U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

MESTRADO EM CIÊNCIA DA INFORMAÇÃO

Classificação de conjuntos de dados de investigação com base em seus registos de metadados

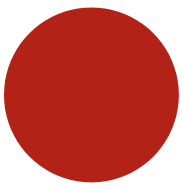
Raquel de Faria Oliveira Damasceno Ribeiro

M

2019

UNIDADES ORGÂNICAS ENVOLVIDAS

FACULDADE DE ENGENHARIA
FACULDADE DE LETRAS



Classificação de conjuntos de dados de investigação com base em seus registos de metadados

Raquel de Faria Oliveira Damasceno Ribeiro

Dissertação realizada no âmbito do Mestrado em Ciência da Informação, orientada pelo Professor Doutor João Miguel Rocha da Silva e coorientada pela Mestre Joana Patrícia de Sousa Rodrigues

Faculdade de Engenharia e Faculdade de Letras

Universidade do Porto

Membros do Júri

Presidente: Prof.a Doutora Maria Cristina de Carvalho Alves Ribeiro

Professora Associada da Faculdade de Engenharia da Universidade do Porto

Orientador: Prof. Doutor João Miguel Rocha da Silva

Professor Auxiliar Convidado da Faculdade de Engenharia da Universidade do Porto

Arguente: Prof.a Doutora Mariana Curado Malta

Professora Adjunta do Instituto Superior de Contabilidade e Administração do Porto

julho de 2019

O obstáculo é o caminho.

Provérbio Zen

Agradecimentos

Eu havia preparado uma secção de agradecimento bem detalhada. Entretanto, penso agora em ser breve.

Como é clássico, o agradecimento à família é fundamental. São o motivo não somente de conclusão do curso, mas o apoio em toda a caminhada do dia a dia. Todos são importantes, em um momento uns se fazem mais presentes, em outras ocasiões somos até mesmo surpreendidos com apoios inesperados. É claro que existem aqueles a quem podemos contar sempre, sempre. Em reconhecimento a todos os familiares não os vou nominar.

Do início da jornada, preciso agradecer em primeiro lugar ao amigo Gustavo Almeida. Respondeu prontamente a um pedido de carta de apresentação, que se diga de passagem, a meus olhos, parecia uma carta de motivação. Em segundo lugar, à minha grande amiga, Ana Maria Figueira, que também não poupou esforços de última hora para sintetizar em poucas palavras todo o trabalho que havíamos desenvolvido. E ainda, o apoio do meu chefe, Rafael Scalia Guedes, que tornou essa trajetória muito mais amena.

Não posso deixar de reconhecer o apoio indireto de pessoas que nem sequer sabem que o fizeram. Cada "não" que recebi em virtude das inúmeras ideias de inovação no campo profissional, contribuiu afirmativamente para essa empreitada. Talvez muito mais do que os "sim". Foram o impulso para eu estar aqui.

Existem aqueles que fizerem parte da caminhada efetivamente. Tive a oportunidade de compartilhar tarefas com praticamente todas as pessoas com quem estudei. Foi um aprendizado sem precedentes, especialmente voltado a interagir com uma cultura muito diferente da minha, além dos diferentes perfis principalmente em virtude da idade e da experiência profissional em "conflito"com o curso.

Meu especial agradecimento à secretária do curso, Sandra Reis, que muito mais do que orientações académicas ofereceu-me apoio pessoal e uma visão mais ampliada da vida.

Raquel de Faria Oliveira Damasceno Ribeiro

Resumo

Tão importante quanto acumular conhecimento é também assegurar-se de que este conhecimento estará disponível quando necessário de forma eficaz e eficiente. A recuperação da informação no ambiente digital sinaliza oportunidades inovadoras e desafios de democratização da informação. O acesso aberto, mais do que uma filosofia, tornou-se uma demanda pública.

A presente dissertação estuda o uso das classificações como forma de recuperar a informação digital nos repositórios de dados. O trabalho propõe a construção de um classificador de documentos a partir de metadados textuais como título, palavras-chave e descrição dos objetos digitais. São utilizados os sistemas de classificação e os tesauros, estruturas clássicas da Ciência da Informação como suporte ou subsídio ao modelo desenvolvido.

A fim de validar o classificador é apresentada uma ferramenta de *crowdsourcing*. E para apresentar os resultados os esforços são direcionados no sentido de oferecer interoperabilidade à solução.

Através da construção de um portal, demonstra de forma prática os resultados das experiências realizadas e aborda o paradigma de recuperação da informação a partir da classificação dos objetos digitais com vistas à democracia e literacia informacional. Pressupõe ainda, a construção de uma ferramenta aplicável às organizações e pessoas.

Como conclusão do estudo podem ser destacados dois pontos. O primeiro ponto diz respeito ao aumento na quantidade de objetos que são classificados aquando da utilização de um tesauro. O segundo ponto refere-se à importância do descritor linguagem de cada objeto digital em virtude da mudança na quantidade de objetos classificados. Essa quantidade pode ser na maioria dos casos incrementada, mas em alguns casos pode sofrer um decréscimo.

Palavras-chave: Recuperação da informação, literacia informacional, democratização da informação, classificador de documentos, tesauro

Abstract

As important as accumulating knowledge is ensuring that it will be available when needed effectively and efficiently. The recovery of information in the digital environment signals innovative opportunities and challenges of information democratization. Open access, more than a philosophy, has become a public demand.

The present dissertation studies the use of classifications as a way of retrieving digital information in data repositories. The work proposes the construction of a classifier of documents from textual metadata such as title, keywords and description of digital objects. Classification systems and thesauri are used, classic structures of Information Science as support or subsidy to the developed model.

In order to validate the classifier a crowdsourcing tool is presented. And to present the results efforts are directed towards offering interoperability to the solution.

Through the construction of a portal, it demonstrates in a practical way the results of the experiments carried out and approaches the paradigm of information retrieval from the classification of digital objects with a view to democracy and information literacy. It also presupposes the construction of a tool applicable to organizations and individuals.

As a conclusion of the study two points can be highlighted. The first point concerns the increase in the number of objects that are classified when using a thesaurus. The second point refers to the importance of the language descriptor of each digital object due to the change in the number of classified objects. This amount can in most cases be increased, but in some cases it may suffer a decrease.

Key-words: Information retrieval, information literacy, information democratization, document classifier, thesaurus

Lista de Figuras

1.1 "Árvore de objetivos" . . . 2

2.1 "Open Science Taxonomy", https://www.fosteropenscience.eu/foster-taxonomy/open-science . . . 9
2.2 "DCC Curation Lifecycle Model", http://www.dcc.ac.uk/resources/curation-lifecycle-model . . . 10
2.3 "Encoding", http://www.dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/?v=elements#subject . . . 13
2.4 "Encoding", http://www.dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/?v=elements#subject . . . 14
2.5 "Elementos e vocabulários DC", https://glennas.wordpress.com/2010/01/31/dublin-core-metadata-initiative-dcmi-learning-resources/ . . . 15
2.6 "Exemplos estrutura *JSON* e *XML*", https://www.w3schools.com/js/$js_{j}son_{x}ml.asp$ 16
2.7 Medeiros, 2014, Tipologia das Classificações . . . 21

3.1 Etapas de construção do *site* . . . 26
3.2 Comparando o mesmo registo em *XML* e *JSON* . . . 29
3.3 Comparando metadados em formato *XML* e *JSON* . . . 30
3.4 Comparando metadados em formato *XML* e *JSON* . . . 31
3.5 Tabelas *SQL* - Importação de dados . . . 32
3.6 Tabelas *SQL* - Conversão de dados . . . 33
3.7 Tabelas *SQL* - Totalização de dados . . . 34
3.8 Diagrama Entidade Relacionamento . . . 36
3.9 Tabelas *SQL* - Dados dos sistemas de classificação . . . 37
3.10 Tabelas *SQL* - Dados de *harvesting* . . . 38
3.11 Tabelas *SQL* - Dados já classificados . . . 39
3.12 Classificação de livro: FEUP x FLUP . . . 40
3.13 Comparativo *FOS*: 2002 e 20015 . . . 41
3.14 Experiências do classificador . . . 42

4.1 Objetivos do portal . . . 44
4.2 Página inicial . . . 45
4.3 Classificação a partir da origem dos dados . . . 46
4.4 Validação geral - simples . . . 48
4.5 Validação cruzada - por domínio . . . 49
4.6 Pesquisa web . . . 50
4.7 Auto validação . . . 51
4.8 Estatísticas . . . 52

4.9 Estatísticas por domínio 53
4.10 Estatísticas por domínio (detalhada) 54
4.11 Estatísticas por linguagem - *Frascati* 55
4.12 Estatísticas por linguagem - *Frascati* (detalhada) 55
4.13 Estatísticas por linguagem - *Unesco* 55
4.14 Estatísticas por linguagem - *Unesco* (detalhada) 56
4.15 Estatísticas por *join* - *Unesco* 56
4.16 Estatísticas por *join* - *Unesco* (detalhada) 57

5.1 Conclusões 60
5.2 Trabalhos futuros 63

Lista de Tabelas

3.1 Número de objetos *XML* e *JSON* 28
3.2 Mnemônicos 35
3.3 Tabelas *"tbPi"*: nomes e descrição 35
3.4 Tabelas *"tbPs"*: nomes e descrição 36
3.5 Tabelas *"tbPr"*: nomes e descrição 37

Abreviaturas e Símbolos

BOAI	Budapest Open Access Initiative
C&T	Ciência e Tecnologia
CID	Ciências da Informação e Documentação
CDD	Classificação Decimal de Dewey
CDU	Classificação Decimal Universal
DCMI	Dublin Core Metadata Initiative
DDC	Dewey Decimal Classification
DDI	Data Documentation Initiative
DER	Diagrama de Entidade Relacionamento
ECM	Enterprise Content Management
EOSC	European Open Science Cloud (EOSC)
ERA	European Research Area
EUDAT	European Data Infrastructure
FAIR	Findable, Accessible, Interoperable and Reusable
FEUP	Faculdade de Engenharia da Universidade do Porto
FOS	Fields of Science and Technology
FOSTER	Facilitate Open Science Training for European Research
GDI	Gestão de Dados de Investigação
HTML	Hypertext Markup Language
ICA	International Council on Archives
ICPSR	Inter University Consortium for Political and Social Research
IEEE	Instituto de Engenheiros Eletricistas e Eletrônicos
IMT	Internet Assigned Numbers Authority
INESC TEC	Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation - Linked Data
LCC	Library of Congress Classification
LCSH	Library of Congress Subject Headings
MeSH	Medical Subject Headings
NLM	National Library of Medicine Classification
OAI	Open Archives Initiative
OAI-PMH	Open Access Initiative – Protocol for Metadata Harvesting
OSI	Open Science Institute
OSF	Open Society Foundations
OWL	Ontology Web Language
RCCAP	Repositório Científico de Acesso Aberto de Portugal
RDA	Research Data Alliance
RDF	Resources Descriptions Framework

RDFS	Resources Descriptions Framework Schema
RI	Repositórios Institucionais
SGDB	Sistema Gerenciador de Banco de Dados
SOC	Sistema de Organização do Conhecimento
SQL	Structured Query Language
TKD	Title - Description - Keyword
UDC	Universal Decimal Classification
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	Extensible Markup Language

Conteúdo

1 Introdução **1**
- 1.1 Contexto e motivação . . . 1
- 1.2 Problemáticas . . . 1
- 1.3 Objetivos e resultados esperados . . . 2
- 1.4 Metodologia . . . 3
- 1.5 Estrutura da Dissertação . . . 4

2 Revisão Bibliográfica **5**
- 2.1 Ciência e dados abertos . . . 5
  - 2.1.1 e-Science . . . 5
  - 2.1.2 Open-Science . . . 6
  - 2.1.3 Ciclo de vida dos dados . . . 9
- 2.2 Descrição de recursos informacionais . . . 11
  - 2.2.1 Metadados . . . 12
  - 2.2.2 Esquemas de metadados . . . 13
  - 2.2.3 Web semântica . . . 15
  - 2.2.4 XML, JSON e JSON-LD . . . 16
- 2.3 Repositórios . . . 17
  - 2.3.1 Contexto e definição . . . 17
  - 2.3.2 Características . . . 18
  - 2.3.3 Iniciativas . . . 19
  - 2.3.4 Suporte a metadados específicos de domínio . . . 20
- 2.4 Classificação . . . 21
  - 2.4.1 Classificações bibliográficas . . . 21
  - 2.4.2 Classificações científicas . . . 22
  - 2.4.3 Organização do conhecimento no ambiente digital . . . 23

3 Abordagem metodológica **25**
- 3.1 Objetivo e descrição . . . 25
- 3.2 Um sistema *on line* para classificação de documentos . . . 25
  - 3.2.1 Ferramentas e tecnologias . . . 25
  - 3.2.2 Etapas . . . 27

4 Portal *Web* de classificação **43**
- 4.1 Objetivos . . . 43
- 4.2 Arquitetura . . . 44
  - 4.2.1 Classificador . . . 44
  - 4.2.2 Validador - *Crowdsourcing* . . . 46

4.2.3 Pesquisa 47
4.3 Funcionalidades 50
4.3.1 Auto validação 50
4.3.2 Estatísticas 50

5 Conclusões **59**
5.1 Visão prática 59
5.2 Trabalhos futuros 62
5.2.1 Classificação automática de documentos 62
5.2.2 Validação do classificador 65
5.2.3 Apresentação dos resultados 65
5.3 Considerações finais 66

Referências **69**

Capítulo 1

Introdução

1.1 Contexto e motivação

A presente dissertação foi desenvolvida no âmbito do Mestrado em Ciência da Informação, lecionado na Faculdade de Engenharia da Universidade do Porto - FEUP, sob o título: "Classificação de conjunto de dados de investigação com base em seus registos de metadados"

A recuperação da informação sempre foi motivo de meu interesse, uma vez que minha formação é relativa à Ciência da Computação. Ao lidar com as disciplinas que tratavam do tema durante o curso, o interesse foi ampliado e assim surgiu a motivação para esta dissertação.

Como é um tema muito amplo e pode ser abordado sob várias perspetivas foi necessário limitá-lo. A restrição foi pautada na utilização de metadados aplicados à gestão de dados de investigação. A escolha decorreu da utilização da ferramenta Dendro[1], aplicação desenvolvida na FEUP. Trata-se de um repositório digital para gestão de dados de investigação.

1.2 Problemáticas

A produção científica está inserida no contexto de produção e disseminação de informação de dados na *Web*. Os materiais produzidos no âmbito das Instituições Académicas são compartilhados por toda a comunidade e os "Repositórios Institucionais"são as ferramentas de armazenamento, distribuição e pesquisa da informação produzida.

A capacidade de produzir e gerir dados requer investimentos não apenas financeiros, mas também a otimização das tarefas e recursos, além da construção de ferramentas para apoiar os investigadores na gestão dos dados produzidos.

O processo colaborativo, especialmente a reutilização de dados, encontra suporte na produção de metadados. É um processo complicado, dispendioso e que não atrai a atenção dos investigadores que regra geral estão preocupados em realizar as pesquisas e deixam a demanda de registo dos dados para o segundo plano.

[1] http://dendro.fe.up.pt

Figura 1.1: "Árvore de objetivos"

Assim, surge a necessidade de gestão dos dados, planeada de modo a assegurar a reutilização dos dados, nomeadamente a curadoria digital, tema tratado no estudo da Ciência da Informação.

A curadoria de dados trabalha com a diversidade de domínios, naturezas e origens diferentes dos dados gerando complexidade nas diversas fases do ciclo de vida dos objetos digitais. A complexidade é aumentada pelo fato de não ser um processo automatizado e precisar da manipulação humana, o que acarreta mais dificuldade a todo o processo.

Portanto, lidar com o processo de gestão da informação torna-se um desafio, especialmente na fase de recuperação da informação. Não basta que a informação esteja devidamente depositada e acessível. É preciso assegurar-se de que será encontrada quando for buscada de forma rápida e eficiente.

1.3 Objetivos e resultados esperados

O objetivo desta dissertação está representado na árvore de objetivos conforme ilustra a figura 1.1.

É composta por três níveis, sendo o objetivo principal dividido em três objetivos no segundo nível e mais três ramificações. Destaca-se como objetivo principal, a proposição de um sistema de classificação para projetos de investigação de diversos domínios a partir dos seus metadados.

O objetivo principal pode ser detalhado em três outras metas que contribuem para o propósito inicial. A primeira consiste na classificação dos conjuntos de dados de investigação com base nas características de cada conjunto; a segunda, passa por recomendar os melhores descritores para novos projetos de investigação; a terceira trata da integração da abordagem proposta ao *workflow* da gestão de dados de investigação do *INESC TEC*.

Esta dissertação recairá sobre os dois primeiros objetivos do segundo nível, que são suportados por outros objetivos, essenciais em todo o processo. É de salientar que o terceiro objetivo nesse nível não será abordado e enquadra-se em propostas para trabalhos futuros.

O primeiro objetivo que trata da classificação dos registos é composto num terceiro nível por dois outros objetivos que contribuem para a consolidação da classificação pretendida. Passa essencialmente por selecionar o sistema de classificação que melhor se aplica aos dados a serem classificados e na aplicação da terminologia do sistema escolhido aos conjuntos de dados a classificar.

O segundo objetivo a ser alcançado vai de encontro à validação do sistema de recomendação proposto. A ferramenta para identificar a viabilidade do resultado apresentado será a utilização de experiências com investigadores e curadores.

Mediante a exposição dos objetivos pretendidos, espera-se que seja possível alcançá-los trabalhando de maneira sistémica de modo a obter resultados que possam consolidar o estudo. Espera-se sobretudo, a concretização dos objetivos através da implementação de uma solução prática e factível.

1.4 Metodologia

A metodologia que dá suporte ao projeto de dissertação apresentado é pautada no método QUANTI-qualitativo. A abordagem procura através da complementaridade dos dois métodos revelar os padrões implícitos nos dados a analisar.

Embora o estudo tenha características de pesquisa qualitativa e quantitativa simultâneas, é maioritariamente assente na metodologia quantitativa. O conjunto de dados produzido pela abordagem quantitativa serve de apoio aos dados qualitativos.

Para a abordagem qualitativa, a pesquisa será iniciada através do método exploratório possibilitando ao pesquisador a aproximação ao objeto alvo da pesquisa e habilitando-o a escolher as técnicas adequadas à sua pesquisa e no caso do trabalho será assentada na pesquisa bibliográfica.

A pesquisa bibliográfica fornece a base teórica e desnuda o caminho trilhado por outros pesquisadores e será realizada em livros, artigos de jornal, revistas científicas e outros documentos já publicados. A dissertação começa pelo desenvolvimento da revisão da literatura de modo a fazer o levantamento das iniciativas que abordam o tema e o contexto do assunto tratado. É averiguado o que já foi publicado, como a matéria foi abordada, as teorias já existentes e a confrontação das diferentes abordagens.

Seguidamente à pesquisa exploratória decorre a pesquisa descritiva, visando um estudo mais detalhado e o aprofundamento do caso. O estudo começa por escolher o repositório e o sistema de classificação a serem utilizados nas análises. Posteriormente serão selecionadas as ferramentas que possam dar suporte ao estudo. Por fim, serão produzidos os dados a serem utilizados na análise quantitativa e qualitativa.

Pretende-se dessa maneira buscar a análise e compreensão da complexidade da produção de metadados em conjunto de dados de investigação através de um estudo aprofundado.

1.5 Estrutura da Dissertação

Para além da introdução, esta dissertação contém quatro capítulos e mais uma secção destinada às referências bibliográficas. No primeiro capítulo são contempladas a introdução, assuntos como o contexto e a motivação, as problemáticas levantadas, além dos objetivos e resultados esperados e por fim, a estrutura definida para a dissertação.

O segundo capítulo destina-se à revisão de literatura e está estruturado em quatro tópicos principais. O primeiro introduz a questão da Ciência e dados abertos. É apresentada a contextualização do tema, a abordagem ao conceito de *e-Science* e os principais avanços históricos são ressaltados. Em alinhamento com o formato aberto aborda a gestão de dados de investigação e a curadoria de dados.

O segundo tópico lida com a descrição dos recursos informacionais e começa por fazer uma pequena analogia entre a informação na Biblioteconomia e no meio digital e destaca a importância dos metadados para identificação, organização e recuperação da informação armazenada ou depositada na *web*. Evidencia os esquemas de metadados, com especial ênfase na iniciativa *DCMI – Dublin Core Metadada Initiative*. Em virtude da necessidade cada vez maior de localizar e processar as informações produzidas no meio digital, aborda as ontologias, o *Resources Descriptions Framework (RDF)* e a *Ontology Web Language (OWL)*. Prossegue discorrendo sobre a troca de dados entre aplicativos e máquinas mediante os padrões *Extensible Markup Language (XML)*, *JavaScript Object Notation (JSON)* e *JavaScript Object Notation – Linked Data (JSON-LD)*.

Na terceira sessão enfatiza a importância dos repositórios institucionais como ferramentas que tratam de todo o processo da comunicação científica e seu papel na gestão do conhecimento científico. Expõe um problema recorrente nos repositórios originado na diversidade de domínios de publicações científicas e salienta a necessidade de uma infraestrutura capaz de atender a todas as áreas do conhecimento, ora de modo amplo, ora voltado ao atendimento das necessidades específicas de cada domínio.

O tópico final refere-se às classificações abordando as mais diferentes categorizações: dos saberes ou filosóficas; classificações científicas, também chamadas de taxonómicas, relativas à classificação dos seres; e as classificações bibliográficas. Descreve sucintamente alguns sistemas de classificação e faz o aporte à organização do conhecimento relativamente às classificações.

No terceiro capítulo é referido o sistema *on line* para classificação de documentos. São especificadas as etapas seguidas desde a escolha do repositório e sistema de classificação a ser utilizado, passando pela recolha de dados e detalhes de todo o processo para construção do classificador.

O quarto capítulo apresenta a consolidação de todo o trabalho realizado consubstanciado no "Portal *web* de classificação". O Portal é a ferramenta visual e interativa que permite expor todo o processo de desenvolvimento do classificador através da visualização dinâmica dos procedimentos executados e os respetivos resultados mediante dados estatísticos.

Por fim, no capítulo que encerra a presente dissertação é realizada a conclusão do trabalho, as abordagens para trabalhos futuros e as considerações finais. Logo em seguida há a secção reservada às referências bibliográficas.

Capítulo 2

Revisão Bibliográfica

2.1 Ciência e dados abertos

Tão importante quanto acumular conhecimento é também assegurar-se de que este conhecimento estará disponível quando necessário de forma eficaz e eficiente.

A capacidade de produzir conhecimento foi potenciada pelas ferramentas tecnológicas da Idade Contemporânea. A era da informação, além de oferecer oportunidades inovadoras lança desafios à democratização da informação.

