBA.

[Bell et al., 2005] Bell, S., Foster, N. F., and Gibbons, S. (2005). Reference librarians and the success of institutional repositories. *Reference Services Review*, 33:283–290.

[Bollen and Van de Sompel, 2006] Bollen, J. and Van de Sompel, H. (2006). An architecture for the aggregation and analysis of scholarly usage data. In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '06, pages 298–307, New York, NY, USA. ACM.

[Bollen et al., 2005] Bollen, J., Van de Sompel, H., Smith, J. A., and Luce, R. (2005). Toward alternative metrics of journal impact: A comparison of download and citation data. *Inf. Process. Manage.*, 41:1419–1440.

[COMMUNITIES, 2007] COMMUNITIES, C. O. T. E. (2007). Communication from the commission to the european parliament, the council and the european economic and social committee: on scientific information in the digital age: access, dissemination and preservation.

[Feijen et al., 2007] Feijen, M., Horstmann, W., Manghi, P., Robinson, M., and Russell, R. (2007). Driver: Building the network for accessing digital repositories across europe. *Ariadne*.

[Ferreira et al., 2008] Ferreira, M., Rodrigues, E., Baptista, A. A., and Saraiva, R. (2008). Carrots and sticks : some ideas on how to create a successful institutional repository. *D-Lib Magazine*, 14(1/2).

[Fowler, 2002] Fowler, M. (2002). *Patterns of enterprise application architecture*. Addison-Wesley Professional.

[Kahn and Wilensky, 2006] Kahn, R. and Wilensky, R. (2006). A framework for distributed digital object services. *International Journal on Digital Libraries*, 6:115–123. 10.1007/s00799-005-0128-x.

[Lagoze and de Sompel, 2001] Lagoze, C. and de Sompel, H. V. (2001). The open archives initiative: Building a low-barrier interoperability framework. *Digital Libraries, Joint Conference on*, 0:54–62.

[Metje and Hilse, 2009] Metje, D. and Hilse, H.-W. (2009). Specification: Data format and exchange for oa statistics. Technical document, version 0.5.

[Needleman, 2006] Needleman, M. H. (2006). The niso standardized usage statistics harvesting initiative (sushi). *Serials Review*, 32(3):216 – 217.

[Organization, 2004] Organization, N. I. S. (2004). Ansi/niso z39.88 - the openurl framework for context-sensitive services. Technical report.

[Pauwels, 2009] Pauwels, B. (2009). Neeo technical guidelines for the exchange of usage metadata. Technical document, version 1.4.

[Rodrigues, 2005] Rodrigues, E. (2005). O repositórium e a política de auto-arquivo da universidade do minho.

[Rodrigues et al., 2010] Rodrigues, E., Saraiva, R., Ribeiro, C., and Fernandes, E. M. (2010). Os repositórios de dados científicos : estado da arte. Technical report.

[Smith et al., 2003] Smith, M., Barton, M., Bass, M., Branschofsky, M., McClellan, G., Stuve, D., Tansley, R., and Walker, J. H. (2003). Dspace: An open source dynamic digital repository. *D-Lib Magazine*, 9(1).

[Van de Sompel and Beit-Arie, 2001] Van de Sompel, H. and Beit-Arie, O. (2001). Open linking in the scholarly information environment using the openurl framework. *New Review of Information Networking*, 7:59–76.

[Van de Sompel and Hochstenbach, 1999a] Van de Sompel, H. and Hochstenbach, P. (1999a). Reference linking in a hybrid library environment. part 1: Frameworks for linking. *D-Lib Magazine*, 5.

[Van de Sompel and Hochstenbach, 1999b] Van de Sompel, H. and Hochstenbach, P. (1999b). Reference linking in a hybrid library environment. part 2: Sfx, a generic linking solution. *D-Lib Magazine*, 5.

[Van de Sompel and Hochstenbach, 1999c] Van de Sompel, H. and Hochstenbach, P. (1999c). Reference linking in a hybrid library environment. part 3: Generalizing the sfx solution in the sfxghent & sfxlanl experiment. *D-Lib Magazine*, 5.

[Verhaar and Vanderfeesten, 2010] Verhaar, P. and Vanderfeesten, M. (2010). Ke usage statistics guidelines: Guidelines for the aggregation and exchange of usage data. Technical report, SURFfoundation.

[Wikipedia, 2011] Wikipedia (2011). Workflow — wikipedia, the free encyclopedia. [Em linha; acedido em 2011-10-03].

Apêndice A

Análise aos eventos de utilização

Informação referente à análise feita aos eventos de utilização de 17 repositórios em regime de SARI até à data de 2011-01-12. São apresentadas as contagens feitas às 2 tabelas que contêm informação sobre utilização (consulta de metadados e *downloads*), tabelas essas usadas pelo add-on *MINHO STATS*.