Com o surgimento da *web* novas maneiras de disponibilizar os resultados de pesquisas através da publicação *on line* tornaram-se cada vez mais populares e a disseminação da informação mais eficaz. O acesso foi democratizado e a quantidade de dados ofertada foi incrementada. Essa nova realidade, associada ao poder da ciência computacional traz novas perspetivas à comunidade académica.

2.1.1 e-Science

A *internet* evoluiu e junto a ela os serviços *web* também foram se aperfeiçoando. Tornou-se em poucas décadas uma ferramenta indispensável nas mais diversas áreas do conhecimento humano inclusive no mundo científico.

A gama de informações que circula pela rede e a diversidade de interesses em áreas correlatas e também em áreas aparentemente disjuntas potenciaram desafios de conectividade e processamento dos dados produzidos em grande escala e de forma isolada, especialmente relacionados com a produção científica. Nesse contexto surge o conceito de *e-science*.

Segundo Hey and Hey (2006), o termo *e-science* foi apresentado por John Taylor, diretor geral do *Research Councils in the UK Office of Science and Technology*. Taylor percebeu que muitas áreas da ciência poderiam beneficiar de uma infraestrutura comum para apoiar colaborações multidisciplinares e distribuídas.

No âmbito do *Institute of Electrical and Electronics Engineers*, organização profissional dedicada ao avanço da tecnologia, podem ser encontradas, no sítio que reúne as conferências internacionais da entidade, duas definições para *e-Science*. Na primeira transcrição assim o define: ”o

e-Sience promove a inovação em pesquisas colaborativas em todas as disciplinas durante todo o ciclo de vida de pesquisa". IEEE - Institute of Electrical and Electronics Engineers (2014).

A segunda transcrição explicita uma definição mais alargada do termo e-Science, tal como formulada pelo IEEE: "*eScience* estuda, aprova e aprimora o processo contínuo de inovação em métodos de pesquisa intensivos em computação ou intensivos em dados; normalmente, isso é feito de forma colaborativa, geralmente usando infraestruturas distribuídas. O *eScience* engloba todas as áreas de pesquisa e aborda todas as etapas do ciclo de vida da pesquisa, desde a formulação das questões de pesquisa, passando por simulações e análise de dados em grande escala, descoberta científica, compartilhamento a longo prazo, reutilização e reaplicação dos resultados, bem como as ferramentas, processos e conhecimentos relevantes". IEEE - Institute of Electrical and Electronics Engineers (2014)

Para Hey (2002), a *e-Science* lida com a colaboração global em áreas chave da ciência e da infraestrutura que sejam capazes de suportar as necessidades advindas, nomeadamente a arquitetura, chamada de *Grid*, baseada em dados distribuídos, computação e colaboração. O autor destaca a importância dos programas internacionais em *e-Science* que tiveram início no Reino Unido e também outras iniciativas como a *Cyberinfrastructure* nos Estados Unidos e a *e-Infrastructure* na Europa.

Assim, *e-Science* não é uma nova disciplina científica, mas um suporte tecnológico que fomenta a ciência e lança um novo paradigma para a pesquisa científica a partir da ciência baseada em dados e em rede a partir da colaboração. A computação desempenha um papel proeminente na pesquisa científica contribuindo para suportar uma infraestrutura capaz de garantir entre outros serviços, a gestão de fluxos de trabalho, o processamento e o compartilhamento de dados.

A colaboração, por sua vez, impõe a necessidade de compartilhamento de informações. As ferramentas utilizadas durante o ciclo de vida da pesquisa devem disponibilizar recursos que permitam a qualquer interessado fazer uso de todo o conteúdo gerado e aceder a toda a infraestrutura relativa ao projeto. Portanto, a tecnologia de acesso aberto é fundamental no desenvolvimento da ciência colaborativa.

2.1.2 Open-Science

A base para o movimento *Open-Science*, Ciência aberta, é fundamentalmente, a questão do acesso livre aos conteúdos académicos relevantes, especialmente dados produzidos através de pesquisa financiada por dinheiro público.

Bartling and Friesike (2014) apresentam no seu livro uma discussão sobre o termo *Open Science* identificando-o com diferentes escolas de pensamento: escola de infraestrutura, abordando a arquitetura tecnológica; escola pública, envolvida com a acessibilidade da criação do conhecimento; escola de mensuração, preocupada com a medição do impacto da nova filosofia; escola democrática, tratando do acesso ao conhecimento e, por fim, escola pragmática, dedicada à pesquisa colaborativa.

Independente da categorização, a *Open Science* é um termo recorrente na comunidade científica. É também abrangente, uma vez que implica diferentes abordagens para o futuro da criação

e disseminação do conhecimento. Encontra-se em Araya and Vidotti (2010) a exposição da trajetória e evolução da produção e disseminação da informação digital, as quais são apresentadas seguidamente.

The Santa Fe Convention for the Open Archives Initiative [1], a Convenção de Santa Fé, no Novo México—EUA, embora já descontinuada, foi o referencial inicial para troca de conteúdo científico utilizando recursos tecnológicos informacionais e comunicacionais, em outubro de 1999. Em agosto do mesmo ano foi criado o sistema para armazenamento, recuperação e disseminação de documentos eletrónicos, *ArXiv*, primeiro repositório baseado na filosofia de arquivos abertos.

A iniciativa alavancou esforços internacionais e após apenas um ano foi celebrada a *Open Archives Initiative – OAI*, Iniciativa dos Arquivos Abertos, objetivando o desenvolvimento de padrões de interoperabilidade a fim de facilitar a disseminação eficiente de conteúdos digitais.

No seguimento do movimento de acesso aberto à literatura científica e aos periódicos académicos, vê-se em dezembro de 2001, a atuação do *Open Science Institute – OSI*, Instituto Sociedade Aberta, a atual *Open Society Foundations - OSF*, Fundação da Sociedade Aberta. Como resultado, foi apresentada a *Budapest Open Access Initiative - BOAI*, Declaração de Budapeste em fevereiro de 2002 [2].

O interesse mundial no intercâmbio de publicações científicas deu origem, em abril de 2003, à *Bethesda Statement on Open Access Publishing*, Declaração de Bethesda [3] com o propósito de tornar a informação mais ampla e oferecer a garantia de sua disponibilidade. A declaração apresenta a definição de "Acesso Livre", além de conclusões e recomendações dos seguintes grupos de trabalhos: instituições e agências de financiamento; bibliotecas e editores; cientistas e sociedades científicas.

O paradigma de acesso livre recebeu, em outubro de 2003, as contribuições de uma nova iniciativa: a Declaração de Berlim. Nesta, o conceito de acesso livre foi remodelado, adotando como contribuições ao padrão aberto, o resultado de pesquisas científicas originais, dados não processados e metadados, além de representações digitais de materiais pictóricos e gráficos e material académico multimédia. A *Berlin Declaration* [4], ou Declaração de Berlim sobre Acesso Livre ao Conhecimento nas Ciências e Humanidades, propôs duas condições relativamente às contribuições ao acesso livre.

A primeira condição visa garantir o direito ao acesso gratuito, irrevogável e mundial por parte dos autores a todos os utilizadores do material disponibilizado. E ainda, uma licença para copiar, usar, distribuir, transmitir e exibir o trabalho publicamente e realizar e distribuir obras derivadas, em qualquer suporte digital para qualquer propósito responsável, sujeito à correta atribuição da autoria.

A segunda condição diz respeito ao depósito de uma versão completa da obra e todos os materiais suplementares num formato eletrónico normalizado e apropriado em pelo menos um

[1] http://www.openarchives.org/sfc/sfc_entry.htm
[2] https://www.budapestopenaccessinitiative.org/read
[3] http://legacy.earlham.edu/ peters/fos/bethesda.htm
[4] https://openaccess.mpg.de/Berlin-Declaration

repositório que utilize normas técnicas adequadas (como as definições *Open Archive*). É recomendado ainda, que o depósito seja mantido por uma instituição académica, sociedade científica, organismo governamental ou outra organização estabelecida que pretenda promover o acesso livre, a distribuição irrestrita, a interoperabilidade e o arquivo a longo prazo.

O Projeto piloto de acesso aberto da Comissão Europeia nasceu em 2008, intitulado *OpenAIRE - Open Access Infrastructure for Research in Europe*[5]. Busca formas inovadoras de comunicar e monitorizar a pesquisa académica através de uma implementação efetiva de Ciência Aberta.

Em 2012 foi apresentado o consórcio *EUDAT - European Data Infrastructure*[6]. Tem o propósito de conceber uma solução pan-europeia para o desafio da proliferação de dados nas comunidades científicas e pesquisas da Europa. Está inserido numa rede distribuída por 15 países europeus. Atua na gerência de dados de pesquisa e recursos de armazenamento e possui alguns dos supercomputadores mais poderosos da Europa.

Como iniciativa da Comissão Europeia, da Fundação Nacional de Ciência do Governo dos Estados Unidos, do Instituto Nacional de Padrões e Tecnologia e do Departamento de Inovação do Governo Australiano foi efetivada em 2013, a *RDA - Research Data Alliance*[7], Aliança de Dados de Pesquisa. Tem como objetivo construir a infra-estrutura social e técnica para permitir o compartilhamento aberto e a reutilização de dados.

No contexto do *European Research Area - ERA*, Espaço Europeu da Investigação, atendendo as diretrizes do projeto *Horizon 2020* a fim de cumprir políticas de acesso aberto e regras de participação, foi implementado a partir de 2014 o *Facilitate Open Science Training for European Research - FOSTER* [8]. Composto por onze parceiros em seis países tem como objetivo principal, contribuir para uma mudança real e duradoura no comportamento dos pesquisadores europeus para garantir que o *Open Science* se torne a norma. O projeto foi aprimorado, recebendo o nome *FOSTER Plus*. Este lida com comunidades e instituições de pesquisa, além de organizações de financiamento a promover a integração de princípios e práticas de acesso aberto, seguindo as políticas de acesso aberto do *ERA* e *Horizon 2020*. O programa atende toda a comunidade académica, com especial incidência em jovens cientistas, pessoal académico e decisores políticos.

De acordo com a iniciativa *FOSTER*, a definição de *Open Science* é: ”o movimento para tornar a pesquisa científica, os dados e a disseminação acessíveis a todos os níveis de uma sociedade inquiridora”.

Como se pode observar na Figura 2.1, a taxonomia do *Open Science* proposta pela organização está dividida em: *Open Access*, *Open Data*, *Open Reproducible Research*, além da subdivisão do próprio *Open Science*, nomeadamente em: *Open Science Definition*, *Open Science Evaluation*, *Open Science Guidelines*, *Open Science Policies*, *Open Science Projects* e *Open Science Tools*.

Ainda no âmbito da União Europeia, há que se referir a *European Open Science Cloud (EOSC)*[9],

[5] https://www.openaire.eu/about
[6] https://eudat.eu/what-eudat
[7] https://www.rd-alliance.org/about-rda
[8] https://www.fosteropenscience.eu/about
[9] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

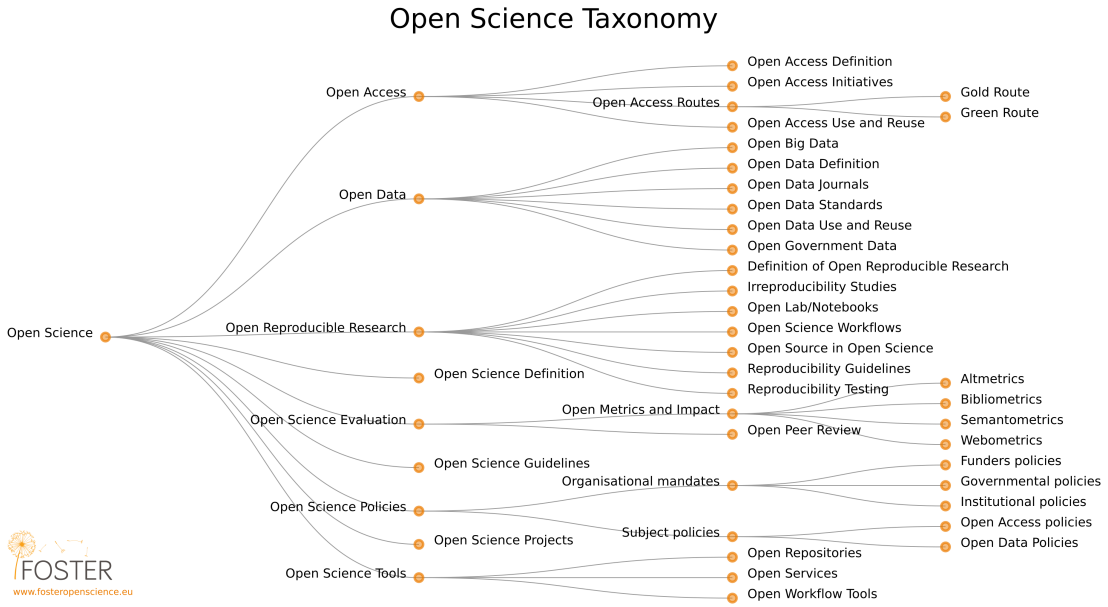


Figura 2.1: "Open Science Taxonomy", https://www.fosteropenscience.eu/foster-taxonomy/open-science

a nuvem para dados de pesquisa na Europa. Visa criar um ambiente confiável para hospedar e processar dados de pesquisa para apoiar a ciência da União Europeia em seu papel de liderança global. Na presente iniciativa reside o projeto *GO FAIR*. Oferece um ecossistema aberto e inclusivo para indivíduos, instituições e organizações a fim de implementarem os princípios de dados do *FAIR*, acrônimo para *Findable*, *Accessible*, *Interoperable* e *Reusable*, tornando os dados localizáveis, acessíveis, interoperáveis e reutilizáveis.

O fortalecimento do acesso aberto foi potenciado em 2018 com a iniciativa *Plan S*, apoiada pelo consórcio internacional de financiadores de pesquisa, *coAllition S*[10]. De acordo com esse Plano, a partir de 2021 todas as publicações científicas cujas pesquisas tenham sido financiadas por subvenções públicas deverão ser publicadas em plataformas ou periódicos de Acesso Aberto.

2.1.3 Ciclo de vida dos dados

Os dados de pesquisa são em grande parte disponibilizados em formato aberto de modo que os resultados possam ser compartilhados e reaproveitados. O paradigma aberto é amplamente aceite nas comunidades científicas. Esforços para assegurar o processamento adequado dos dados e a qualidade das informações na gestão dos dados através de políticas, padrões e práticas é uma meta de investigadores, pesquisadores e curadores.

Nesse sentido, a Gestão de Dados de Investigação (GDI) torna-se fundamental para que o processo seja consistente e os dados sejam armazenados, preservados e compartilhados com segurança, agregando mais qualidade nas atividades de pesquisa e a despender menor esforço.

[10] https://www.coalition-s.org/

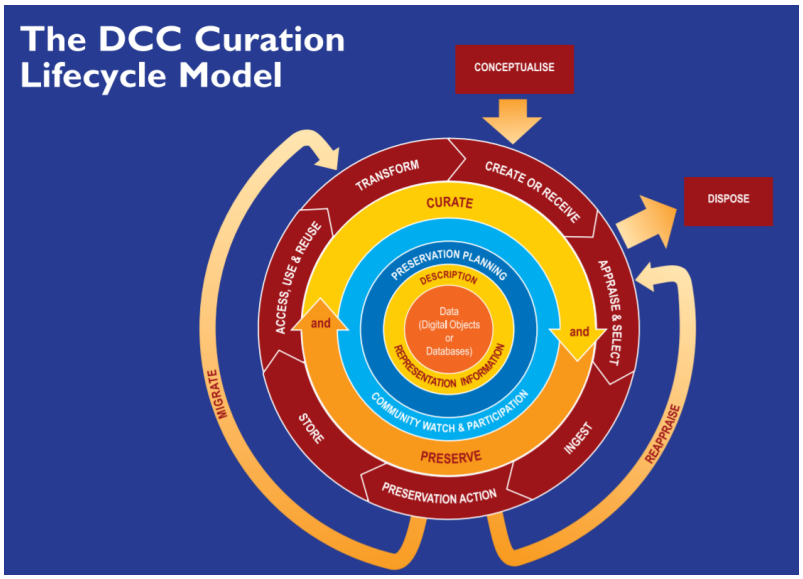


Figura 2.2: "DCC Curation Lifecycle Model", http://www.dcc.ac.uk/resources/curation-lifecycle-model

A GDI implica não somente produzir dados, mas também assegurar que esses resultados fiquem disponíveis a outros pesquisadores mediante o reaproveitamento das informações coletadas e processadas, imprimindo maior celeridade a novas descobertas e reduzindo esforços resultantes de experimentos duplicados. Em outras palavras, busca a reprodutibilidade dos resultados de investigação, um dos pilares da pesquisa científica.

A curadoria digital procura oferecer suporte ao desafio de gerir dados de investigação adequadamente durante todo o ciclo de vida da informação. Ocorre contudo, que identificar os benefícios da curadoria não é suficiente, sendo necessário colocá-la em prática no ambiente de trabalho, o que exige soluções técnicas, ferramentas e recursos de aprendizado.

O *Data Curation Center - DCC*, Centro de Curadoria Digital é uma instituição do Reino Unido que disponibiliza suporte, implementação de políticas, infraestrutura e serviços para organizações de pesquisa através de parcerias no Reino Unido, Europa, Austrália e Estados Unidos.

Na figura 2.2 é apresentado o ciclo de vida do modelo apresentado pelo *DCC*. Higgins (2008) explica as fases representadas na imagem, a começar pela figura central do ciclo, os dados. Refere que estes podem ser objetos digitais ou bases de dados. Prossegue descrevendo as ações que permeiam a base do esquema, inerentes a todas as atividades do ciclo. São atividades gerenciais e administrativas da curadoria. Sistematiza então, as próximas fases em duas categorias: ações sequenciais e ocasionais.

Como ações sequenciais apresenta:

- **Conceptualização:** planeamento da criação dos dados, desde a captura até o armazenamento;

- **Criação e receção:** criar dados atribuindo-lhes metadados e recebê-los seguindo políticas pré-determinadas;
- **Avaliação e seleção:** avaliar e selecionar os dados para curadoria e preservação a longo prazo;
- **Ingestão:** transferir dados para um arquivo, repositório, servidor ou *data center*;
- **Preservação:** garantir a integridade, confiabilidade, autenticidade dos dados através de ações como limpeza de dados, validação, atribuição de metadados de preservação e formatos de arquivos aceitáveis;
- **Armazenamento:** armazenamento dos dados de forma segura atendendo padrões relevantes;
- **Acesso, uso e reuso:** garantir a acessibilidade dos dados com aplicação de controle de acessos e autenticação;
- **Transformação:** criar novos dados a partir dos originais, como por exemplo, seleções, consultas e migração para novos formatos.

E como ações ocasionais:

- **Eliminação:** dados não selecionados podem ser transferidos para outros arquivos ou repositórios ou serem destruídos de maneira segura;
- **Reavaliação:** dados podem ser adicionados para avaliação adicional e nova seleção;
- **Migração:** com objetivo de evitar a obsolescência de hardware e software os dados podem ser migrados para formatos diferentes.

O *DCC* enfatiza que a curadoria e a preservação dos dados são processos contínuos que necessitam planeamento durante todo o ciclo de vida dos objetos digitais para que permaneçam autênticos, confiáveis e utilizáveis de forma a assegurar sua integridade.

2.2 Descrição de recursos informacionais

Na Biblioteconomia, as informações relativas à indexação e catalogação são utilizadas como forma de localizar os itens de uma coleção, tal como livros, revistas ou jornais. Os catálogos auxiliam a descoberta e localização de itens em um acervo através de propriedades dos itens como por exemplo autor, título, assunto, data de publicação, entre outros.

No meio digital existem os metadados e também os esquemas para identificação, organização e recuperação da informação armazenada ou depositada na *web*. Os esquemas regem a estrutura dos registos de metadados que estão associados aos recursos de uma coleção.

2.2.1 Metadados

Para (Alves, 2010), o termo metadados ganhou destaque na década de 90 e passou a ser utilizado para descrever recursos informacionais sendo amplamente utilizado em diversas áreas do conhecimento. Contudo, a autora enfatiza que o termo foi criado nos anos de 1960 por Jack E. Myers.

Segundo Baca (2016), os metadados não só identificam e descrevem um objeto de informação com também documentam como esse objeto se comporta, qual a sua função e uso, sua relação com outros objetos de informação e como foi gerido ao longo do tempo.

Formenton et al. (2018) acredita que a concepção de metadados encontra sua origem na catalogação de bibliotecas, com o objeto de descrever um recurso informacional de maneira única. O processo busca garantir a recuperação da informação em qualquer ambiente/sistema: convencional ou digital.

No âmbito do *DCC*, Higgins (2007) fornece a definição e a importância dos metadados: "Os metadados são a espinha dorsal da curadoria digital. Sem eles, um recurso digital pode ser irrecuperável, não identificável ou inutilizável. Metadados são informações descritivas ou contextuais que se referem ou estão associadas a outro objeto ou recurso. Isso geralmente toma a forma de um conjunto estruturado de elementos que descrevem o recurso informacional e auxilia na sua identificação, localização e recuperação, ao mesmo tempo em que facilita a gestão de conteúdo e acesso".

Duval et al. (2002) destaca três maneiras de associar metadados a recursos informacionais:

- **Metadados incorporados:** os metadados são criados juntamente com a criação do recurso.
- **Metadados associados:** os metadados são gerados em arquivos separados e acoplados aos recursos descritos
- **Metadados de terceiros:** são armazenados em repositórios por meio de uma organização que pode não ter acesso ao conteúdo dos recursos.

Os autores fazem ainda outra diferenciação importante entre os metadados: entradas objetivas e subjetivas. Os primeiros, como por exemplo autoria, data de criação e versão, entre outros, por terem uma natureza objetiva podem inclusive, serem gerados por máquina. Outras informações sobre os objetos, como palavras-chave, resumos, por exemplo, estão sujeitos a diferentes pontos de vista.

É importante enfatizar que organizações diferentes gerem metadados de recursos de acordo com interesse próprios, o que pode causar diferenças, levando a ambiguidades e confusões.

A subjetividade pode ser ainda maior quando se trata de descritores específicos de um domínio. Um metadado que descreve uma característica pedagógica por exemplo, pode ilustrar a questão subjetiva de certos descritores, uma vez que a característica pedagógica pode depender de uma filosofia educacional.

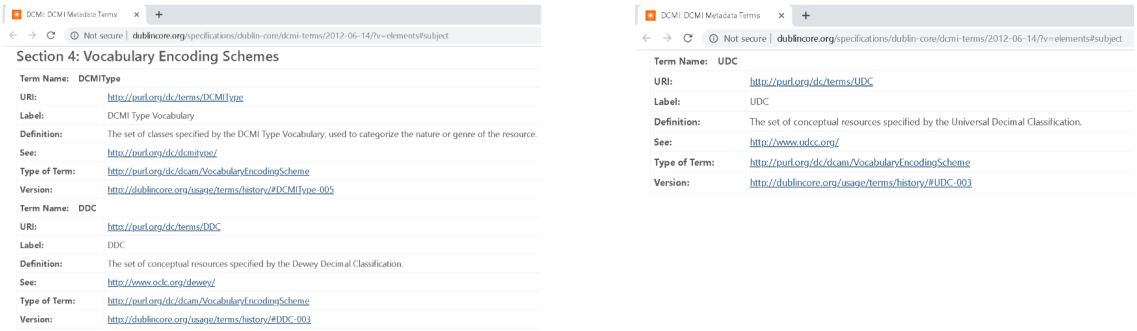


Figura 2.3: "Encoding", http://www.dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/?v=elements#subject

2.2.2 Esquemas de metadados

Os esquemas de metadados de acordo com Higgins (2007), surgiram devido às necessidades específicas de comunidades de utentes. Os esquemas estabelecem padrões de estrutura e padrões de conteúdo de metadados para descrição de recursos. É um processo democrático em que a comunidade, através do consenso, estabelece regras para submissão, aprovação e publicação de novos elementos. O apoio de entidades de excelência com reconhecimento nacional ou internacional asseguram mais visibilidade e aceitação dos padrões entre os quais podem ser citados, a *Library Congress* e o *International Council on Archives - ICA*.

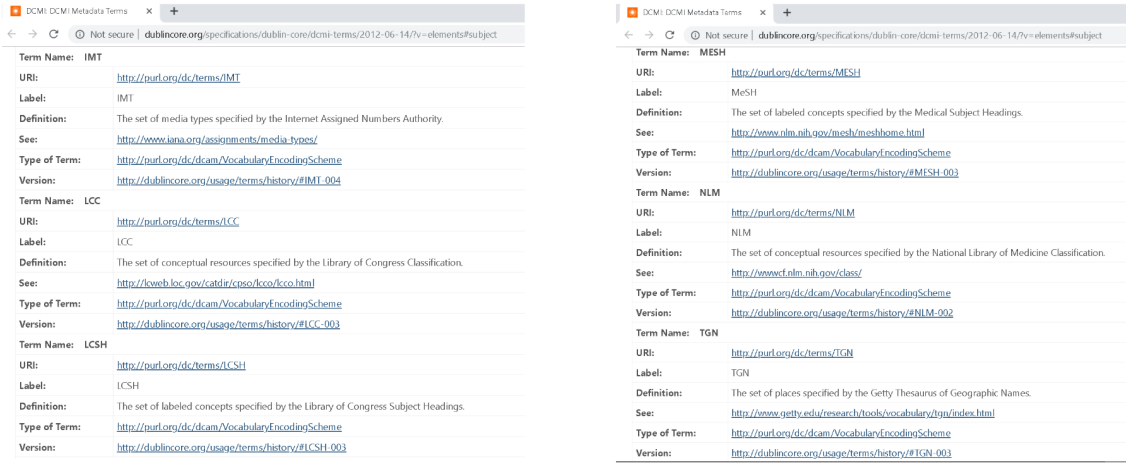
Existem padrões de metadados de aplicação genérica, utilizados para descrever e pesquisar recursos informacionais na *web* utilizados por ampla quantidade de domínios permitindo a interoperabilidade de áreas diferentes. Por outro lado, existem esquema de metadados altamente especializados para comunidades específicas desenvolvidos por especialistas de um domínio de conhecimento de modo a atender as especificidades daquela área.

A iniciativa que merece destaque relativamente ao desenvolvimento e disponibilização de padrão de metadados de repercussão internacional teve origem na década de 1990 em Dublin, Ohio, nos EUA. Conhecida como *Dublin Core Metadata Initiative - DCMI* é uma organização aberta que apoia a inovação no design de metadados e as melhores práticas na ecologia de metadados.

O *Dublin Core - DC*, é o padrão adotado pela *DCMI* sendo recomendado pelo consórcio *World Wide Web Consortium - W3C* [11] *W3C*. É um modelo simples de descrição de recursos digitais constituído por dois grupos de descritores. O primeiro grupo, *DC Metadata Element Set*[12], é o mais utilizado sendo composto por 15 elementos. Outro grupo, o *DC Metada Terms*[13] é um grupo de descritores mais refinados.

Além dos refinamentos, o *DC* possui um conjunto de esquemas de codificação usados para auxiliar a correta interpretação do valor de um elemento. Há dois tipos, esquemas de codificação de vocabulário e esquemas de codificação de sintaxe.

[11] https://www.w3.org/RDF/
[12] http://dublincore.org/documents/dces/
[13] http://dublincore.org/documents/dcmi-terms/



DCMI: DCMI Metadata Terms — Not secure | dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/?v=elements#subject

Term Name:	**IMT**
URI:	http://purl.org/dc/terms/IMT
Label:	IMT
Definition:	The set of media types specified by the Internet Assigned Numbers Authority.
See:	http://www.iana.org/assignments/media-types/
Type of Term:	http://purl.org/dc/dcam/VocabularyEncodingScheme
Version:	http://dublincore.org/usage/terms/history/#IMT-004
Term Name:	**LCC**
URI:	http://purl.org/dc/terms/LCC
Label:	LCC
Definition:	The set of conceptual resources specified by the Library of Congress Classification.
See:	http://lcweb.loc.gov/catdir/cpso/lcco/lcco.html
Type of Term:	http://purl.org/dc/dcam/VocabularyEncodingScheme
Version:	http://dublincore.org/usage/terms/history/#LCC-003
Term Name:	**LCSH**
URI:	http://purl.org/dc/terms/LCSH
Label:	LCSH
Definition:	The set of labeled concepts specified by the Library of Congress Subject Headings.
Type of Term:	http://purl.org/dc/dcam/VocabularyEncodingScheme
Version:	http://dublincore.org/usage/terms/history/#LCSH-003

DCMI: DCMI Metadata Terms — Not secure | dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/?v=elements#subject

Term Name:	**MESH**
URI:	http://purl.org/dc/terms/MESH
Label:	MeSH
Definition:	The set of labeled concepts specified by the Medical Subject Headings.
See:	http://www.nlm.nih.gov/mesh/meshhome.html
Type of Term:	http://purl.org/dc/dcam/VocabularyEncodingScheme
Version:	http://dublincore.org/usage/terms/history/#MESH-003
Term Name:	**NLM**
URI:	http://purl.org/dc/terms/NLM
Label:	NLM
Definition:	The set of conceptual resources specified by the National Library of Medicine Classification.
See:	http://wwwcf.nlm.nih.gov/class/
Type of Term:	http://purl.org/dc/dcam/VocabularyEncodingScheme
Version:	http://dublincore.org/usage/terms/history/#NLM-002
Term Name:	**TGN**
URI:	http://purl.org/dc/terms/TGN
Label:	TGN
Definition:	The set of places specified by the Getty Thesaurus of Geographic Names.
See:	http://www.getty.edu/research/tools/vocabulary/tgn/index.html
Type of Term:	http://purl.org/dc/dcam/VocabularyEncodingScheme
Version:	http://dublincore.org/usage/terms/history/#TGN-003

Figura 2.4: "Encoding", http://www.dublincore.org/specifications/dublin-core/dcmi-terms/2012-06-14/?v=elements#subject

Os esquemas de codificação de vocabulário indicam que o valor é um termo de um vocabulário controlado, expresso em tabelas de valores. As figuras, 2.3 e 2.4 apresentam os esquemas.

Entre os exemplos destacados, há os seguintes vocabulários:

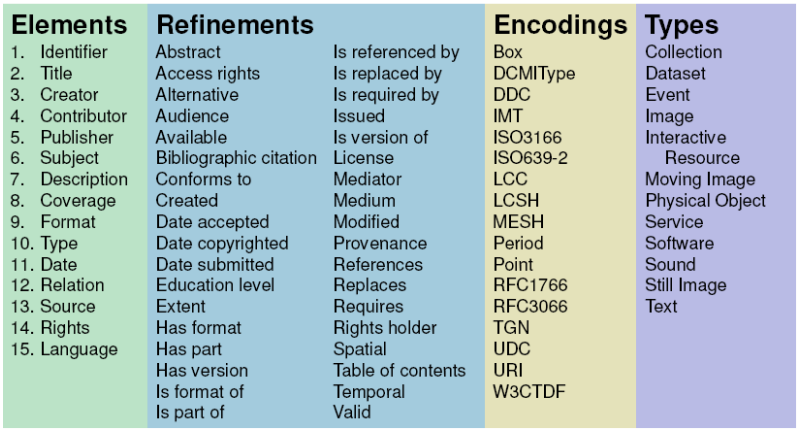
1. *Library of Congress Subject Headings - LCSH*
2. *Medical Subject Headings - MeSH*
3. *Dewey Decimal Classification - DDC*
4. *Library of Congress Classification - LCC*
5. *Universal Decimal Classification - UDC*
6. *Thesaurus of Geographic Names - TGN*
7. *National Library of Medicine Classification - NLM*
8. *Internet Assigned Numbers Authority - IMT*

Os esquemas de codificação de sintaxe apresentam a indicação do formato de um valor de acordo com uma notação formal. Um exemplo, seria uma *string* a indicar o valor de uma data, como "2019-12-01"

Na figura 2.5 estão ilustrados os elementos e os esquemas de codificação do *DC*.

No âmbito das Ciências Sociais, o projeto *Data Documentation Initiative - DDI* [14], criado em 1995, estabelece um padrão internacional para o conteúdo, preservação, transporte e preservação de documentação em bases de dados em Ciências Sociais. Uma importante funcionalidade oferecida pelo programa é o mapeamento para outros padrões, especialmente o *Dublin Core*.

[14] https://www.ddialliance.org/about/about-the-alliance.



Elements	Refinements		Encodings	Types
1. Identifier	Abstract	Is referenced by	Box	Collection
2. Title	Access rights	Is replaced by	DCMIType	Dataset
3. Creator	Alternative	Is required by	DDC	Event
4. Contributor	Audience	Issued	IMT	Image
5. Publisher	Available	Is version of	ISO3166	Interactive Resource
6. Subject	Bibliographic citation	License	ISO639-2	Moving Image
7. Description	Conforms to	Mediator	LCC	Physical Object
8. Coverage	Created	Medium	LCSH	Service
9. Format	Date accepted	Modified	MESH	Software
10. Type	Date copyrighted	Provenance	Period	Sound
11. Date	Date submitted	References	Point	Still Image
12. Relation	Education level	Replaces	RFC1766	Text
13. Source	Extent	Requires	RFC3066	
14. Rights	Has format	Rights holder	TGN	
15. Language	Has part	Spatial	UDC	
	Has version	Table of contents	URI	
	Is format of	Temporal	W3CTDF	
	Is part of	Valid		

Figura 2.5: "Elementos e vocabulários DC", https://glennas.wordpress.com/2010/01/31/dublin-core-metadata-initiative-dcmi-learning-resources/

2.2.3 Web semântica

A *Web* foi desenvolvida inicialmente para disponibilizar informações somente aos humanos. A crescente necessidade de recuperar a informação implicou novas formas de processamento da informação, nomeadamente através das máquinas. Assim, surgiu o conceito de *Web semântica*, que nas palavras de Araya and Vidotti (2010) "é o desenvolvimento de linguagens para a expressar informação de forma processável por uma máquina". Significa em outras palavras, atribuir semântica aos dados.

Os esquemas de metadados tratados anteriormente descrevem a estrutura dos registos de metadados, mas não conseguem atribuir-lhes semântica. Uma ontologia estabelece a semântica aos recursos descritos e representados pelos metadados. Para Maculan et al. (2017), as ontologias são uma forma de representar conhecimento e são capazes de fazer a desambiguação de termos. Nesse sentido, o propósito de uma ontologia é estabelecer uma semântica comum a determinadas comunidades de forma a compartilhar suas informações, ou seja, formalizar o conhecimento de uma comunidade para que possa ser recuperado.

De acordo com Decker et al. (2000), as ontologias fornecem um entendimento comum dos tópicos que podem ser comunicados entre pessoas e sistemas de aplicativos, mediante conceitos formais compartilhadas de domínios particulares. São usadas por exemplo, no comércio eletrónico, integração vertical de mercados ou mecanismos de pesquisa. Os autores destacam que as ontologias usam a hierarquização de conceitos dentro de um domínio através da descrição das propriedades cruciais de cada um desses conceitos. O processo consiste na atribuição de valores à conceptualização proposta e utilização de sentenças lógicas.

Para Seiji and Bittencourt (2015), as ontologias são geralmente expressas em uma linguagem baseada em lógica, de modo que distinções precisas, consistentes e significativas possam ser feitas entre as classes, propriedades e relações. Normalmente são utilizadas a lógica de predicados ou a lógica descritiva.

```
{"employees":[
  { "firstName":"John", "lastName":"Doe" },
  { "firstName":"Anna", "lastName":"Smith" },
  { "firstName":"Peter", "lastName":"Jones" }
]}
```

```
<employees>
  <employee>
    <firstName>John</firstName> <lastName>Doe</lastName>
  </employee>
  <employee>
    <firstName>Anna</firstName> <lastName>Smith</lastName>
  </employee>
  <employee>
    <firstName>Peter</firstName> <lastName>Jones</lastName>
  </employee>
</employees>
```

Figura 2.6: "Exemplos estrutura *JSON* e *XML*", https://www.w3schools.com/js/js$_{j}son_{x}ml.asp$

À medida que a disponibilização de recursos *web* foi aumentando e se aperfeiçoando, a necessidade de os controlar impôs a criação de mecanismos cada vez mais eficazes para localizar e processar toda a informação produzida. Conforme foi descrito, as ontologias foram criadas para oferecer semântica, ou seja, significado aos metadados.

A estruturação da informação entretanto estava em franco desenvolvimento e carecia de novas ferramentas. Seiji and Bittencourt (2015) asseveram que para oferecer suporte a esse processo de evolução havia a necessidade de uma base para processamento dos metadados assegurando a interoperabilidade das descrições entre diferentes aplicações. Assim, foi necessária a criação de padrões tanto para a sintaxe quanto para a semântica, a padronização de formas de acesso e a especificação de vocabulários comuns.

2.2.4 XML, JSON e JSON-LD

A troca de dados entre aplicativos diferentes ou entre máquinas distintas precisa seguir um padrão para que o intercâmbio de informações ocorra satisfatoriamente. No contexto da *Web* são referenciados dois padrões, nomeadamente *Extensible Markup Language - XML* e *JavaScript Object Notation - JSON*. Ambos possuem semelhanças e diferenças.

Importa contudo ressaltar que, independente da estrutura dos padrões, ambas apresentam um formato de troca de dados independente de linguagens de programação. Compartilham também duas outras características: são auto-descritivas e hierárquicas.

A próxima figura 2.6, apresenta de forma bem simples e didática, a comparação entre os dois modelos através do exemplo de criação de uma matriz de três funcionários. Na parte superior da imagem está implementado o exemplo em *XML* e na parte inferior, o mesmo exemplo em *JSON*.

O *XML* destaca-se por ser uma linguagem de anotação de documentos sendo o mais indicado para aplicações orientadas a documentos. Enquanto o *XML* é mais indicado para troca de documentos, o *JSON* oferece mapeamento direto para objeto de linguagem de programação.

Em 2014 o *W3C* aprovou um novo padrão baseado no formato *JSON*, chamado *JSON-LD*. Segundo o *schema.org* é o formato mais recente para expressar dados estruturados. Foi projetado para ser usado como *JSON*, sem conhecimento de *RDF* ou para ser usado como *RDF* se desejado, e também para uso com outras tecnologias de dados vinculados, como por exemplo *SPARQL*.

2.3 Repositórios

No seguimento dos conceitos apresentados anteriormente, nomeadamente dados abertos, ontologias e metadados, é abordado em seguida, o que vem a ser um repositório enquadrando-o no contexto até então exposto.

2.3.1 Contexto e definição

De acordo com Baptista (2017), "Os repositórios digitais científicos são uma das formas mais comuns de implementar o acesso aberto; constituem o que normalmente se designa por via verde para o acesso aberto".

No entendimento de Sobral and dos Santos (2017), os repositórios digitais são sistemas informacionais multifacetados em virtude de terem propósitos, objetivos dos mantenedores e público alvo diferentes. Relativamente ao propósito podem ser classificados como académicos/científicos; artísticos, culturais e sociais. Quanto ao conteúdo podem ser temáticos ou institucionais. Sob a perspetiva documental, serem monudocumentais ou multidocumentais. Em relação à administração podem ser consorciados, centralizados, descentralizados e governamentais. Quanto aos tipos de dados enquadrarem-se como textuais ou multimédia. Sob a natureza da informação tidos como primários, agregadores ou de fontes secundárias. E finalmente, uma última classificação no que tange ao sistema de avaliação, serem *peer-reviewed* ou não avaliados.

A abordagem relativa à natureza facetada dos repositórios proposta por estes autores não é objeto do presente trabalho, que tem o foco dirigido aos Repositórios Institucionais (RI). Os RI além de tratarem de todo o processo da comunicação científica têm a atribuição de gerir o conhecimento científico. Chapman et al. (2009), afirma que um repositório institucional coleta, gere e dissemina materiais produzidos em uma instituição.

No estudo intitulado "Repositórios Institucionais de Acesso Aberto: Análise do Cenário nos Países Ibero-Americanos", Cocco (2012) conclui: "A maioria (82,7%) dos repositórios institucionais dos países Ibero-Americanos foram desenvolvidos por Universidades. É possível observar que 36,8% têm as bibliotecas e centros de documentação e informação como responsáveis pela gestão e 83,9% utilizam o software *DSpace*, e registaram mais de 119 tipos de coleções, sendo que 6,7% das coleções são publicações de carácter científico para o depósito dos documentos e 58,6% utilizam o processo de auto-arquivamento pelos autores".

Para Chapman et al. (2009), os *RI* incorporam metadados que podem ser extraídos de sites *web*, mapeados de planilhas ou banco de dados. Assim, podem ser mapeados e convertidos de outros sistemas ou obtidos a partir do próprio criador, pela equipa da biblioteca ou do repositório. Complementarmente destaca a proveniência dos metadados que podem ter origem em várias disciplinas.

A definição transcrita de um trabalho conjunto da Universidade do Minho e da Universidade do Porto em Furtado et al. (2017) oferece uma abordagem mais ampla aos repositórios institucionais. "Um repositório institucional é uma infra-estrutura mantida por uma organização, tal como uma universidade ou um centro de investigação, com o propósito de colecionar e preservar a sua produção científica, técnica ou administrativa, e de lhe dar visibilidade. Os repositórios institucionais podem contribuir para iniciativas mais alargadas, como a agregação da publicação científica a nível nacional. Os repositórios têm também uma função importante na auditoria das organizações".

2.3.2 Características

Para Rodrigues et al. (2010), os repositórios institucionais lidam com a curadoria de dados relativos a documentos de literatura científica, atividades de pesquisa e depósito de dados científicos. Mas, além das características particulares de conteúdo, os repositórios tratam também de aspetos políticos, legais e éticos provenientes do acesso e reutilização dos dados científicos.

Nesse ponto convém destacar a nomenclatura aplicada aos repositórios institucionais que lidam com dados e por esse motivo serem conhecidos como repositório de dados.

Os mesmos autores apresentam três níveis no processo de curadoria dos dados de repositórios de dados. O primeiro nível garante o armazenamento fiável do conjunto de dados. O segundo nível assegura a descrição do conjunto de dados, ou à coleção, através de metadados. E por fim, o terceiro nível trata da descrição individual dos itens que compõem o conjunto de dados.

Os repositórios institucionais armazenam objetos representados por ficheiros. Entretanto, as aplicações científicas podem ter dados científicos associados. Assim, os repositórios precisam suprir a necessidade dos investigadores de armazenarem os dados produzidos no âmbito das pesquisas realizadas.

Um tipo de serviço oferecido pelos repositórios institucionais que merece ser referido é o processo de auto-arquivo no qual o próprio autor faz o depósito de um trabalho científico. Durante o procedimento de depósito o utilizador descreve e caracteriza os objetos a serem depositados entre os quais estão inclusas as permissões, licenças, tipos de acesso, integração com outros sistemas e por último e não menos importante, os metadados. A atribuição de metadados segue geralmente esquemas de metadados adotados por cada instituição em específico [15].

Cabe destacar uma funcionalidade bastante importante relativamente aos RI, que potencializa a troca de informação e assegura a interoperabilidade dos conteúdos abertos. Trata-se de protocolos

[15] http://projeto.rcaap.pt/index.php/lang-pt/como-auto-arquivar-documentos/introducao-3

como o *Open Access Initiative – Protocol for Metadata Harvesting (OAI-PMH)*. Esse protocolo permite a coleta de metadados de forma interoperável e está presente em diversos repositórios.

2.3.3 Iniciativas

A implementação de repositórios institucionais tem vindo a crescer fomentada principalmente por organismos internacionais em apoio ao acesso aberto de dados. Dentre as inúmeras iniciativas nesse sentido destacam-se entre outras:

- ***B2Share***[16]. O projeto é a solução do *EUDAT* para comunidades científicas e cientistas armazenarem e compartilharem dados de pesquisa de pequena escala e de diversos contextos.
- ***ICPSR - Inter University Consortium for Political and Social Research***[17]. Composto por cerca de 776 universidades, agências governamentais e outras instituições, é um repositório que armazena, organiza e fornece acesso a dados científicos com o objetivo de validar os resultados de pesquisas e a reutilização dos dados há mais de cinco décadas.
- ***RCCAP* - Repositório Científico de Acesso Aberto de Portugal**[18]. A iniciativa portuguesa de acesso aberto em Portugal nasceu em julho de 2008 promovida pela *UMIC* - Agência para a Sociedade do Conhecimento e operacionalizada pela *FCCN* - Fundação para a Computação Científica Nacional, com o apoio da Universidade do Minho, tendo sido apresentada e lançada oficialmente em dezembro do mesmo ano.
- ***re3data - Registry of Research Data Repositories***[19]. O re3data.org, lançado em 2012, é um conjunto de repositórios de dados de pesquisa composto por diferentes disciplinas acadêmicas de abrangência internacional. É financiado pela *German Research Foundation* - Fundação Alemã de Pesquisa. Apresenta repositórios para o armazenamento permanente e acesso de conjuntos de dados a pesquisadores, órgãos financiadores, editores e instituições académicas.
- ***DSpace***[20]. Lançado em 2002, é uma plataforma de software de código aberto disponibilizada para instituições acadêmicas sem fins lucrativos e comerciais que criam repositórios digitais abertos. Foi desenvolvido por *MIT Libraries e Hewlett-Packard*. O *DSpace* tornou-se um projeto do *DuraSpace* em 2009, quando as organizações *Fedora Commons* e *DSpace* fundiram-se para formar o *DuraSpace*.
- ***Fighshare***[21]. Voltado para académicos, instituições e editores o *Fighshare*, sob a égide do acesso aberto ao conhecimento, oferece serviços de gestão de dados de pesquisa académica e disseminação de dados.