As contagens de cada repositório são representadas por "sari_rcaap_ACRONIMO", acrônimo esse que pode ser usado para consultar a informação do repositório presente no portal RCAAP usando o URL: http://www.rcaap.pt/repositoryInfo.jsp?id=ACRONIMO

downloads	942849
views	1036090
total	1978939

Tabela A.1: Resultado sumário da análise de eventos de utilização de 17 SARIs

sari_rcaap_rinsa	downloads	133
	views	311
	total	444
sari_rcaap_iconline	downloads	85238
	views	98358
	total	183596
sari_rcaap_ripcb	downloads	50405
	views	48442
	total	98847
sari_rcaap_rlneg	downloads	69638
	views	52047
	total	121685
sari_rcaap_ripv	downloads	1794
	views	4083
	total	5877
sari_rcaap_veritati	downloads	39803
	views	53890
	total	93693
sari_rcaap_digituma	downloads	18377
	views	11140
	total	29517
sari_rcaap_arca	downloads	7249
	views	25442
	total	32691
sari_rcaap_rips	downloads	17841
	views	21994
	total	39835
sari_rcaap_rchporto	downloads	9174
	views	10631
	total	19805
sari_rcaap_bibliotecadigitalipb	downloads	305997
	views	471629
	total	777626
sari_rcaap_repul	downloads	234861
	views	102991
	total	337852
sari_rcaap_rhff	downloads	26381
	views	23931
	total	50312
sari_rcaap_ubithesis	downloads	11213
	views	14082
	total	25295
sari_rcaap_comum	downloads	25168
	views	50427
	total	75595
sari_rcaap_rispa	downloads	39577
	views	46692
	total	86269

Tabela A.2: Resultado detalhado da análise de eventos de utilização de 17 SARIs

Apêndice B

Registo OAI-PMH (CTXO)

```
<?xml version="1.0" encoding="UTF-8"?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://www.openarchives.org/OAI
    /2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd"
    >
  <responseDate>2011-09-30T14:38:11Z</responseDate>
  <request metadataPrefix="ctxo" verb="ListRecords">http:
      //comum.rcaap.pt/oai-stats/request</request>
  <ListRecords>
    <record>
      <header>
        <identifier>oai:comum.rcaap.pt:download_64</
            identifier>
        <datestamp>2009-11-02T16:59:02Z</datestamp>
      </header>
      <metadata>
        <ctx:context-objects xmlns:ns3="http://www.w3.org
            /2001/XMLSchema-instance" xmlns:ctx="info:ofi/
            fmt:xml:xsd:ctx" xmlns:dcterms="http://
            dublincore.org/documents/dcmi-terms/"
            ns3:schemaLocation="info:ofi/fmt:xml:xsd:ctx
            http://www.openurl.info/registry/docs/info:ofi
            /fmt:xml:xsd:ctx">
          <ctx:context-object identifier="oai:comum.rcaap
              .pt:download_64" timestamp="2009-11-02
              T16:59:02Z">
            <ctx:referent>
              <ctx:identifier>http://comum.rcaap.pt/
                  handle/123456789/12</ctx:identifier>
              <ctx:metadata-by-val>
                <ctx:format>http://dublincore.org/
                    documents/dcmi-terms/</ctx:format>
                <ctx:metadata>
                  <dcterms:type>article</dcterms:type>
```

```
                    <dcterms:rights>open access</
                        dcterms:rights>
                    <dcterms:language>eng</dcterms:language
                        >
                    <dcterms:author>Afonso, Joao</
                        dcterms:author>
                    <dcterms:author>Veiga, Pedro</
                        dcterms:author>
                  </ctx:metadata>
                </ctx:metadata-by-val>
              </ctx:referent>
              <ctx:requester>
                <ctx:identifier>
                    data:0a28242b48dea9247c25c1438ad4a571</
                    ctx:identifier>
                <ctx:identifier>data:193.136.44.0</
                    ctx:identifier>
                <ctx:metadata-by-val>
                  <ctx:format>http://dublincore.org/
                      documents/dcmi-terms/</ctx:format>
                  <ctx:metadata>
                    <dcterms:spatial>PT</dcterms:spatial>
                  </ctx:metadata>
                </ctx:metadata-by-val>
              </ctx:requester>
              <ctx:service-type>
                <ctx:metadata-by-val>
                  <ctx:format>http://dublincore.org/
                      documents/dcmi-terms/</ctx:format>
                  <ctx:metadata>
                    <dcterms:format>info:eu-repo/semantics/
                        objectFile</dcterms:format>
                  </ctx:metadata>
                </ctx:metadata-by-val>
              </ctx:service-type>
              <ctx:resolver>
                <ctx:identifier>http://comum.rcaap.pt/oai-
                    stats/request</ctx:identifier>
              </ctx:resolver>
            </ctx:context-object>
          </ctx:context-objects>
        </metadata>
      </record>
    </ListRecords>
</OAI-PMH>
```