[16] https://www.eudat.eu/services/b2sharet
[17] https://www.icpsr.umich.edu/icpsrweb/content/about/
[18] http://projeto.rcaap.pt/index.php/lang-pt/sobre-o-rcaap/enquadramento
[19] https://www.re3data.org/about
[20] https://duraspace.org/dspace/about/
[21] https://knowledge.figshare.com/articles/item/what-is-figshare

- ***Zenodo***[22]. O consórcio OpenAire em parceria com o *CERN - European Organization for Nuclear Research*, desenvolveu o repositório *Zenodo*, que torna a partilha, curadoria e publicação de dados e software uma realidade para todos os pesquisadores. O *Zenodo* não impõe nenhum requisito quanto a formato, tamanho, restrições de acesso ou licença. Permite a publicação de conteúdo fechado e restrito, para que os artefatos possam ser capturados e armazenados com segurança enquanto a pesquisa estiver em andamento. Dessa forma, os *links* protegidos podem ser compartilhados com os revisores e o conteúdo também pode ser embargado e aberto automaticamente quando o documento associado é publicado.

2.3.4 Suporte a metadados específicos de domínio

A diversidade de domínios de publicações científicas exige que os RI estejam preparados a atender as mais variadas disciplinas.

Um estudo apresentado pelo *DCC* Ltd (2010) identificou quatro áreas disciplinares no compartilhamento de dados de pesquisa, reutilização e viabilidade a longo prazo, nomeadamente: Artes e Humanidades, Ciências Sociais, Ciências da Vida e Ciências Físicas.

Além de reconhecer as diferenças disciplinares existentes nos processos de curadoria e partilha de dados destacou ainda outros fatores que influenciam a curadoria de dados:

1. Tipo e quantidade de dados produzidos
2. Singularidade dos dados e a potencialidade em serem reutilizados
3. Património e práticas das comunidades de pesquisa
4. Adoção de formatos e esquemas de metadados e padrões por cada comunidade

De acordo com Shakeri and Gracy (2014), ainda sobre o estudo citado, o achado mais crítico refere-se à conclusão de que uma abordagem genérica para a curadoria de dados não é aplicável a todas as disciplinas. Assim, seria necessário o desenvolvimento de estratégias específicas de acordo com o domínio de cada comunidade de pesquisa, a partir da investigação local das necessidades, expectativas e práticas de dados de seus pesquisadores.

A *RDA - Research Data Alliance*[23] aborda as especificidades das disciplinas científicas. "De todas as disciplinas científicas, desde a herança cultural até a agricultura, da engenharia aeroespacial à biologia marinha, cada disciplina possui uma abordagem específica para o planeamento, aquisição, processamento e armazenamento dos dados coletados e gerados.

No mesmo sítio da comunidade *RDA* pode ser encontrada a definição de sete disciplinas que permeiam o compartilhamento de dados: Agricultura; Linguística; Ciências Biomédicas; Química; Humanidades Digitais; Biblioteconomia, Arquivística e Ciência da Informação; Ciências Sociais.

Nesse cenário de múltiplos domínios a especificidade é um fator que cria necessidades diversas ensejando a criação de metadados específicos a cada domínio.

[22] https://about.zenodo.org/
[23] https://www.rd-alliance.org/rda-disciplines

CLASSIFICAÇÕES	De acordo com o seu conteúdo	- **Classificações enciclopédicas** (CDU, CDD, LCC) - **Classificações especializadas** (Classificação médica da *National Library of Medicine*, Classificação da OCDE)
	De acordo com a sua estrutura	- **Classificações enumerativas** (LCC, CDD) - **Classificações por facetas** (Classificação de Colon) - **Classificações mistas** (CDU)

Figura 2.7: Medeiros, 2014, Tipologia das Classificações

2.4 Classificação

O estudo das classificações passa pelas classificações dos saberes, ou filosóficas, classificações científicas, também chamadas de taxonómicas, relativas à classificação dos seres e às classificações bibliográficas. As classificações bibliográficas surgiram no século XIX e até meados do século XX foram consideradas as principais estruturas de organização do pensamento (Simões, 2010).

2.4.1 Classificações bibliográficas

Relativamente à Ciência da Informação, as classificações bibliográficas são naturalmente, objeto de estudo aprofundado. (Medeiros, 2014) apresenta na figura 2.7, a tipologia das classificações bibliográficas relativamente ao seu conteúdo e estrutura.

A divisão quanto ao conteúdo está em conformidade com a sua abrangência. Quando procuram lidar com todas as áreas do conhecimento são chamadas enciclopédicas. E quando tratam uma determinada área particular de conhecimento são designadas classificações especializadas.

Quanto à sua estrutura, podem ser enumerativas se listarem de forma linear e exaustiva todas as matérias abrangidas. São consideradas facetadas quando decompõem todos os domínios e suas respetivas partes componentes. Há ainda as classificações mistas que são o resultado da aglutinação dos dois sistemas enunciados, ou seja, caracterizam-se por uma componente enumerativa aliada ao emprego de facetas.

(Simões, 2010) destaca que a classificação bibliográfica tem como principal objetivo organizar o conhecimento humano em grandes classes epistemológicas e ao mesmo tempo tem o papel de organizá-lo fisicamente nas estantes numa biblioteca. E não menos importante, a classificação tem ainda a função de recuperar a informação.

No âmbito da Teoria da Classificação, Campos and Gomes (2003) chamam a atenção para a teoria de Ranganathan ao apresentar uma nova forma de organizar o universo de assuntos. A

classificação dicotômica/binária ou decatômica são refutadas tendo a vista que os assuntos dos documentos não fazem parte de um domínio único de conhecimento. Assim, Ranganathan apresenta a policotomia ilimitada. Em outras palavras, apresenta a Teoria da Classificação Facetada. Para clarificar sua ideia, faz analogia com a árvore baniana na qual do tronco original formam-se muitos outros troncos secundários de tempos em tempos.

As primeiras grandes classificações datam dos fins do século XIX e princípios do século XX. Entre as mais conhecidas podem ser distinguidas, a Classificação Decimal de *Dewey* (CDD), a Classificação da Biblioteca do Congresso (LCC), a Classificação Bibliográfica de *Bliss* e a Classificação Decimal Universal (CDU).

A título de exemplo, serão brevemente explanados dois grandes sistemas bibliográficos:

- ***Universal Decimal Classification***. A Classificação Decimal Universal, é um reconhecido sistema de classificação bibliográfica a qual (Simões, 2010) distingue seu objetivo inicial como a compilação de um repositório universal de bibliografia ordenado por temas. A CDU é considerada um instrumento de divulgação de toda a literatura científico-técnica produzida até então. Contudo, a autora tece críticas à estrutura e ao conteúdo da ferramenta especialmente a rigidez das classes e notações e a respetiva dificuldade em representar novos conceitos.

- ***Library of Congress Classification***. O sistema de classificação da Biblioteca do Congresso surgiu de um modelo adotado por Thomas Jefferson em sua biblioteca particular, a qual viria mais tarde ser incorporada pela Biblioteca do Congresso dos Estados Unidos e consequentemente sua classificação. O Sistema foi adotado por grandes bibliotecas académicas naquele país.

2.4.2 Classificações científicas

Ainda conforme (Simões, 2010), as classificações naturalistas pretendem esgotar todo o conhecimento relativo a um tema na respetiva classe. Cita o exemplo da classificação proposta por Lineu ao apresentar a hierarquia da classificação científica dos seres vivos. Salienta também o fato dessas classificações simplesmente enumerarem os assuntos, enfatizando a relação hierárquica, concorrendo, deste modo, para outra característica – a exclusividade.

Souza (2006) analisou esquemas de classificação bibliográfica, tabelas de classificação de áreas do conhecimento específicas de Ciência e Tecnologia e também esquemas de classificação dos saberes no contexto da educação com o objetivo de identificar as grandes áreas do conhecimento em Ciência e Tecnologia (C&T).

O estudo concluiu que há consenso na agregação de áreas em grandes áreas do conhecimento em C&T. Entretanto, ressalta que há dificuldade na definição do número e na ordem de apresentação das grandes áreas em virtude da função da natureza do objeto de representação, e também da finalidade da organização do conhecimento.

Segundo suas palavras, "a pesquisa na área de organização e representação do conhecimento no contexto de gestão e avaliação não deixa de ser ainda um desafio para os cientistas da informação".

Sem entrar no mérito do desafio que o tema sugere, são descritas a seguir, duas iniciativas de classificação científica no campo da tecnologia, pesquisa e inovação.

- ***Organisation for Economic Co-operation and Development***[24]. A Organização para Cooperação e Desenvolvimento Económico (OCDE) é uma organização internacional que trabalha para construir melhores políticas fundamentada na pesquisa e desenvolvimento experimental. Busca através do trabalho criativo e sistemático aumentar o estoque de conhecimento - incluindo o conhecimento da humanidade, cultura e sociedade.

  Entre os inúmeros serviços oferecidos pela Organização está o conhecido Manual *Frascati*, ferramenta essencial para formuladores de políticas de ciência e inovação em todo o mundo. O manual disponibiliza a classificação *Fields of Science and Technology* para compilar estatísticas de Pesquisa e Desenvolvimento.

- ***Association for Computing Machinery - ACM***[25]. A *ACM* é uma Associação Científica e Educacional de âmbito internacional dedicada ao avanço da arte, ciência, engenharia e aplicação da tecnologia da informação. Além de atender a interesses profissionais e públicos, promove o intercâmbio aberto de informações e os mais altos padrões profissionais e éticos.

  O sistema de classificação da entidade, *Computing Classification System (CCS)* é voltado para assuntos da Ciência da Computação. O primeiro sistema de classificação da *ACM* foi publicado em 1964. Em 1982 publicou um sistema inteiramente novo e em 2012 uma nova estrutura poli-hierárquica e uma abordagem mais aprofundada foi disponibilizada. O esquema de 1998 foi entretanto, mapeado para o de 2012.

2.4.3 Organização do conhecimento no ambiente digital

A organização do conhecimento é um desafio ininterrupto e que a cada dia torna-se mais complexo, pois a produção de conhecimento é um processo contínuo, transformado sobretudo pela evolução da tecnologia. Campos and Gomes (2003) afirmam que a forma de representação e organização dos diferentes domínios de conhecimentos implicam diretamente na recuperação da informação em meios eletrónicos. Portanto, a organização do conhecimento está intimamente ligada à recuperação da informação.

As classificações conforme tratado na secção anterior são uma forma incontestável de organização do conhecimento. Novas estruturas de organização do conhecimento, como é o caso das taxonomias, tesauros e ontologias são assentadas nos princípios das classificações.

[24] https://www.oecd.org/sti/inno/frascati-manual.htm
[25] https://www.acm.org/about-acm/acm-history

Para Medeiros (2014), ”uma taxonomia é, por definição, uma classificação ou categorização de um conjunto de coisas organizadas de forma hierárquica. No campo das Ciências da Informação e Documentação - CID, uma taxonomia é uma lista de termos preferenciais com estrutura hierárquica”.

A autora explica que “os tesauros constituem-se como sistemas de organização do conhecimento inseridos nos grupos de relações.[...] A adaptação tecnológica dos tesauros enquanto esquemas de organização da informação tornou-se absolutamente fundamental, sobretudo para responder aos novos desafios trazidos pelo advento da *web*, em particular da "Web Semântica”.

Para Maculan et al. (2017), no âmbito da Ciência da Informação, "as ontologias são também um tipo de Sistema de Organização do Conhecimento (SOC), desenvolvidas como um modelo conceptual e base de conhecimento de um domínio, para serem interpretáveis pela máquina".

No estudo sobre instrumentos para organização do conhecimento, Maria and Tristão (2004) analisa o sistema de classificação facetada e os tesauros. Conclui que o tesauro e o sistema de classificação apresentam diferentes níveis e profundidade de organização, mas que podem coexistir em um sistema de recuperação de informação, complementando um as deficiências do outro.

Para Souza and Pestana (2017), os tesauros são sistemas estruturados que podem potencializar a recuperação da informação sendo caracterizados por "forte estrutura semântica". Podem ser referenciados como a "primeira linguagem de indexação mais flexível e adaptável ao contexto digital".

Capítulo 3

Abordagem metodológica

3.1 Objetivo e descrição

A presente iniciativa de classificação de recursos informacionais começou por ser idealizada a atender especificamente à aplicação Dendro, uma solução *open-source* desenvolvida pelo InfoLab da FEUP para construção de um repositório digital e gestão de dados de investigação.

A aplicação é utilizada para gerir dados de investigação de diferentes domínios através da inserção de metadados de diversas ontologias. A fim de facilitar o trabalho do investigador durante a inclusão de seu projeto na plataforma Dendro, foi verificada a necessidade de uma funcionalidade que pudesse sugerir automaticamente os melhores descritores de acordo com cada novo projeto.

Assim, a ideia é construir uma ferramenta que dê suporte ao Dendro, inicialmente a partir da classificação dos domínios de cada projeto. Ou seja, permitir uma pré-configuração automática do Dendro tornando a plataforma mais interativa de acordo com as necessidades de cada investigador.

A título de exemplo, se o investigador estiver a incluir um projeto que trata de conteúdo áudio visual, o Dendro poderá sugerir o preenchimento de descritores presentes na ontologia *Audio Visual Contents*.

3.2 Um sistema *on line* para classificação de documentos

A figura 3.1 ilustra a metodologia aplicada na construção do sistema, nomeadamente as ferramentas e tecnologias utilizadas e as etapas executadas.

Assim, as próximas subseções discorrem em detalhes a figura mencionada, sistematizando a escolha das ferramentas e tecnologias utilizadas e as respectivas etapas.

3.2.1 Ferramentas e tecnologias

As ferramentas utilizadas no decorrer do projeto não foram selecionadas previamente. Em virtude da demanda, os softwares foram selecionados levando-se em consideração dois quesitos principais: software de acesso aberto que não tivesse custo agregado e também facilidade de

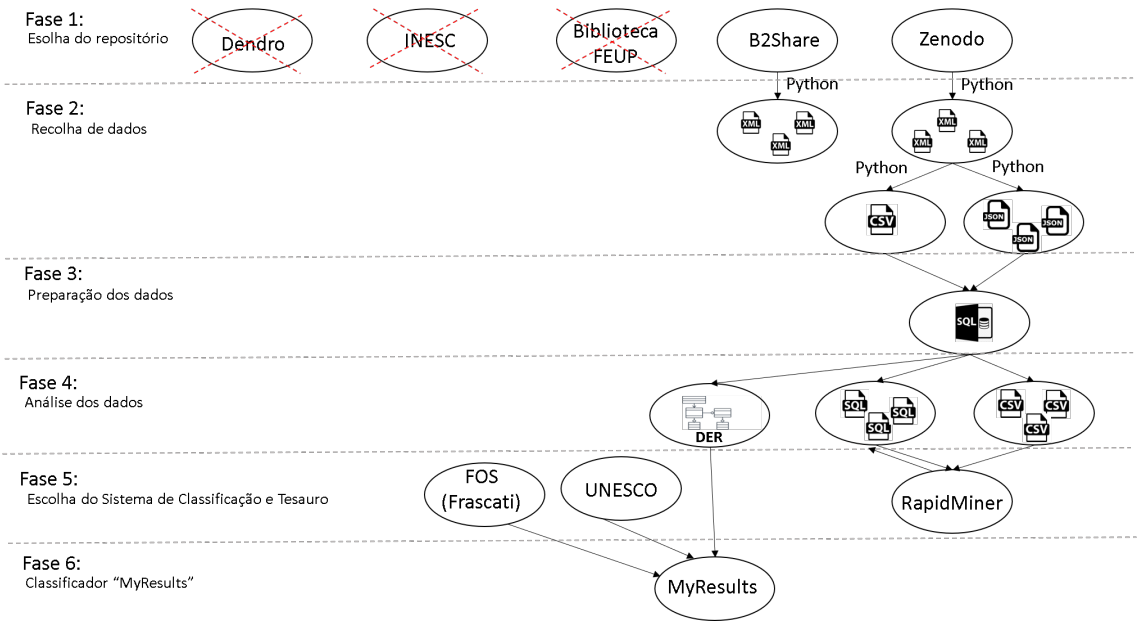


Figura 3.1: Etapas de construção do *site*

uso, nomeadamente experiência anterior em lidar com a ferramenta de forma a garantir maior produtividade.

A primeira ferramenta utilizada foi o *Python* a fim de executar a extração dos dados através do protocolo *OAI-PMH*.

A etapa seguinte, de extração de dados na *API* em formato *JSON* também foi desenvolvida com o *Python*. Toda a codificação utilizou a versão 3.7 no ambiente *JetBrains PyCharm Community Edition 2018.2.4 x64*.

Naquela altura, era suposto utilizar a plataforma de software *RapidMiner* e a preparação dos dados começou a ser executada no próprio *Python*. Entretanto, foi constatada a necessidade de utilização de uma base de dados. Inicialmente verificou-se a possibilidade de utilizar uma biblioteca de suporte a banco de dados no *Python*, como por exemplo *Sqlite*, uma vez que o *Python* não oferece suporte nativo a *SQL*. Outra possibilidade levantada foi a utilização de um banco de dados *NoSQL*, como por exemplo, o *MongoDb*.

Antes porém, foi realizada uma experiência prévia no *SQL Server* e verificou-se que o suporte a *JSON* na plataforma era muito intuitivo. Porém, o fator decisivo na escolha dessa ferramenta foi a experiência já adquirida em lidar com bases de dados *SQL*, notadamente o *SQL Server*. A prática adquirida em trabalhar com o software proporcionava um rendimento muito superior frente ao aprendizado de uma nova ferramenta.

Outro requisito fundamental à definição do software de banco de dados foi a integração ao *RapidMiner*. A versão utilizada do *SQL Server* é a de número 14.0.1000.169 juntamente com o *Microsoft SQL Server Management Studio*, versão 14.0.17213.0. Por fim, outro ponto positivo na escolha da ferramenta decorreu da conexão do *SQL Server* com o *RapidMiner* ser muito fácil de

configurar. Ainda sobre a escolha do *SQL* convém ressaltar que o ambiente utilizado para executar o *Python* era demasiadamente pesado e frequentemente apresentava mensagem de insuficiência de memória no portátil em uso.

Relativamente ao *RapidMiner* foi adquirida uma licença a título educacional através da Universidade do Porto, intitulada *RapidMiner Studio Educational 8.0.001*. Cumpre ressaltar que não foi realizado nenhum estudo sobre a utilização das ferramentas em ambiente não institucional especialmente no que concerne a custos de licenciamento.

Durante o desenvolvimento do projeto sentiu-se a necessidade de uma ferramenta para visualização dos caminhos permeados e análise dos resultados. O banco de dados começou a crescer significativamente e a visualização do resultado das *queries* precisava de uma ferramenta visual. Assim, optou-se pela utilização do *Microsoft Visual Studio Community 2017*, versão 15.5.6 e *Microsoft .NET Framework*, versão 4.7.03056.

3.2.2 Etapas

O desenvolvimento do projeto, conforme se pode observar na figura, é composto por seis etapas. Começa com a escolha do repositório, recolha, preparação e análise dos dados, seguida da escolha do sistema de classificação e tesauro culminando com a construção do portal chamado de "MyResults".

3.2.2.1 Escolha do repositório

O sistema classificador a ser desenvolvido conforme explicado anteriormente, tem o objetivo de fornecer suporte à aplicação Dendro, nomeadamente a sugestão dos melhores descritores de metadados a cada projeto de investigação a ser criado na plataforma.

O primeiro passo deste projeto consistiu em escolher a fonte de dados a ser utilizada no processo classificatório. A seleção naturalmente, recaiu sobre o próprio repositório. Decorre que na análise inicial, a quantidade de registos do referido repositório foi considerada insatisfatória por apresentar um número pequeno de registos.

De seguida, buscou-se a adesão de dados do Repositório do Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência - *INESC TEC*. A solução também foi desconsiderada em virtude da quantidade insuficiente de registos.

A busca de outro repositório caminhou em direção ao *B2Share*, infraestrutura colaborativa de dados da iniciativa *EUDAT*. À partida, a plataforma continha registos que ofereciam possibilidade a um trabalho mais abrangente. Contudo, outro entrave surgiu. Mesmo através de suporte à recolha de dados através de *API's* e protocolo *OAI-PMH*, os metadados disponibilizados não ofereciam mapeamento ao padrão *Dublin Core*, informação confirmada após contatos através de *email* com a entidade para esclarecimento desta questão.

Mediante a restrição do *B2Share* foi avençada a alternativa de utilização dos dados do sistema da Biblioteca da FEUP a qual traduziu-se também inexequível, pois não havia como aceder aos

XML	JSON	Perda
253.801	217.640	14.25%

Tabela 3.1: Número de objetos *XML* e *JSON*

dados daquele departamento, uma vez que a plataforma não conta com uma *API* para recolha dos dados.

Por fim, foi idealizada a utilização de dados do Repositório *Zenodo*, que após os testes iniciais mostrou-se adequado à tarefa classificatória em decurso. Em virtude da maior quantidade de dados a opção pareceu mais favorável à iniciativa proposta.

3.2.2.2 Recolha dos dados

Conforme relatado, o repositório escolhido para realizar o trabalho foi o *Zenodo*, mas antes de se chegar a essa conclusão ocorreu a recolha de dados no *B2Share*.

O procedimento de recolha dos dados ocorreu através do protocolo *OAI-PMH*, portanto em formato *XML* usando a linguagem *Python*. Os ficheiros foram depositados em um diretório local no portátil utilizado no estudo e posteriormente na plataforma *GitHub*. A extração de dados no *B2Share* resultou no total de 9.845 ficheiros.

Entretanto, nova recolha foi executada, desta vez no *Zenodo*. À semelhança da primeira tarefa, foram usados os mesmos procedimentos, protocolo *OAI-PMH* e a linguagem *Python*. Os ficheiros foram armazenados no portátil de estudo e também na plataforma *GitHub*. A extração resultou em 253.801 ficheiros em formato *XML*.

De seguida, novamente no *Python*, após a conclusão do *download*, os ficheiros foram convertidos em um único ficheiro em formato *CSV*.

Após a recolha e análise dos dados em *XML*, verificou-se no sítio do *Zenodo* a possibilidade de extração dos dados do repositório no formato *JSON*. Verificou-se também que os dados no formato *JSON* apresentavam maior riqueza de descritores de metadados.

Além dos tradicionais descritores de metadados, os dados no formato *JSON* possuíam atributos relativos ao controle de versões dos ficheiros, característica importante que permite execução de diversos filtros. Ou seja, pode-se trabalhar com todas as versões de cada projeto do repositório ou por exemplo, escolher a versão mais atualizada de cada projeto.

Assim, em virtude da possibilidade de dados mais completos, uma nova recolha de dados foi executada. A partir do nome de cada ficheiro *XML*, foi feito o *download* do respetivo ficheiro em *JSON*.

O processo de coletar os ficheiros *JSON* foi um processo demorado com erros de conexão motivadas pela política de segurança do repositório. Teve portanto que ser retomado diversas vezes.

É apresentado na tabela 3.1 os dados estatísticos da fase inicial de recolha no *Zenodo*. A quantidade de ficheiros *JSON* é menor devido a estas falhas de conexão. A diferença na quantidade de ficheiros no formato *XML* e *JSON* chega a uma taxa de 14.25% de perda.

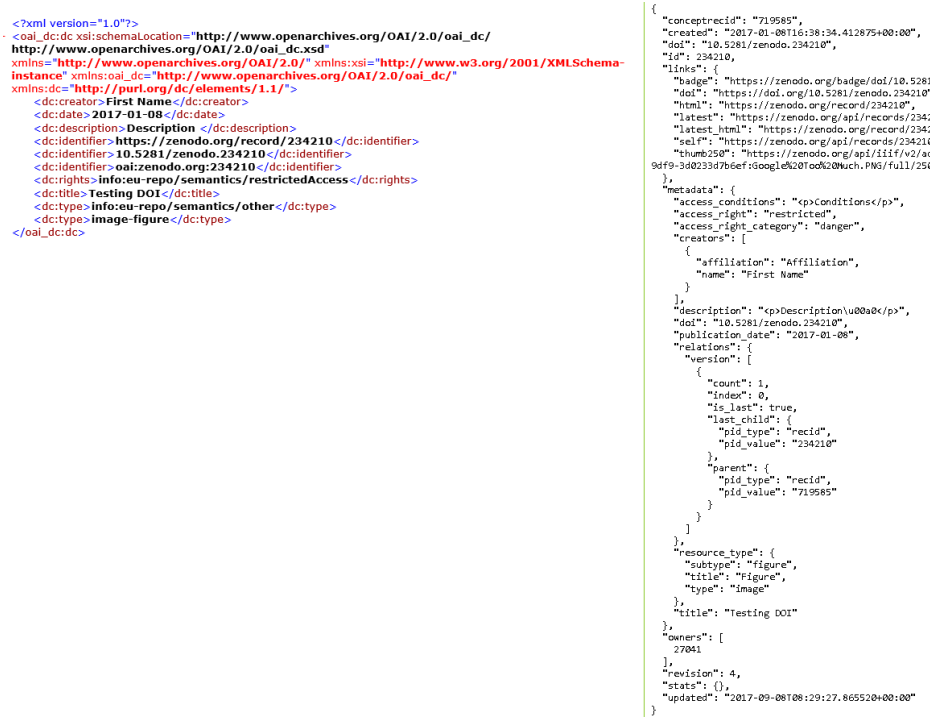


Figura 3.2: Comparando o mesmo registo em *XML* e *JSON*

A título de curiosidade, ao se comparar a quantidade de ficheiros extraídos do repositório *B2Share* no total de 9.845 ficheiros e a quantidade de 253.801 ficheiros no *Zenodo* pode-se perceber facilmente o aumento significativo no universo da amostragem.

A fim de clarificar a diferença entre os descritores em ambos os formatos, a figura 3.2 exibe a comparação do registo relativo ao projeto com identificação igual a 234210.

Do lado esquerdo é apresentado o registo no formato *XML* e do lado direito da figura, está representado o mesmo objeto digital no formato *JSON*. Visualmente pode-se perceber a maior quantidade de informação no ficheiro em formato *JSON*.

3.2.2.3 Preparação dos dados

Após a recolha dos dados, o próximo passo era lidar com os dados. Conforme referido anteriormente, a primeira opção foi utilizar o *Python* para preparar os dados. Mostrando-se inviável foi utilizado o *SQL Server*.

Tanto os dados em *XML* já agrupados no ficheiro *CSV* quanto os dados no formato *JSON* foram importados para a base de dados. Verificou-se entretanto, que a conversão dos ficheiros em *XML* para o ficheiro *CSV* apresentava erros.

Uma opção seria importar os ficheiros em formato nativo *XML* para a base de dados. A preocupação em lidar com um padrão internacional como é o caso dos resultados OAI-PMH em *XML* acabou sendo colocada em segundo plano e a escolha recaiu em trabalhar com os ficheiros *JSON*. A escolha deteve-se à possível oportunidade de que os dados complementares pudessem oferecer

Formato XML	Formato JSON

<dc:contributor>The SAFE Project</dc:contributor>	"metadata": {
	"contributors": [{
	"affiliation": "Imperial College London",
	"name": "The SAFE Project",
	"orcid": "0000-0003-3378-2814",
	"type": "ContactPerson"
	}],

<dc:creator>Fayle, Tom</dc:creator>
<dc:creator>Ewers, Robert</dc:creator>
	"metadata": {
	"creators": [{
	"name": "Fayle, Tom"
	}, {
	"affiliation": "Imperial College London",
	"name": "Ewers, Robert"
	}],
	
	"metadata": {"owners": [
	48652],

<dc:date>2018-04-30</dc:date>	"created": "2018-04-30T12:17:36.309699+00:00", "metadata": { "publication_date": "2018-04-30", "updated": "2018-07-03T12:56:52.094657+00:00",
<dc:description>Description: Bait card records of ant ...	"metadata": { "description": "Description: Bait card records of ant communities ...

<dc:rights>info:eu-repo/semantics/openAccess</dc:rights>
<dc:rights>https://creativecommons.org/licenses/by/4.0/</dc:rights>
	"metadata": {
	"access_right": "open",
	"access_right_category": "success",
	"license": {
	"id": "CC-BY-4.0"},

<dc:title>How does forest conversion and fragmentation affect ant communities and the ecosystem processes that they mediate?</dc:title>	"metadata": { "title": "How does forest conversion and fragmentation affect ant communities and the ecosystem processes that they mediate?",

<dc:relation>doi:10.5281/zenodo.1237729</dc:relation>
<dc:relation>url:https://zenodo.org/communities/safe</dc:relation>
	"metadata": {
	"doi": "10.5281/zenodo.1237730",
	"relations": {
	"version": [{
	"count": 1,
	"index": 0,
	"is_last": true,
	"last_child": {
	"pid_type": "recid",
	"pid_value": "1237730"},
	"parent": {
	"pid_type": "recid",
	"pid_value": "1237729"}}]},
	"revision": 3,
	"stats": {
	"downloads": 5.0,
	"unique_downloads": 5.0,
	"unique_views": 15.0,
	"version_downloads": 5.0,
	"version_unique_downloads": 5.0,
	"version_unique_views": 15.0,
	"version_views": 15.0,
	"version_volume": 1438660.0,
	"views": 15.0,
	"volume": 1438660.0 },

Figura 3.3: Comparando metadados em formato *XML* e *JSON*

um ganho significativo na análise dos dados. Convém destacar que o mapeamento dos campos entre os dois formatos foi previamente elaborado, tendo como base o padrão *Dublin Core*, conforme se pode verificar na figura 3.4.

Ainda sobre a escolha, um motivo que corroborou a dispensa do formato *XML* prendeu-se à capacidade do formato *JSON* oferecer uma estrutura mais ágil de manipulação e uma hierarquização muito fácil de lidar, especialmente no *SQL Server*.

Cumpre destacar que além da perda inicial de registos durante o *harvesting* no sítio do repositório para download dos ficheiros em *JSON*, outros 96 ficheiros foram ignorados por apresentarem erros de formatação. Assim, a quantidade total de ficheiros válidos no *SQL* é exatamente igual a 217.447, uma vez que três ficheiros foram adicionados ao *SQL* manualmente e outros 100 também foram postos de lado por não apresentarem os metadados mínimos necessários ao trabalho. Cada ficheiro deveria possuir ao menos um de três metadados, nomeadamente *Title*, *Description* ou *keyword*.

Formato XML	Formato JSON
<dc:subject>ant</dc:subject> <dc:subject>scavenging</dc:subject> <dc:subject>community composition</dc:subject>	"metadata": { "keywords": ["ant", "scavenging", "community composition"],

<dc:type>info:eu-repo/semantics/other</dc:type>
<dc:type>dataset</dc:type>
	"metadata": {
	"resource_type": {
	"title": "Dataset",
	"type": "dataset"},

<dc:identifier>https://zenodo.org/record/1237730</dc:identifier>
<dc:identifier>10.5281/zenodo.1237730</dc:identifier>
<dc:identifier>oai:zenodo.org:1237730</dc:identifier>
	"files": [{
	"bucket": "1f6dc58f-4060-4d0d-8efc-0edba601706b",
	"checksum": "md5:05a209726c2a727558c71cb1dc1c2908",
	"key": "template_Fayle_BaitCards.xlsx",
	"links": {
	"self": "https://zenodo.org/api/files/1f6dc58f-4060-4d0d-8efc-0edba601706b/template_Fayle_BaitCards.xlsx"},
	"size": 287732,
	"type": "xlsx"}],
	
	"id": 1237730,
	
	"links": {
	"badge": "https://zenodo.org/badge/doi/10.5281/zenodo.1237730.svg",
	"bucket": "https://zenodo.org/api/files/1f6dc58f-4060-4d0d-8efc-0edba601706b",
	"conceptbadge": "https://zenodo.org/badge/doi/10.5281/zenodo.1237729.svg",
	"conceptdoi": "https://doi.org/10.5281/zenodo.1237729",
	"doi": "https://doi.org/10.5281/zenodo.1237730",
	"html": "https://zenodo.org/record/1237730",
	"latest": "https://zenodo.org/api/records/1237730",
	"latest_html": "https://zenodo.org/record/1237730",
	"self": "https://zenodo.org/api/records/1237730" },
	
	"conceptdoi": "10.5281/zenodo.1237729",
	"conceptrecid": "1237729",
	
	"doi": "10.5281/zenodo.1237730",
	
	"metadata": {
	"related_identifiers": [{
	"identifier": "10.5281/zenodo.1237729",
	"relation": "isVersionOf",
	"scheme": "doi"
	}],
	"communities": [{
	"id": "safe"}
	],

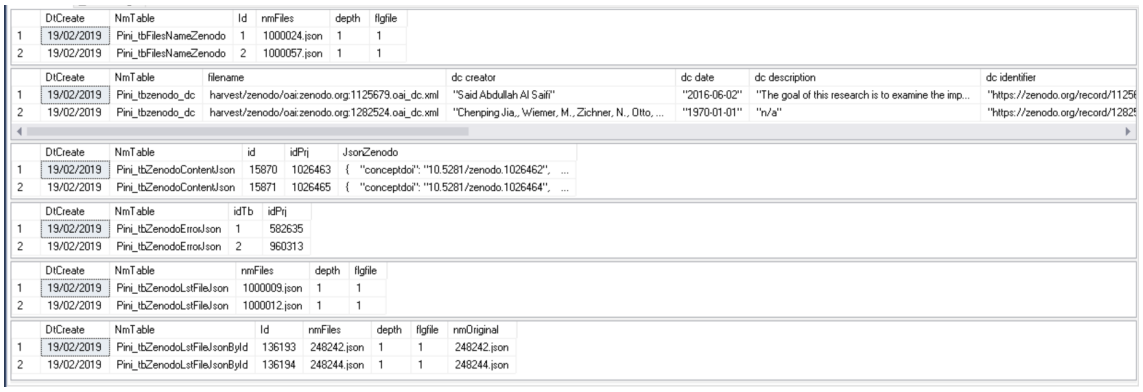
Figura 3.4: Comparando metadados em formato *XML* e *JSON*

As tabelas iniciais que compõem a base de dados podem ser divididas em categorias. No primeiro grupo, conforme se verifica na figura 3.5, estão listadas a tabelas geradas a partir dos ficheiros originais que foram baixados do *Zenodo*. Estas tabelas representam os resultados da importação direta dos dados.

No segundo grupo estão listadas as tabelas que foram criadas a partir da identificação de campos importantes do ficheiro com as informações *JSON*. Em outras palavras, foram analisados os possíveis metadados que poderiam ser objeto de estudo mais detalhado. A partir da análise, aqueles referenciados como importantes, foram acondicionados em tabelas próprias. A figura 3.6 demonstra as tabelas criadas nessa fase do projeto.

Em ambos os casos, estão exemplificadas duas instâncias de cada tabela com a respetiva data de criação da tabela, seu nome e os respetivos atributos.

A título de exemplo, pode-se observar no destaque em amarelo da figura, a referência à tabela do banco de dados com o nome *"Pini2_tbZenodoAlmostAll"*. A primeira coluna indica a data de criação da tabela, isto é, dia 02/04/2019. A segunda coluna tem a informação do nome da tabela e a terceira faz referência a uma chave de controlo interna. Todos os demais campos, entre eles, *idPrj*,*conceptDoi*,*conceptId*,*created*,*doi*,*id*,*linkBadge*, são aqueles selecionados a partir do ficheiro *JSON* como campos da *"Pini2_tbZenodoAlmostAll"*. Essa tabela é importante porque reúne os



	DtCreate	NmTable	Id	nmFiles	depth	flgfile
1	19/02/2019	Pini_tbFilesNameZenodo	1	1000024.json	1	1
2	19/02/2019	Pini_tbFilesNameZenodo	2	1000057.json	1	1

	DtCreate	NmTable	filename	dc creator	dc date	dc description	dc identifier
1	19/02/2019	Pini_tbzenodo_dc	harvest/zenodo/oai:zenodo.org:1125679.oai_dc.xml	"Said Abdullah Al Saifi"	"2016-06-02"	"The goal of this research is to examine the imp...	"https://zenodo.org/record/1125
2	19/02/2019	Pini_tbzenodo_dc	harvest/zenodo/oai:zenodo.org:1282524.oai_dc.xml	"Chenping Jia,, Wiemer, M., Zichner, N., Otto, ...	"1970-01-01"	"n/a"	"https://zenodo.org/record/1282

	DtCreate	NmTable	id	idPrj	JsonZenodo
1	19/02/2019	Pini_tbZenodoContentJson	15870	1026463	{ "conceptdoi": "10.5281/zenodo.1026462", ...
2	19/02/2019	Pini_tbZenodoContentJson	15871	1026465	{ "conceptdoi": "10.5281/zenodo.1026464", ...

	DtCreate	NmTable	idTb	idPrj
1	19/02/2019	Pini_tbZenodoErrosJson	1	582635
2	19/02/2019	Pini_tbZenodoErrosJson	2	960313

	DtCreate	NmTable	nmFiles	depth	flgfile
1	19/02/2019	Pini_tbZenodoLstFileJson	1000009.json	1	1
2	19/02/2019	Pini_tbZenodoLstFileJson	1000012.json	1	1

	DtCreate	NmTable	Id	nmFiles	depth	flgfile	nmOriginal
1	19/02/2019	Pini_tbZenodoLstFileJsonById	136193	248242.json	1	1	248242.json
2	19/02/2019	Pini_tbZenodoLstFileJsonById	136194	248244.json	1	1	248244.json

Figura 3.5: Tabelas *SQL* - Importação de dados

metadados mais importantes de cada objeto digital a ser classificado.

Na figura 3.7 pode-se verificar outro grupo de tabelas do banco de dados utilizado para uma análise mais aprofundada dos metadados originais. O objetivo nessa altura era estabelecer relação entre os metadados e tentar avaliar enviesamento dos dados.

A limpeza dos dados foi realizada tanto no *SQL* aquando da geração dos *scripts* como também dentro do próprio *RapidMiner*. O processo deteve-se aos procedimentos básicos de retirar registos inválidos, substituir caracteres de formatação, como por exemplo *tags HTML* a configurar texto em negrito, espaçamento e afins. No *RapidMiner* foram utilizados os operadores *Filter Example*, *Select Attributes*, *Example Range*.

3.2.2.4 Análise dos dados

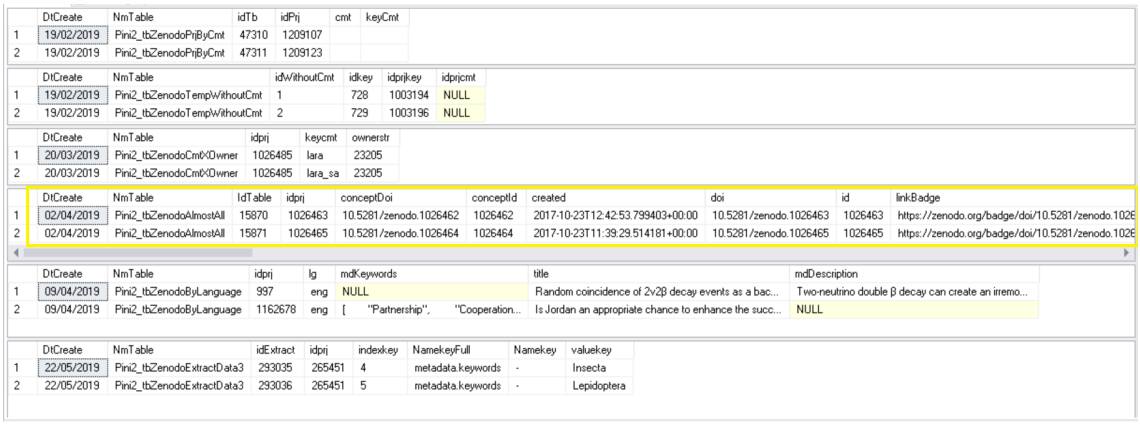
A fase relativa à análise dos dados está descrita em duas subseções. São abordadas as experiências realizadas no *RapidMiner* e a preparação do Diagrama de Entidade Relacionamento.

A. Experiências no *RapidMiner*

As experiências iniciais no *RapidMiner* prenderam-se a encontrar padrões a partir dos operadores *FP-Growth*, *Create Association*, *Correlation Matrix*. A princípio, os *datasets* utilizados foram ficheiros *CSV*, como aquele gerado a partir do *harvest* inicial ao *Zenodo* e também ficheiros gerados pelo *SQL*.

Cumpre salientar que a primeira licença adquirida da ferramenta tinha limitação à quantidade de dados a serem processados e somente 10.000 registos poderiam ser trabalhados por vez. Essa limitação interferia diretamente na escolha dos dados e respetiva amostragem. O problema foi contudo resolvido ao entrar em contato com o serviço de apoio e a licença sofrer *upgrade*.

A dificuldade em importar os ficheiros gerados pelo *SQL* no formato *csv* para o *RapidMiner* proporcionou a descoberta do serviço nativo de conexão direta com a base de dados. Facilmente configurável, a conexão facilitou sobremaneira a execução de novas experiências. Assim, foi possível criar *tables* para troca de informações nos dois sentidos. Criavam-se as tabelas de dados no *SQL*, que eram lidas dentro do *RapidMiner* e também no sentido contrário. Alguns projetos no



	DtCreate	NmTable	idTb	idPrj	cmt	keyCmt
1	19/02/2019	Pini2_tbZenodoPrjByCmt	47310	1209107		
2	19/02/2019	Pini2_tbZenodoPrjByCmt	47311	1209123		

	DtCreate	NmTable	idWithoutCmt	idkey	idprjkey	idprjcmt
1	19/02/2019	Pini2_tbZenodoTempWithoutCmt	1	728	1003194	NULL
2	19/02/2019	Pini2_tbZenodoTempWithoutCmt	2	729	1003196	NULL

	DtCreate	NmTable	idprj	keycmt	ownerstr
1	20/03/2019	Pini2_tbZenodoCmtXOwner	1026485	lara	23205
2	20/03/2019	Pini2_tbZenodoCmtXOwner	1026485	lara_sa	23205

	DtCreate	NmTable	IdTable	idprj	conceptDoi	conceptId	created	doi	id	linkBadge
1	02/04/2019	Pini2_tbZenodoAlmostAll	15870	1026463	10.5281/zenodo.1026462	1026462	2017-10-23T12:42:53.799403+00:00	10.5281/zenodo.1026463	1026463	https://zenodo.org/badge/doi/10.5281/zenodo.1026
2	02/04/2019	Pini2_tbZenodoAlmostAll	15871	1026465	10.5281/zenodo.1026464	1026464	2017-10-23T11:39:29.514181+00:00	10.5281/zenodo.1026465	1026465	https://zenodo.org/badge/doi/10.5281/zenodo.1026

	DtCreate	NmTable	idprj	lg	mdKeywords	title	mdDescription
1	09/04/2019	Pini2_tbZenodoByLanguage	997	eng	NULL	Random coincidence of 2ν2β decay events as a bac...	Two-neutrino double β decay can create an irremo...
2	09/04/2019	Pini2_tbZenodoByLanguage	1162678	eng	["Partnership", "Cooperation...	Is Jordan an appropriate chance to enhance the succ...	NULL

	DtCreate	NmTable	idExtract	idprj	indexkey	NamekeyFull	Namekey	valuekey
1	22/05/2019	Pini2_tbZenodoExtractData3	293035	265451	4	metadata.keywords	-	Insecta
2	22/05/2019	Pini2_tbZenodoExtractData3	293036	265451	5	metadata.keywords	-	Lepidoptera

Figura 3.6: Tabelas *SQL* - Conversão de dados

executados no aplicativo permitiam a escrita dos resultados diretamente na base de dados com o operador *Write Database*.

Outros tipos de projetos como *Data Mining*, *Clusters*, *Decision Tree*, *Randon Tree*, *Deep Learning*, *Cross Validation*, *Data to Similiraty* foram executados.

Os resultados encontrados no *RapidMiner* não foram considerados satisfatórios. A dificuldade em construir um classificador, especialmente por não se ter uma variável etiquetada ou uma variável anotada impossibilitaram o uso dessa ferramenta. Mediante essa constatação, os esforços voltaram-se ao *SQL*.

B. Preparação do *DER - Diagrama de Entidade Relacionamento*

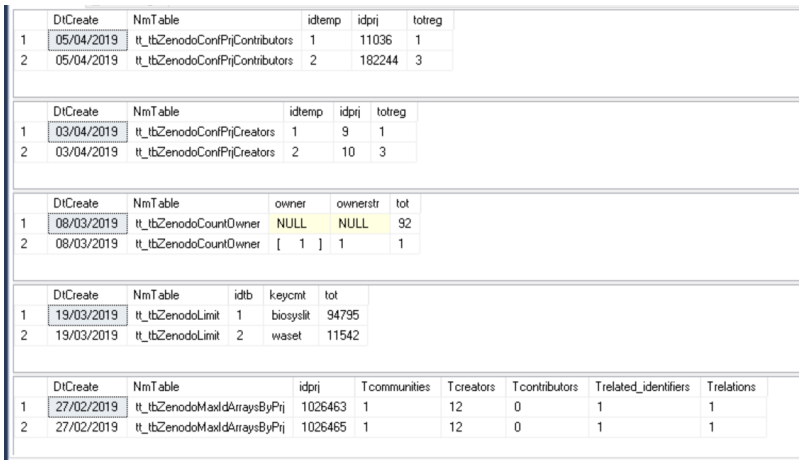
De volta ao *SQL Server* foi construído o Diagrama de Entidade Relacionamento e as tabelas do banco de dados para o sistema classificador proposto foram projetadas conforme se pode observar na figura 3.8.

Deve-se relatar o fato de várias outras tabelas terem sido construídas, especialmente para produzirem *queries* para o *RapidMiner* e também como resultado dos projetos que decorreram na plataforma. Como os resultados daquele software não estão sendo relatados, as diversas *table* foram também omitidas.

No diagrama referido, as tabelas podem ser divididas em três grupos principais. As que começam por *"tbPi"* referem-se aos dados propriamente ditos e a sílaba *"i"* é acrônimo para a palavra *"information"*. Assim, armazenam as informações em bruto, que são aquelas extraídas do repositório. Portanto são as informações do *Zenodo*.

As tabelas começadas por *"tbPs"* são aquelas que gerem as informações pertinentes às parametrizações dos sistemas de classificação formais e a sílaba *"s"* refere-se ao termo *"science"*. São portanto, responsáveis por armazenar dados relativos aos sistemas de classificação clássicos, como por exemplo, os nomes dos campos de acordo com os idiomas.

Antes da explicação de cada tabela, cabe aqui um aparte para apresentar, de acordo com a recomendação de boas práticas de programação, a utilização de mnemônicos para representar os



	DtCreate	NmTable	idtemp	idprj	totreg
1	05/04/2019	tt_tbZenodoConfPrjContributors	1	11036	1
2	05/04/2019	tt_tbZenodoConfPrjContributors	2	182244	3

	DtCreate	NmTable	idtemp	idprj	totreg
1	03/04/2019	tt_tbZenodoConfPrjCreators	1	9	1
2	03/04/2019	tt_tbZenodoConfPrjCreators	2	10	3

	DtCreate	NmTable	owner	ownerstr	tot
1	08/03/2019	tt_tbZenodoCountOwner	NULL	NULL	92
2	08/03/2019	tt_tbZenodoCountOwner	[1]	1	1

	DtCreate	NmTable	idtb	keycmt	tot
1	19/03/2019	tt_tbZenodoLimit	1	biosyslit	94795
2	19/03/2019	tt_tbZenodoLimit	2	waset	11542

	DtCreate	NmTable	idprj	Tcommunities	Tcreators	Tcontributors	Trelated_identifiers	Trelations
1	27/02/2019	tt_tbZenodoMaxIdArraysByPrj	1026463	1	12	0	1	1
2	27/02/2019	tt_tbZenodoMaxIdArraysByPrj	1026465	1	12	0	1	1

Figura 3.7: Tabelas *SQL* - Totalização de dados

nomes dos objetos do banco de dados. A tabela 3.2 lista os mnemônicos e os respetivos conceitos relacionados.

Voltando às tabelas do banco de dados, a começar pelos objetos com a assinatura *"tbPi"*, ou seja, *information*, pode-se observar na tabela 3.3 uma breve descrição de cada objeto *table*. Na figura 3.10 estão demonstradas a amostra de dois atributos de cada um dos objetos *table*.

Relativamente ao grupo *science*, *"tbPs"*, pode ser observado assim como no grupo anterior, a tabela 3.4 fazendo uma breve descrição de cada objeto *table*. A figura 3.9, exibe a amostra de dois atributos de cada um dos objetos *table*. Estes são os objetos derivados do *harvesting*.

Para interligar os dois "mundos", ou seja, o mundo real dos dados dos repositórios e o mundo real dos sistemas de classificação já concebidos na Ciência da Informação, tem-se o terceiro grupo apelidado de *"tbPr"*, significando *"run"*. Essas tabelas armazenam a parametrização do conjunto de dados que será classificado e o resultado do classificador.

Após preencher as tabelas para execução do classificador, nomeadamente as intituladas *"tbPs"* e *"tbPi"*, conforme detalhado anteriormente, o resultado do classificador é finalmente armazenado nas tabelas finais, como se vê na tabela 3.5 e na figura 3.11.

3.2.2.5 Escolha do sistema de classificação e tesauro

A classificação é um tema bastante complexo, por isso mesmo é um assunto de extrema importância na Ciência da Informação.

A nomenclatura utilizada neste trabalho pode tornar-se um pouco confusa, uma vez que a referência aos sistemas de classificação pode ter dois significados nesse contexto. O trabalho propõe um classificador, ou seja, uma técnica capaz de aplicar uma etiqueta aos objetos digitais do repositório em análise de acordo com um sistema de classificação clássico, isto é um sistema de classificação já utilizado no contexto da Ciência da Informação.

Assim, faz-se necessário estar alerta quando o trabalho faz referência a um ou outro. Preferencialmente é utilizado o termo classificador para indicar o sistema em desenvolvimento e a

Mnemônico	Conceito
Act	action
Clf	classification
Cpt	concept
Ctt	content
DO	digital object
Exc	exclude
Fct	facet
Fld	field
Gol	goal
Hvt	harvest
Lcs	license
Lct	location
Lst	list
Lvl	level
Mdd	metadata
Pbt	publication
Pfd	prefered
Pmt	parameter
Pre	pre
Rld	related
Rsc	resource
Roo	root
Slf	self
Sys	system
Vrs	version

Tabela 3.2: Mnemônicos

Tabela	Descrição
tbPiCttMdd	Conteúdo de metadados
tbPiDO	Objetos digitais
tbPiDOMdd	Metadados dos Objetos digitais
tbPiFld	Nomenclatura da faceta origem dos dados
tbPiFldRld	Campos relacionados
tbPiHvtAct	Atributos relacionados à ação de *harvesting*
tbPiHvtLct	Atributos relacionados aos locais de *harvesting*
tbPiHvtLctFld	Informações acerca do *harvesting* e respetivos sistemas de classificação
tbPiLgg	Tipos de linguagens ou idiomas possíveis
tbPiRscTpe	Tipos de Recursos

Tabela 3.3: Tabelas *"tbPi"*: nomes e descrição

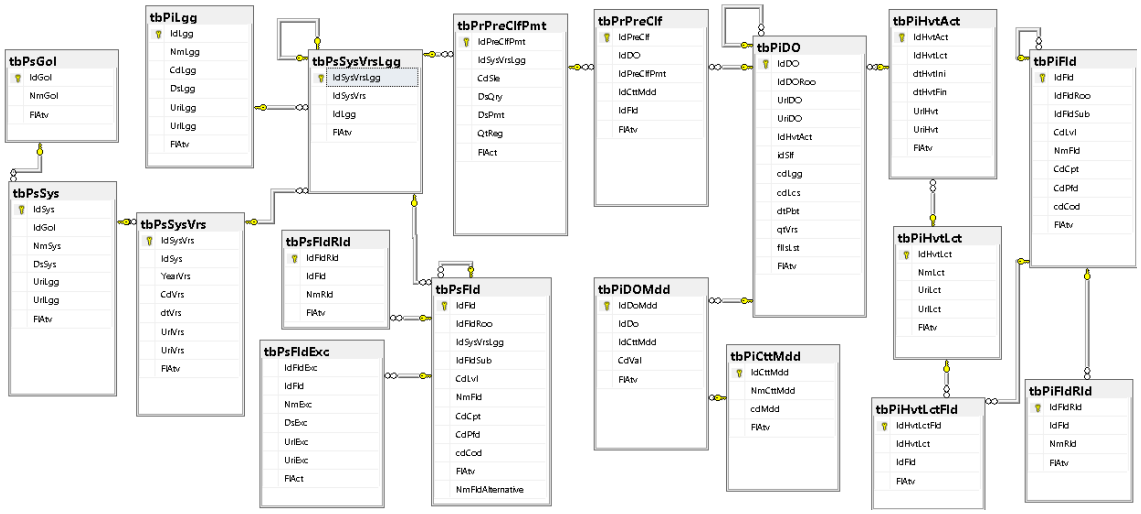


Figura 3.8: Diagrama Entidade Relacionamento

expressão sistema de classificação para abordar os sistemas de classificação tradicionais da Ciência da Informação.

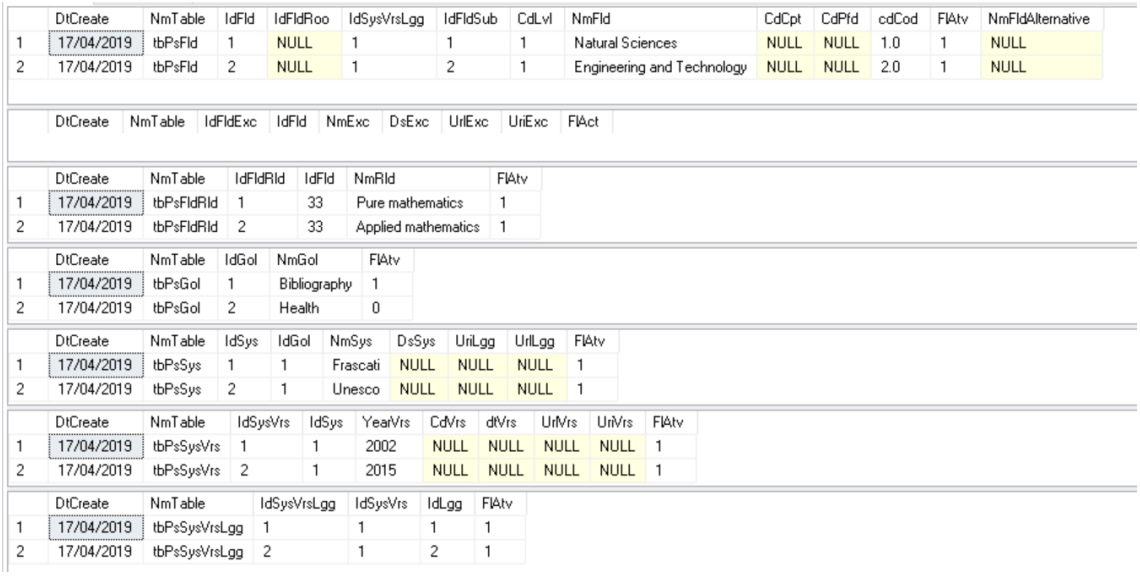
Após esse esclarecimento, e a título de exemplo da problemática da classificação bibliográfica, pode-se verificar na figura 3.12 a dificuldade relativa a esse tipo de classificação. O exemplo apresenta a comparação da classificação do mesmo livro, intitulado "Informação, uma história, uma teoria, um dilúvio"de James Gleick, mesma editora e mesma versão nas bibliotecas das duas faculdades da Universidade do Porto, FEUP e FLUP. É um exemplo interessante, pois embora pertencentes à mesma instituição as referidas bibliotecas apontaram classificações diferentes para o mesmo recurso informacional.

Do lado direito da figura está o *screenshot* da Bibioteca da FEUP indicando os códigos de classificação da *CDU* "316.77"e "316.42"e do lado esquerdo está destacado o código CDU "1"na Biblioteca FLUP.

Retomando a questão da escolha do melhor sistema de classificação a ser utilizado no classificador em desenvolvimento, foram considerados os sistemas de classificação bibliográficos e não bibliográficos.

Tabela	Descrição
tbPsFld	Nomenclatura dos sistemas de classificação
tbPsFldExc	Campos a serem excluídos, ou seja, desconsiderados durante a classificação
tbPsFldRld	Campos relacionados a cada atributo do sistema de classificação
tbPsGol	Objetivo, finalidade de cada sistema de classificação
tbPsSys	Sistema de classificação propriamente dito, como nome, *URI*, *URL*
tbPsSysVrs	Versão do referido sistema de classificação
tbPsSysVrsLgg	Linguagem de acordo com a versão do sistema de classificação

Tabela 3.4: Tabelas *"tbPs"*: nomes e descrição



	DtCreate	NmTable	IdFld	IdFldRoo	IdSysVrsLgg	IdFldSub	CdLvl	NmFld	CdCpt	CdPfd	cdCod	FlAtv	NmFldAlternative
1	17/04/2019	tbPsFld	1	NULL	1	1	1	Natural Sciences	NULL	NULL	1.0	1	NULL
2	17/04/2019	tbPsFld	2	NULL	1	2	1	Engineering and Technology	NULL	NULL	2.0	1	NULL

	DtCreate	NmTable	IdFldExc	IdFld	NmExc	DsExc	UrlExc	UriExc	FlAct

	DtCreate	NmTable	IdFldRld	IdFld	NmRld	FlAtv
1	17/04/2019	tbPsFldRld	1	33	Pure mathematics	1
2	17/04/2019	tbPsFldRld	2	33	Applied mathematics	1

	DtCreate	NmTable	IdGol	NmGol	FlAtv
1	17/04/2019	tbPsGol	1	Bibliography	1
2	17/04/2019	tbPsGol	2	Health	0

	DtCreate	NmTable	IdSys	IdGol	NmSys	DsSys	UriLgg	UrlLgg	FlAtv
1	17/04/2019	tbPsSys	1	1	Frascati	NULL	NULL	NULL	1
2	17/04/2019	tbPsSys	2	1	Unesco	NULL	NULL	NULL	1

	DtCreate	NmTable	IdSysVrs	IdSys	YearVrs	CdVrs	dtVrs	UrlVrs	UriVrs	FlAtv
1	17/04/2019	tbPsSysVrs	1	1	2002	NULL	NULL	NULL	NULL	1
2	17/04/2019	tbPsSysVrs	2	1	2015	NULL	NULL	NULL	NULL	1

	DtCreate	NmTable	IdSysVrsLgg	IdSysVrs	IdLgg	FlAtv
1	17/04/2019	tbPsSysVrsLgg	1	1	1	1
2	17/04/2019	tbPsSysVrsLgg	2	1	2	1

Figura 3.9: Tabelas *SQL* - Dados dos sistemas de classificação

A opção entretanto recaiu sobre o Sistema de Classificação *Fields of Science and Technology - FOS*. A escolha foi motivada por tratar-se de um sistema bastante simples uma vez que possui apenas dois níveis hierárquicos. Além da simplicidade, o *FOS* foi desenvolvido especialmente para o ambiente de atividades de investigação e desenvolvimento por domínio científico. Portanto, pareceu ser uma boa solução uma vez que a fonte de dados selecionada, nomeadamente os dados de investigação do *Zenodo*, são também, de natureza do domínio científico. Foram analisadas duas versões do *FOS*, nomeadamente a versão de 2002 e também a versão 2015, ambas no idioma em inglês. A figura 3.13 ilustra o comparativo entre as duas versões.

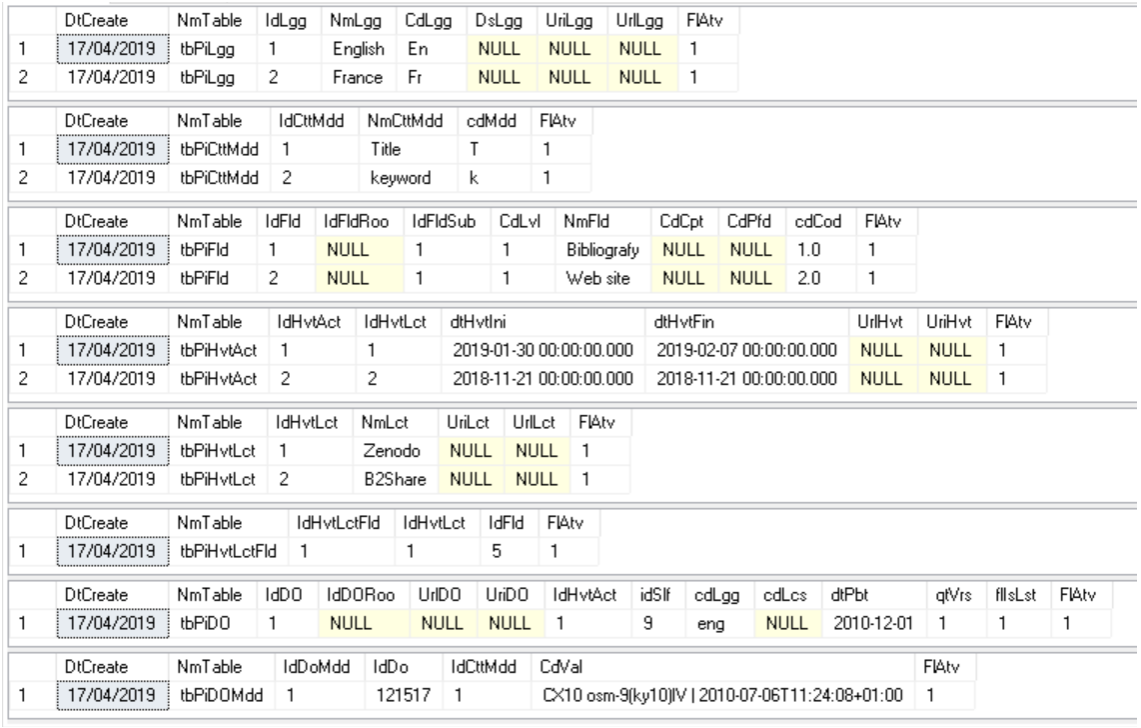
Não é objetivo desse trabalho discutir o melhor sistema de classificação de maneira ampla, mas tão somente, escolher o que melhor se adapta ao objetivo proposto. O trabalho prende-se à escolha de um sistema de classificação a ser utilizado no classificador dos objetos digitais que foram selecionados na etapa anterior.

É importante destacar que a escolha do sistema de classificação a ser utilizado não representou grande peso. Em outras palavras, não foi uma decisão estritamente fundamental tendo em vista que o trabalho foi desenvolvido de modo a ser possível experimentar vários sistemas de classificação. A estrutura de todo projeto foi concebida a permitir testes paralelos com diferentes configurações a fim de comparar e encontrar melhores resultados.

Os primeiros testes classificatórios foram executados a princípio somente com o sistema *FOS*

Tabela	Descrição
tbPrPreClf	Resultados do classificador
tbPrPreClfPmt	Parâmetros da cada amostragem

Tabela 3.5: Tabelas *"tbPr"*: nomes e descrição



	DtCreate	NmTable	IdLgg	NmLgg	CdLgg	DsLgg	UriLgg	UrlLgg	FlAtv
1	17/04/2019	tbPiLgg	1	English	En	NULL	NULL	NULL	1
2	17/04/2019	tbPiLgg	2	France	Fr	NULL	NULL	NULL	1

	DtCreate	NmTable	IdCttMdd	NmCttMdd	cdMdd	FlAtv
1	17/04/2019	tbPiCttMdd	1	Title	T	1
2	17/04/2019	tbPiCttMdd	2	keyword	k	1

	DtCreate	NmTable	IdFld	IdFldRoo	IdFldSub	CdLvl	NmFld	CdCpt	CdPfd	cdCod	FlAtv
1	17/04/2019	tbPiFld	1	NULL	1	1	Bibliografy	NULL	NULL	1.0	1
2	17/04/2019	tbPiFld	2	NULL	1	1	Web site	NULL	NULL	2.0	1

	DtCreate	NmTable	IdHvtAct	IdHvtLct	dtHvtIni	dtHvtFin	UrlHvt	UriHvt	FlAtv
1	17/04/2019	tbPiHvtAct	1	1	2019-01-30 00:00:00.000	2019-02-07 00:00:00.000	NULL	NULL	1
2	17/04/2019	tbPiHvtAct	2	2	2018-11-21 00:00:00.000	2018-11-21 00:00:00.000	NULL	NULL	1

	DtCreate	NmTable	IdHvtLct	NmLct	UriLct	UrlLct	FlAtv
1	17/04/2019	tbPiHvtLct	1	Zenodo	NULL	NULL	1
2	17/04/2019	tbPiHvtLct	2	B2Share	NULL	NULL	1

	DtCreate	NmTable	IdHvtLctFld	IdHvtLct	IdFld	FlAtv
1	17/04/2019	tbPiHvtLctFld	1	1	5	1

	DtCreate	NmTable	IdDO	IdDORoo	UrlDO	UriDO	IdHvtAct	idSlf	cdLgg	cdLcs	dtPbt	qtVrs	fllsLst	FlAtv
1	17/04/2019	tbPiDO	1	NULL	NULL	NULL	1	9	eng	NULL	2010-12-01	1	1	1

	DtCreate	NmTable	IdDoMdd	IdDo	IdCttMdd	CdVal	FlAtv
1	17/04/2019	tbPiDOMdd	1	121517	1	CX10 osm-9(ky10)IV | 2010-07-06T11:24:08+01:00	1

Figura 3.10: Tabelas *SQL* - Dados de *harvesting*

e era suposto executar somente esse tipo de experiência. Em virtude dos primeiros resultados do classificador relativamente ao sistema de classificação, conforme detalhado na seção 4.3.2, intitulada de Estatísticas, foi então referida a possibilidade de utilização de um tesauro como forma de aprimorar os resultados. Assim, foi preciso escolher o tesauro que melhor se adequaria ao projeto.

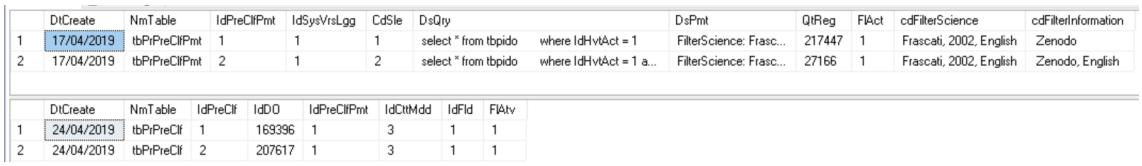
A escolha recaiu sobre o Tesauro da UNESCO[1] em virtude das seguintes razões: acesso aos dados através de descarregamento em ficheiro *XML*, mapeamento dos campos para *Dublin Core* e terminologia multidisciplinar.

Importante destacar que a utilização do tesauro, embora não tenha sido planeada na fase inicial do projeto não impactou a estrutura original do banco de dados. A arquitetura do banco foi planeada a ser bastante flexível sem contudo deixar de cumprir os requisitos fundamentais de normalização dos dados. Assim, foi possível sem nenhuma dificuldade dar um importante passo no processo. Não somente experimentar sistemas de classificação, mas fazer uso de um recurso muito importante da Ciência da Informação, representado pelos tesauros.

3.2.2.6 Portal *MyResults*

Por fim, resta ilustrar as experiências que foram executadas a fim de construir o classificador pretendido. Foram 10 experiências executadas, uma de cada vez, aplicando diferentes filtros. De acordo com o planeamento dos testes foram utilizadas diferentes combinações de sistemas de

[1] http://vocabularies.unesco.org/browser/thesaurus/en/



	DtCreate	NmTable	IdPreCltPmt	IdSysVrsLgg	CdSle	DsQry		DsPmt	QtReg	FlAct	cdFilterScience	cdFilterInformation
1	17/04/2019	tbPrPreCltPmt	1	1	1	select * from tbpido	where IdHvtAct = 1	FilterScience: Frasc...	217447	1	Frascati, 2002, English	Zenodo
2	17/04/2019	tbPrPreCltPmt	2	1	2	select * from tbpido	where IdHvtAct = 1 a...	FilterScience: Frasc...	27166	1	Frascati, 2002, English	Zenodo, English

	DtCreate	NmTable	IdPreClf	IdDO	IdPreCltPmt	IdCttMdd	IdFld	FlAtv
1	24/04/2019	tbPrPreClf	1	169396	1	3	1	1
2	24/04/2019	tbPrPreClf	2	207617	1	3	1	1

Figura 3.11: Tabelas *SQL* - Dados já classificados

classificação ou tesauro relativamente à linguagem dos dados do Repositório *Zenodo* e também da linguagem do tesauro. A imagem 3.14 detalha os testes.

A técnica utilizada foi o teste de comparação da ocorrência dos nomes de cada nível hierárquico do sistema de classificação ou tesauro relativamente aos três metadados textuais, designados *T-K-D*, referindo-se respetivamente aos metadados *Title, Description, Keyword*.

O próximo capítulo explica os objetivos, a arquitetura e as funcionalidades do Portal.

Autor	**Gleick**, James
Título	Informação : uma história, uma teoria, um dilúvio / James **Gleick**
Língua	por
Local	Lisboa
Editor	Temas & Debates
Ano	2012
Descrição	591 p. ; 24 cm
Assunto	Tecnologias da informação
	Informação
CDU	001
ISBN	978-989-644-172-2
Objeto Digital	Ver capa
Cota	001G468i
Existências...	FLUP
Link partilhável	http://catalogo.up.pt:80/F/?func=direct&doc_number=000805668&local_base=FLUP

Autor	**Gleick**, James
Título	Informação : uma história, uma teoria, um dilúvio / James **Gleick** ; Tradução Artur Lopes Cardoso
Língua	por
Local	Lisboa
Editor	Temas e Debates
Ano	2012
Descrição	591 p. : il. ; 24 cm
Assunto	Sociedade da informação
	Ciência da informação
CDU	316.77
	316.42
Ent.Adic.	Cardoso, Artur Lopes
ISBN	978-644-172-2
Cota	316.77 /GLEj/INF
Existências...	FEUP
Link partilhável	http://catalogo.up.pt:80/F/?func=direct&doc_number=000561408&local_base=FEUP

Figura 3.12: Classificação de livro: FEUP x FLUP

COMPARISON OF THE REVISED FOS CLASSIFICATION WITH THAT IN FM 2002

	FOS in FM 2002	Revised FOS
1. Natural Sciences	1.1 Mathematics and computer sciences 1.2 Physical sciences 1.3 Chemical sciences 1.4 Earth and related environmental sciences 1.5 Biological sciences	1.1 Mathematics 1.2 Computer and information sciences 1.3 Physical sciences 1.4 Chemical sciences 1.5 Earth and related environmental sciences 1.6 Biological sciences 1.7 Other natural sciences
2. Engineering and Technology	2.1 Civil engineering 2.2 Electrical engineering, electronics 2.3 Other engineering sciences	2.1 Civil engineering 2.2 Electrical engineering, electronic engineering, information engineering 2.3 Mechanical engineering 2.4 Chemical engineering 2.5 Materials engineering 2.6 Medical engineering 2.7 Environmental engineering 2.8 Environmental biotechnology 2.9 Industrial Biotechnology 2.10 Nano-technology 2.11 Other engineering and technologies
3. Medical and Health Sciences	3.1 Basic medicine 3.2 Clinical medicine 3.3 Health sciences	3.1 Basic medicine 3.2 Clinical medicine 3.3 Health sciences 3.4 Health biotechnology 3.5 Other medical sciences
4. Agricultural Sciences	4.1 Agriculture, forestry, fisheries and allied sciences 4.2 Veterinary medicine	4.1 Agriculture, forestry, and fisheries 4.2 Animal and dairy science 4.3 Veterinary science 4.4 Agricultural biotechnology 4.5 Other agricultural sciences
5. Social Sciences	5.1 Psychology 5.2 Economics 5.3 Educational sciences 5.4 Other social sciences	5.1 Psychology 5.2 Economics and business 5.3 Educational sciences 5.3 Sociology 5.5 Law 5.6 Political Science 5.7 Social and economic geography 5.8 Media and communications 5.7 Other social sciences
6. Humanities	6.1 History 6.2 Languages and literature 6.3 Other humanities	6.1 History and archaeology 6.2 Languages and literature 6.3 Philosophy, ethics and religion 6.4 Art (arts, history of arts, performing arts, music) 6.5 Other humanities

Figura 3.13: Comparativo *FOS*: 2002 e 20015

EXPERIÊNCIA	SISTEMA DE CLASSIFICAÇÃO	AMOSTRA
1	Frascati, 2002, English	Zenodo
2	Frascati, 2002, English	Zenodo, English
3	Frascati, 2015, English	Zenodo
4	Frascati, 2015, English	Zenodo, English
5	Unesco, 2016, English, 2.15 - Mathematics and statistics	Zenodo
6	Unesco, 2016, English, 2.15 - Mathematics and statistics	Zenodo, English
7	Unesco, 2016, France, 2.15 - Mathematics and statistics	Zenodo
8	Unesco, 2016, France, 2.15 - Mathematics and statistics	Zenodo, France
9	Unesco, 2016, Spanish, 2.15 - Mathematics and statistics	Zenodo
10	Unesco, 2016, Spanish, 2.15 - Mathematics and statistics	Zenodo, Spanish

Figura 3.14: Experiências do classificador

Capítulo 4

Portal *Web* de classificação

O capítulo 3 apresentou o projeto do sistema *on line* para classificação de documentos, fases de construção e estrutura ao nível do banco de dados. Nessa secção será apresentado o portal *web* relativo a esse sistema. Encontra-se ainda na fase embrionária mediante o tempo disponível para sua construção.

4.1 Objetivos

O desenvolvimento do portal surgiu da necessidade de apresentar o resultado de todo trabalho realizado. Embora não tenha sido planeado inicialmente, a sua construção foi necessária. É a concretização e consolidação do objetivo inicial e representa o amplo resultado atingido com o árduo trabalho decorrido.

Assim, o portal possibilitou a acomodação das ideias trabalhadas e na figura 4.1 estão reunidas as respostas que se pretende oferecer, os pilares do trabalho desenvolvido, as fases do classificador e as propostas de inovação.

Em poucas palavras, busca-se com o projeto a recuperação da informação com vistas à transparência e privacidade, mas sobretudo à democracia da informação e principalmente à literacia.

Espera-se que o classificador possa ajudar a resolver problemas de recuperação da informação como por exemplo, problemas linguísticos, de semântica, de exaustividade versus especificidade, além de ruídos x silêncio.

O classificador em desenvolvimento procura reunir duas ciências que se complementam: Ciência da Informação e Ciência da Computação. Relativamente à primeira foram selecionados os seguintes recursos: sistemas de classificação e tesauros, utilização de metadados e propostas da iniciativa *DCMI*. Entre os recursos tecnológicos foram destacados o *harvesting*, banco de dados estruturados, ferramentas *web* como *web services* e *web sites*.

O site e o classificador foram estruturados mediante três fases: 1—Classificador, 2—Validador e 3—Pesquisa.

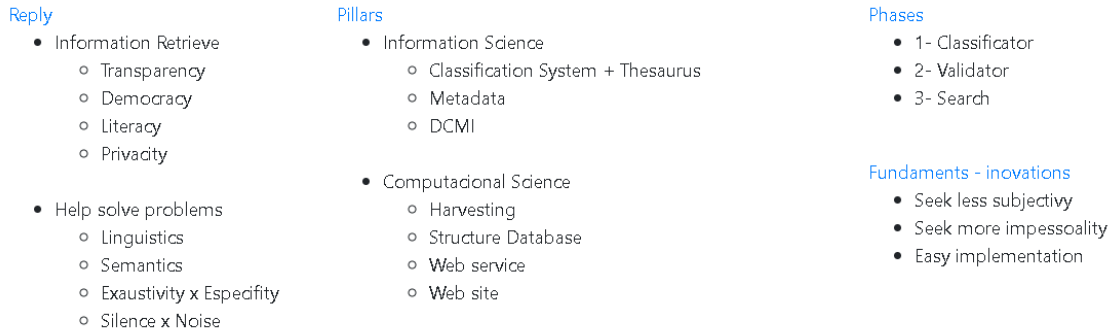


Figura 4.1: Objetivos do portal

Como componentes de inovação espera-se que seja possível, simultaneamente: diminuir a subjetividade no processo de classificação de documentos e consequentemente, aumentar a impessoalidade. Tudo isso através de uma implementação descomplicada.

4.2 Arquitetura

A figura 4.2 demonstra a página inicial do portal. A imagem foi demarcada em três áreas destacadas em vermelho, a simbolizarem o cabeçalho, corpo e rodapé. No cabeçalho estão representados respetivamente, o logótipo do projeto chamado "myResults", supostas informações acerca do projeto e uma secção dedicada a *Newsletter*.

A parte do meio, aqui denominada corpo, expõe a estrutura fundamental do portal e está dividida em três secções, nomeadas *"1 - Classificator"*, *"2 - Validator"* e *"3 - Search"* que neste contexto serão denominadas "1 - Classificador", "2 - Validador"e "3 - Pesquisa".

O rodapé, como é de costume em portais, foi projetado para fazer ligação com grupos de interesse, discorrer sobre o projeto e acessar informação pertinente ao assunto da plataforma.

Portanto, a arquitetura proposta é similar às fases planeadas para o projeto de classificação dos objetos digitais. Cumprem assim, a tarefa de explicar a classificação dos objetos, a validação da classificação e apresentação do resultado, referida pela nomenclatura "Pesquisa".

O Classificador representa a primeira fase e consiste no processo de classificação dos objetos digitais. São necessários três recursos: os objetos que serão classificados, o sistema de classificação e o tesauro previamente escolhidos.

Na segunda etapa tem-se o validador que em outras palavras é uma ferramenta de *crowdsourcing*. Ele realiza o processo de validação de forma clara e explícita. Por fim, a terceira etapa é a ferramenta de busca dos objetos que foram classificados.

4.2.1 Classificador

A etapa de classificação, conforme explicado, fundamenta-se no processo de atribuição de uma ou mais etiquetas classificadoras para cada objeto digital. Os detalhes técnicos foram esclarecidos

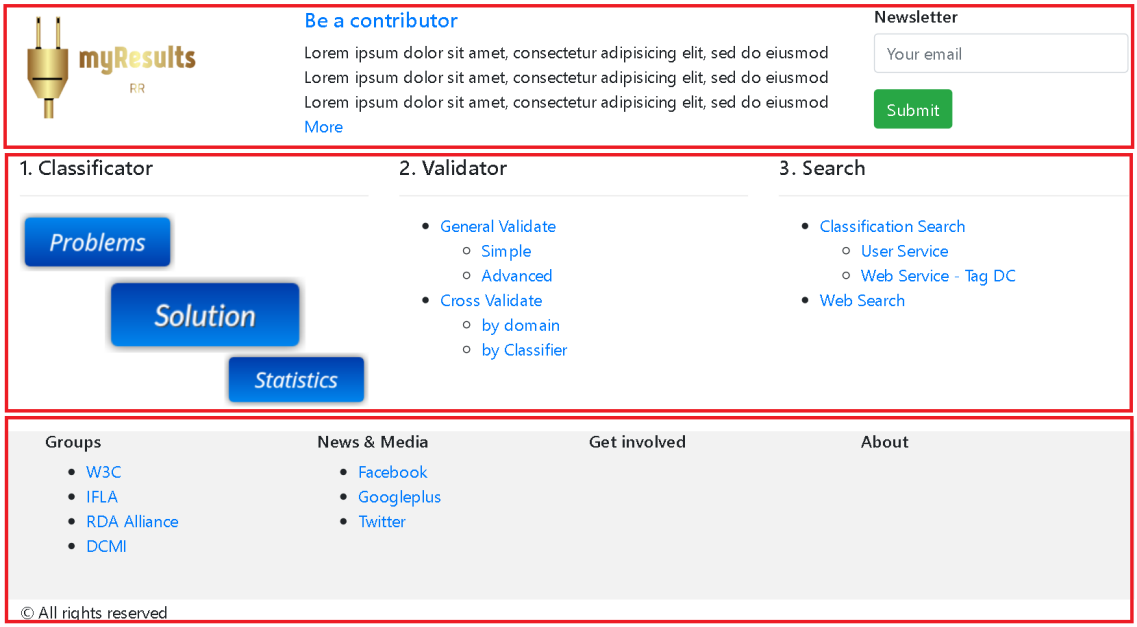


Figura 4.2: Página inicial

no capítulo anterior.

Os botões identificados na figura que reporta a página inicial do portal, chamados de “Problemas”, “Soluções” e “Estatísticas” visam oferecer a contextualização do classificador. Na parte dedicada aos problemas, foram reunidos alguns *screenshots* de sítios *web* que demonstram "dificuldades"durante a pesquisa de recursos informacionais como por exemplo na Universidade do Porto, no *Google*, na *GloboPlay* e no próprio repositório *Zenodo*.

O botão dedicado às soluções apresenta breve estudo de uma possível faceta de classificação relativa à origem dos dados, descrita em seguida. O conteúdo relativo ao botão destinado às estatísticas também está explicado em uma sessão exclusiva no decorrer do capítulo.

A classificação dos objetos digitais, nomeadamente do Repositório *Zenodo* foi construída sob a perspetiva de um sistema de classificação a fim de ser atribuído uma (ou mais do que uma) etiqueta relativa ao domínio científico representado pelo sistema de classificação. Nesse sentido, pode-se falar na classificação relativa à faceta "domínio".

Paralelamente à construção dessa faceta foram identificadas durante a análise dos dados do repositório outras possíveis maneiras de classificar os objetos digitais. A figura 4.3 ilustra uma possível maneira de classificar esses objetos a partir da origem dos dados. Na primeira parte da figura estão os objetos antes de qualquer classificação ou de outra maneira, sem uma etiqueta. Na segunda parte da figura tem-se atribuído aos objetos, uma cor que o identifica de acordo com a terceira parte da figura.

Para construção dessa faceta foram propostos quatro grandes ramos, que representariam a origem dos objetos. Assim, poderiam ser do tipo bibliográficos, de *web sites*, *wikis* e *database*. No caso de serem do tipo bibliográficos poderiam à partida, serem do tipo repositórios, de museus,

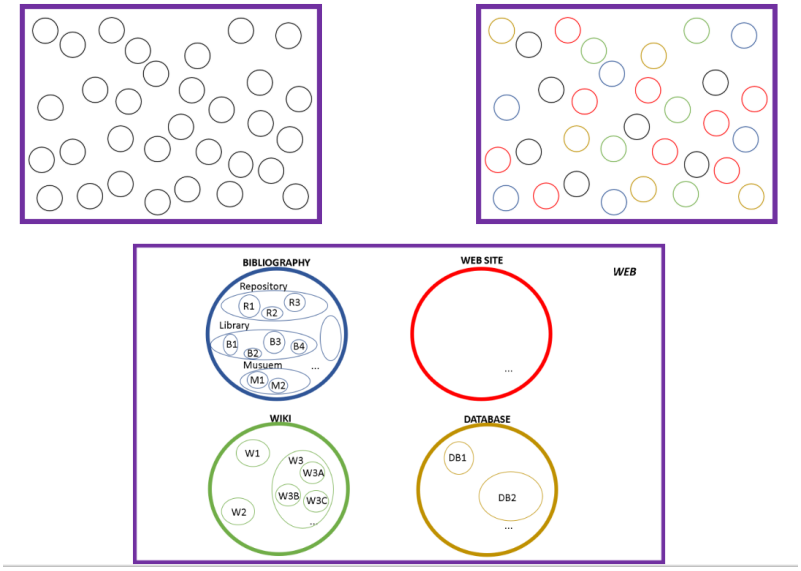


Figura 4.3: Classificação a partir da origem dos dados

de bibliotecas.

Esta faceta foi superficialmente analisada e requer estudo. O mais importante contudo é verificar a possibilidade de identificação dos objetos digitais a partir de sua proveniência e o estudo restringiu-se somente a delinear essa possibilidade.

4.2.2 Validador - *Crowdsourcing*

A partir da etiqueta ou etiquetas previamente atribuídas a cada objeto realizada na primeira etapa, faz-se necessário a validação da classificação em decurso. O projeto idealizou a validação por utentes na forma de *crowdsourcing*.

O empreendimento, ou seja, a validação da classificação proposta tem as características pertinentes a esse tipo de atividade. Pretende-se através de participações autenticadas e voltadas a um público que lida com a Ciência da Informação, especialmente a pessoas envolvidas no processo de indexação e catalogação, unir esforços e produzir resultados altamente coerentes e precisos. Em última análise pretende-se trabalhar colaborativamente.

O exemplo das diferenças nas classificações do mesmo livro nas faculdades FEUP e FLUP demonstram o problema da subjetividade no processo de indexação. Conforme referido anteriormente, espera-se que alguns problemas da indexação podem ser minimizados. A validação do classificador busca atingir uma maior interação entre os profissionais que lidam com a indexação e na medida do possível, algum consenso nessa questão.

O formato proposto nestas validações, é apresentar para cada objeto, as classificações aferidas e dois botões nos quais o utente que está a validar pode simplesmente fazer o clique no botão a confirmar ou rejeitar a classificação atribuída. Caso seja carregado o botão verde, indicando a confirmação, o texto referente passa a ser exibido na cor verde. Por outro lado, se for escolhido

o botão vermelho a indicar a rejeição, o texto automaticamente passa a ser apresentado na cor vermelha.

Foram desenvolvidas dois protótipos para o processo de validação conforme se verifica nas próximas figuras. Na primeira figura 4.4, está exemplificada a "Validação Geral", chamada de "Simples", pois foi projetada para três filtros: sistema de classificação, repositório e uma opção para objetos similares.

Para o campo "Sistema de classificação"são recuperados de maneira dinâmica, todos os sistemas de classificação armazenados no banco de dados. Relativamente ao repositório está restrito ao *Zenodo*, pois é a única fonte de dados do sistema. O filtro pertinente à similaridade dos objetos oferece a possibilidade de buscar somente a última e mais atualizada versão de cada objeto. É facultado ao utente do *Zenodo* publicar todas as atualizações de seu projeto de investigação. Esse filtro procura facilitar o trabalho de validação ao desconsiderar possíveis duplicados.

A figura 4.4 demonstra a validação chamada "Validação Cruzada". Assim como a validação simples, oferece a opção para a escolha do sistema de classificação e para o repositório. Como tem o propósito de fazer a validação a partir de objetos que tenham recebido múltiplas classificações, oportuniza a escolha do número de classificações por objeto, ou seja, que tenham sido identificados como pertencentes a mais de um domínio científico. Esta funcionalidade representa o carácter multidisciplinar do classificador.

Convém esclarecer, conforme se pode deduzir ao observar a figura que ilustra a página inicial do portal, que embora tenham sido planeadas quatro tipos de validações foram apresentadas apenas duas. Relativamente à "Validação Geral"foi demonstrada a opção "Simples"e quanto à "Validação Cruzada"foi disponibilizada a opção "Por Domínio". A "Validação Geral"do tipo "Avançada"foi planeada a possibilitar filtros que possam trazer mais familiaridade ao validador. A "Validação Cruzada por Classificador"poderá estar disponível quando forem aplicados outros tipos de sistemas de classificação aos objetos digitais.

4.2.3 Pesquisa

Assim como na etapa de validação, a implementação da funcionalidade "Pesquisa"foi realizada parcialmente. Foram projetados dois serviços de "Apresentação dos resultados do classificador". O primeiro tem como proposta apresentar a etiqueta com a classificação propriamente dita, ou seja, apresentar ao solicitante uma resposta de acordo com a consulta realizada. O segundo serviço refere-se à construção de um buscador, isto é, uma ferramenta de pesquisa na *web*.

A primeira função faz parte da resposta ao problema formulado na dissertação que se prende à definição de uma classificação de modo a configurar os metadados específicos na plataforma Dendro. Em virtude da flexibilidade impressa à solução foram projetadas duas formas de fornecer a resposta.

A apresentação do resultado pode ocorrer através de um serviço destinado ao utente via consulta ao *site* ou através de um serviço a ser executado diretamente por máquinas.

A pesquisa a ser realizada por qualquer utente ou instituição foi projetada a partir dos requisitos mínimos necessários ao classificador. São necessários os campos textuais como título, *keywords* e

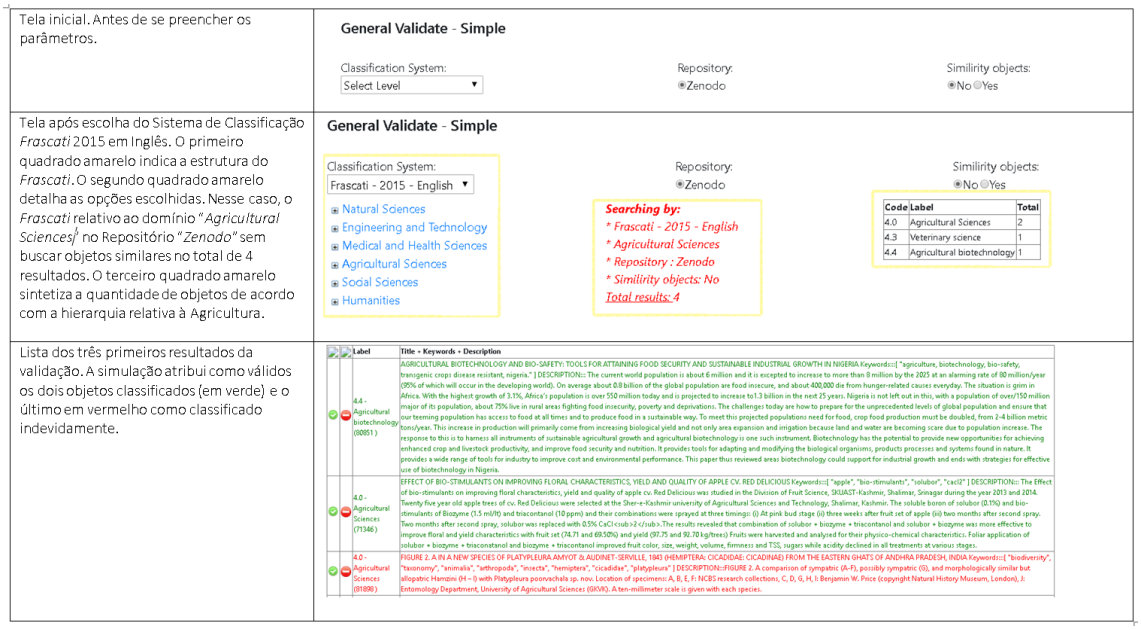


Figura 4.4: Validação geral - simples

descrição e o respetivo sistema de classificação que já tenha sido implementado no portal. Dessa forma é possível oferecer ao solicitante do serviço a etiqueta de classificação pertinente ao objeto que se quer classificar.

Pode-se ainda pré-configurar o classificador, ou seja, através de uma parametrização prévia armazenar as variáveis necessárias ao serviço. Pode-se por exemplo atribuir os sistemas de classificação que cada repositório utiliza.

De modo a atender um dos objetivos desse estudo, que é a indicação de metadados específicos dentro da plataforma Dendro, o resultado estaria disponível através de serviços *web* no formato *Dublin Core* através do descritor <*dc:subject*>. A proposta é exequível de forma a oferecer interoperabilidade uma vez que o serviço seria realizado por meio de um *Web Service* com a *tag* no formato *Dublin Core*. A interoperabilidade pode ser maximizada utilizando-se os esquemas de codificação disponibilizados pela iniciativa *DCMI*, ou seja, utilizando os vocabulários controlados preparados para as classificações bibliográficas.

O segundo serviço, voltado para a construção de um buscador como ferramenta de pesquisa é destinado ao público em geral. A figura 4.6 exemplifica o que foi feito para pesquisas simples ou gerais. De acordo com as etiquetas classificadoras dos objetos já classificados pode oferecer uma nova forma de recuperar informação objetivando a democracia e a literacia da informação.

A figura foi delimitada por marcas na cor amarelo a fim de explicar as funcionalidades da página. Pode-se averiguar no lado esquerdo, duas alternativas que o utente pode escolher. No momento, estão limitadas (e também não implementadas) a utilizar o histórico e restringir a busca à origem do objeto. Outros tipos de filtros, como aqueles apresentados na secção de validação poderiam potenciar ainda mais as possibilidades de resultado, como por exemplo, autor, afiliação,

Figura 4.5: Validação cruzada - por domínio

tipo de recurso, comunidades e outros.

No topo e no centro há a caixa para inserção da palavra ou expressão que se deseja pesquisar, que na demonstração está preenchida com a palavra "*health*". Após o clique no botão "*Execute*"são apresentados os resultados categorizados pelas classes do sistema de classificação utilizado. Relativamente ao caso em questão foram encontrados objetos em quatro Domínios conforme indicam as setas amarelas.

Nesse ponto encontra-se uma alternativa ao paradigma dos buscadores tradicionais, em que se utiliza todo o histórico do utilizador de modo a "adivinhar"o que o utente deseja buscar. Aqui, os resultados são agrupados de acordo com o domínio a que pertencem. Isso permite uma escolha de acordo com o interesse do utente permitindo-lhe verificar de uma forma abrangente as possibilidades para sua consulta.

Essa nova maneira de apresentar os resultados de pesquisas parece ser uma alternativa à limitação das famosas "bolhas de filtro". Parece ainda, ser um novo caminho de modo a proporcionar maior literacia aos utentes de plataformas de busca de objetos digitais. A capacidade de oferecer mais liberdade a quem está a realizar uma pesquisa pode implicar resultados menos limitadores.

A utilização do "Histórico anterior"passa a figurar como uma opção não somente configurável mas também bastante explícita ao utente.

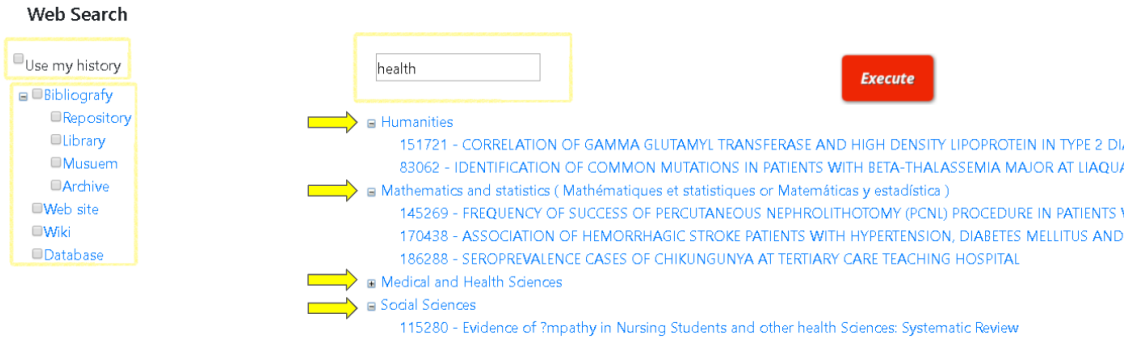


Figura 4.6: Pesquisa web

4.3 Funcionalidades

Entre as funcionalidades presentes nos serviços do portal serão destacadas a seguir, duas delas: a auto validação e os vários resultados enquadrados como estatísticas.

4.3.1 Auto validação

A classificação é importante mas é fundamental existir uma constante e aprimorada avaliação do processo classificatório. Assim, a figura 4.7 procura resumir o funcionamento do classificador e o processo de interação entre as fases projetadas.

A validação ocorre em três sentidos. Explicitamente, entre o classificador e o processo de validação através da ferramenta projetada para este fim, ou seja, através do processo *crowdsourcing*. A validação ocorre também na fase de pesquisa. Aqui poderá ser realizada implicitamente verificando o comportamento do utente por meio de uma permissão assentida ou mesmo explicitamente através de seu *feedback*. Por fim, há ainda lugar para a auto validação do classificador.

A título de exemplo, pode-se referir a classificação que acontece em dois níveis hierarquicamente diferentes e intrinsecamente relacionados através de uma dependência. Um objeto pode ser classificado com a etiqueta "6.0 - Humanities"e "6.1 - History". Como a segunda classificação é derivada da primeira pode-se desconsiderar a classificação relativa ao nível imediatamente superior, ou seja, não é necessário classificar o objeto com a etiqueta "6.0".

Em outras palavras, o classificador fornece subsídios ao validador e à pesquisa e processa o resultado de forma a depurar o processo de classificação.

4.3.2 Estatísticas

À medida que o classificador foi sendo desenvolvido os resultados eram analisados e comparados. Assim, seguem as estatísticas decorrentes dos testes realizados.

Foram estabelecidos quatro tipos de estatísticas conforme se vê na figura 4.8. Os três primeiros estão descritos a seguir, designados por *By domain*, *By languague*, *By join*. A última opção, *By metadata* não foi desenvolvida em sua totalidade.

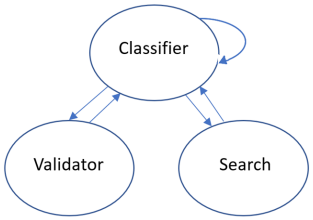


Figura 4.7: Auto validação

Cumpre destacar que as estatísticas são dinâmicas sendo possível escolher parâmetros como por exemplo, o sistema de classificação e a exibição do resultado de forma sintética ou analítica através da opção *Detail*.

Prosseguindo com as demonstrações das estatísticas, a ilustração 4.9 apresenta o resultado do classificador ao utilizar o sistema *Frascati* nas versões 2002 e 2015 de modo sintético. A utilização da nomenclatura *Frascati* acaba por representar uma figura de Linguagem. O verdadeiro nome do sistema é *FOS* conforme referido no capítulo 2 deste trabalho, pois *Frascati* é o nome do Manual que apresenta o sistema de classificação. Portanto, ao referenciar o nome *Frascati*, o portal está referindo-se ao *FOS*.

Este primeiro resultado mostrou de forma inequívoca que o simples uso de um sistema de classificação não é suficiente para um classificador eficiente. Pode-se observar nas duas últimas linhas destacadas que a quantidade de objetos classificados como *Social Sciences* atinge o índice de 0.270411% na versão 2002 e 1.45691% na versão 2015. Ao mesmo tempo, os objetos classificados como *Humanities* cai de 1.68547% na versão 2002 para 0.121409% na versão 2015. Ou seja, na versão 2002 classifica mais objetos na área de *Humanities* enquanto a versão 2015 classifica os objetos na sua maioria como *Social Sciences*.

Embora essa primeira estatística não traga um resultado com boas perspetivas futuras, a figura 4.10 é apresentada para destacar de modo analítico a atribuição das classes nas duas versões do *Frascati*.

Importantíssimo nesse ponto, é clarificar a quantificação dos objetos classificados de modo sintético e analítico. O total de objetos sumarizado em detalhes é maior que o número de objetos somados sem detalhes, porque aquando da totalização detalhada, um mesmo objeto pode receber duas etiquetas simultâneas dentro do mesmo nível. Para exemplificar, pode ser citada a diferença no domínio intitulado "Medical and Health Sciences". Na figura 4.9 são referidos 26 objetos nessa categoria e na figura 4.10, se forem somados os objetos que fazem parte deste domínio, o número aumenta para 30 objetos. Há quatro objetos que receberam duas classificações simultâneas nesse domínio. Os números que os identificam são 83062, 145269, 151721 e 151721.

Assim como esses primeiros resultados estatísticos puderam evidenciar de forma bem clara a efetividade acerca do uso do classificador, os próximos exemplos também conseguem explicitar a importância do descritor linguagem, <dc:language>, na atribuição de metadados.

Assim, a imagem 4.11 sintetiza a aplicação do filtro linguagem durante o processo de classificação. Em ambas versões do *Frascati*, 2002 e 2015, quando são selecionados só objetos de

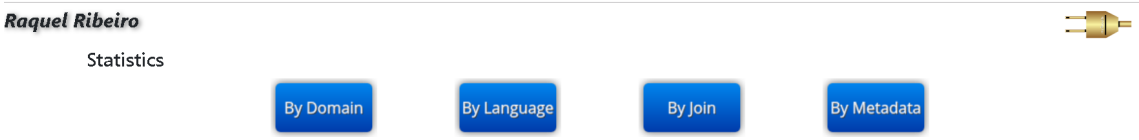


Figura 4.8: Estatísticas

determinado idioma ao se fazer a classificação, a quantidade de objetos classificados aumenta. Relativamente ao *Frascati* 2002, a proporção de objetos classificados aumenta de 1.52911% para 2.6062%. Em relação ao *Frascati* 2015, o aumento é ainda maior, partindo de 1.20535% para 3.08106%.

Na próxima figura, 4.12 semelhante à anterior, ou seja, a exemplificar a utilização do filtro linguagem, de forma detalhada acerca do Sistema *Frascati*, pode-se observar nas linhas em verde o resultado da proporção de objetos classificados sem o *match* de linguagem. A título de curiosidade, são apresentadas as diversas linguagens dos objetos em questão. Ao se verificar as linhas destacadas em alaranjado pode-se comparar que o resultado também aumenta. Passa de 1.03519% para 2.6062% quando se aplica o *Frascati* 2002 em relação ao *match* da linguagem. Relativamente ao *Frascati* 2015, o aumento é ainda maior, na mesma situação, passa de 0.687064% para 3.08106%.

Ainda em relação à utilização da linguagem, ao se verificar a quantidade de objetos classificados aquando da utilização do Tesauro *Unesco*, o aumento no número de objetos classificados aumenta em relação à linguagem *English*, que passa de 1.59257% para 8.56217% e também no idioma *Spanish*, mudando de 0.656252% para 1.09091%. Fato este, exemplificado na figura 4.13.

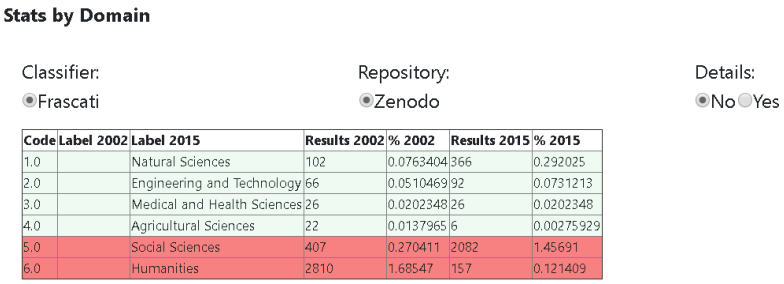
Entretanto, curiosamente, no idioma francês, ocorre o decréscimo na quantidade de objetos classificados, mudando de 0.00735811% para os "inacreditáveis", zero %. A análise superficial do fenómeno leva a crer que a atribuição do descritor linguagem dos objetos foi provavelmente realizada de forma incorreta.

A última imagem relativa à questão linguística, demonstrada na figura 4.14 expõe a estatística anterior de forma detalhada sendo possível notar os mesmos valores além da exibição das demais linguagens encontradas nos respetivos objetos digitais.

A fim de concluir as estatísticas apuradas, são apresentadas a seguir as evidências talvez mais interessantes, quando foi aplicado o Tesauro da *Unesco* relativamente ao Domínio, *2 - Science* e aos conceitos do grupo *2.15 - Mathematics and statistics*. Foram contabilizadas 102 entradas equivalentes a 34 conceitos representados em três idiomas: Inglês, Espanhol e Francês.

Essas estatísticas foram denominadas *join* pois a ideia é utilizar os termos dos tesauros em combinação com os sistemas de classificação de forma a atingir um resultado melhor no classificador em estudo.

A figura 4.15 tem apenas três linhas mas tem grande significado. A primeira linha totaliza a quantidade de objetos classificados ao ser utilizado o sistema *Frascati* 2002, que na ocasião não



Code	Label 2002	Label 2015	Results 2002	% 2002	Results 2015	% 2015
1.0		Natural Sciences	102	0.0763404	366	0.292025
2.0		Engineering and Technology	66	0.0510469	92	0.0731213
3.0		Medical and Health Sciences	26	0.0202348	26	0.0202348
4.0		Agricultural Sciences	22	0.0137965	6	0.00275929
5.0		Social Sciences	407	0.270411	2082	1.45691
6.0		Humanities	2810	1.68547	157	0.121409

Figura 4.9: Estatísticas por domínio

conseguiu etiquetar nenhum objeto. A segunda linha totaliza os objetos classificados tendo como base o *Frascati* 2015, que nesse caso, totalizou 268 objetos.

Enquanto no *Frascati* 2002 a referência à classe hierarquicamente superior, denominada *Natural Sciences* era composta pelos termos *Mathematics and Computer Sciences*, na versão revisada de 2015, os termos foram desmembrados em dois outros: um campo usado estritamente para *Mathematics* e outro para *Computer and information sciences*. Essa simples separação possibilitou a classificação dos 268 objetos referidos.

O ganho mais significativo entretanto, é percebido quando são utilizados os conceitos do tesauro. A experiência demonstra nessa situação, que a utilização dos 34 conceitos relatados, como por exemplo, *Factor analysis* em Inglês ou *Análisis factorial* em Espanhol ou *Analyse factorielle* em Francês impulsionou a descoberta de novos objetos, passando de 268 a 7969 objetos classificados. O número sem análise crítica demonstra um acréscimo de 29.7351%.

É claro que o estudo precisa amadurecer e muitas variáveis necessitam de análise. A derradeira figura, com nomenclatura 4.15 apresenta em detalhes, a quantidade objetos classificados com as *labels* dos conceitos do tesauro e a percentagem de objetos classificados em relação aos primeiros 268 objetos etiquetados anteriormente.

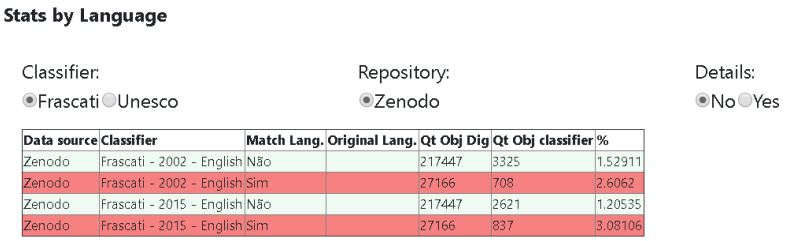
Foram apresentadas até aqui, as estatísticas relativamente ao domínio dos sistemas de classificação, ou seja, das classes que compõe cada sistema, sob a perspetiva da utilização do descritor que identifica a linguagem dos objetos classificados e da utilização de tesauros para complementar o classificador. A última opção que analisaria as classificações sob a óptica da atribuição de pesos de cada metadado utilizado como *input* do classificador não foi concluída a tempo de ser apresentada. É de suma importância realizar essa análise para verificar se a classificação foi atribuída a partir do *Title*, *Description* ou *Keyword* e mesmo das possíveis combinações entre eles. Esse estudo permitiria atribuir um *score* a cada objeto classificado relativamente à incidência do metadado que originou a classificação.

Complementarmente à análise dos *scores*, começou a ser desenvolvido um mecanismo de termos excludentes durante o processo classificatório de forma a assegurar a remoção de possíveis classificações erróneas. Em termos gerais, o classificador precisa ser balanceado. Se por um lado são incluídas opções de modo a proporcionar maior abrangência de cobertura para inclusão dos objetos em determinada *label*, o processo inverso também precisa ser executado de modo a

Code	Label 2002	Label 2015	Results 2002	% 2002	Results 2015	% 2015
1.0	Natural Sciences	Natural Sciences	76	0.0611643	76	0.0611643
1.1	Mathematics and computer sciences	Mathematics	0	0	268	0.215685
1.2	Physical sciences	Computer and information sciences	11	0.00643835	0	0
1.3	Chemical sciences	Physical sciences	5	0.00275929	11	0.00643835
1.4	Earth and related environmental sciences	Chemical sciences	0	0	5	0.00275929
1.5	Biological sciences	Earth and related environmental sciences	12	0.00597847	0	0
1.6		Biological sciences	0	0	12	0.00597847
1.7		Other natural sciences	0	0	0	0
2.0	Engineering and Technology	Engineering and Technology	6	0.00367906	6	0.00367906
2.1	Civil engineering	Civil engineering	60	0.0473679	60	0.0473679
2.10		Nano-technology	0	0	2	0.00183953
2.11		Other engineering and technologies	0	0	0	0
2.2		Electrical engineering, electronic engineering, information engineering	0	0	0	0
2.3	Other engineering sciences	Mechanical engineering	0	0	17	0.0124168
2.4		Chemical engineering	0	0	7	0.00321918
2.5		Materials engineering	0	0	0	0
2.6		Medical engineering	0	0	3	0.00183953
2.7		Environmental engineering	0	0	4	0.00275929
2.8		Environmental biotechnology	0	0	0	0
2.9		Industrial Biotechnology	0	0	0	0
3.0	Medical and Health Sciences	Medical and Health Sciences	4	0.00321918	4	0.00321918
3.1		Basic medicine	0	0	0	0
3.2	Clinical medicine	Clinical medicine	3	0.00183953	3	0.00183953
3.3	Health sciences	Health sciences	23	0.0151761	23	0.0151761
3.4		Medical biotechnology	0	0	0	0
3.5		Other medical sciences	0	0	0	0
4.0	Agricultural Sciences	Agricultural Sciences	3	0.00137965	3	0.00137965
4.1		Agriculture, forestry, and fisheries	0	0	0	0
4.2	Veterinary medicine	Animal and dairy science	19	0.0124168	0	0
4.3		Veterinary science	0	0	1	0.000459882
4.4		Agricultural biotechnology	0	0	2	0.000919764
4.5		Other agricultural sciences	0	0	0	0
5.0	Social Sciences	Social Sciences	139	0.085998	139	0.085998
5.1	Psychology	Psychology and cognitive sciences	131	0.0910567	0	0
5.2	Economics	Economics and business	143	0.0910567	2	0.00137965
5.3	Educational sciences	Educational	3	0.00183953	1250	0.874236
5.4	Other social sciences	Sociology	1	0.000459882	29	0.0211546
5.5		Law	0	0	652	0.450685
5.6		Political Science	0	0	43	0.023454
5.7		Social and economic geography	0	0	0	0
5.8		Media and communications	0	0	0	0
6.0	Humanities	Humanities	153	0.119569	153	0.119569
6.1	History	History and archaeology	2664	1.56498	2	0.000919764
6.2	Languages and literature	Languages and literature	2	0.000919764	2	0.000919764
6.3		Philosophy, ethics and religion	0	0	0	0
6.4		Art (arts, history of arts, performing arts, music)	0	0	0	0
6.5		Other humanities	0	0	0	0

Figura 4.10: Estatísticas por domínio (detalhada)

permitir a melhor etiqueta quando há inúmeras possibilidades. Assim, fica registada a necessidade do que será denominado, calibração do classificador.



Stats by Language

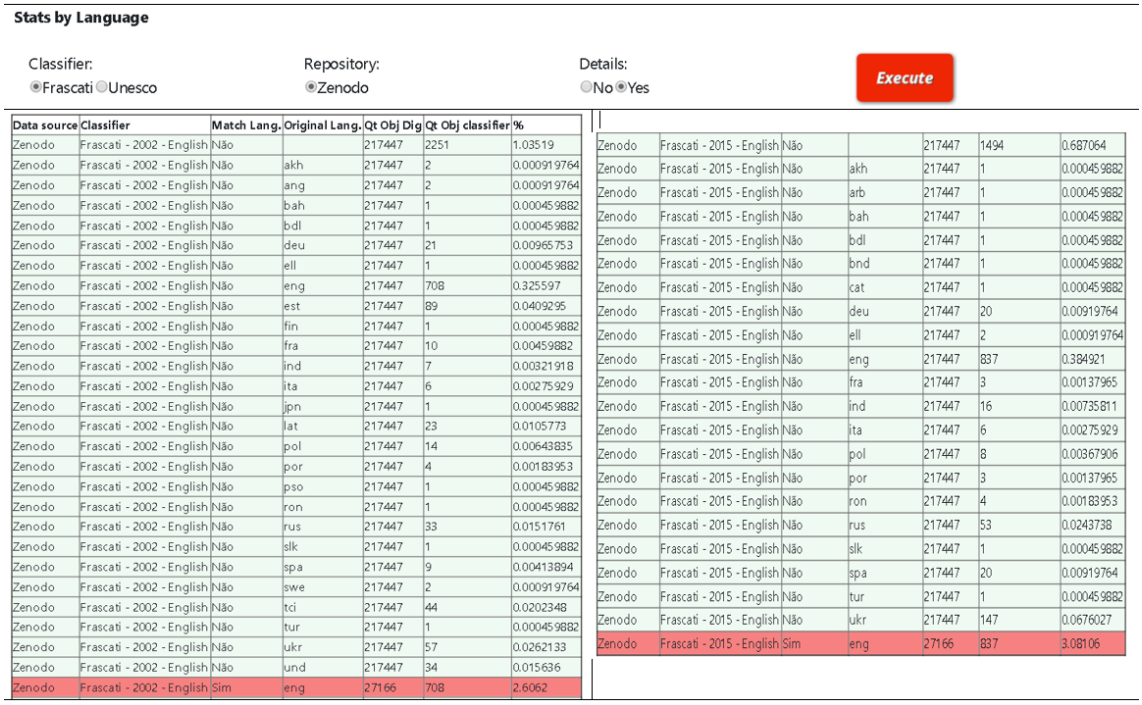
Classifier: ◉Frascati ○Unesco

Repository: ◉Zenodo

Details: ◉No ○Yes

Data source	Classifier	Match Lang.	Original Lang.	Qt Obj Dig	Qt Obj classifier	%
Zenodo	Frascati - 2002 - English	Não		217447	3325	1.52911
Zenodo	Frascati - 2002 - English	Sim		27166	708	2.6062
Zenodo	Frascati - 2015 - English	Não		217447	2621	1.20535
Zenodo	Frascati - 2015 - English	Sim		27166	837	3.08106

Figura 4.11: Estatísticas por linguagem - *Frascati*



Stats by Language

Classifier: ◉Frascati ○Unesco

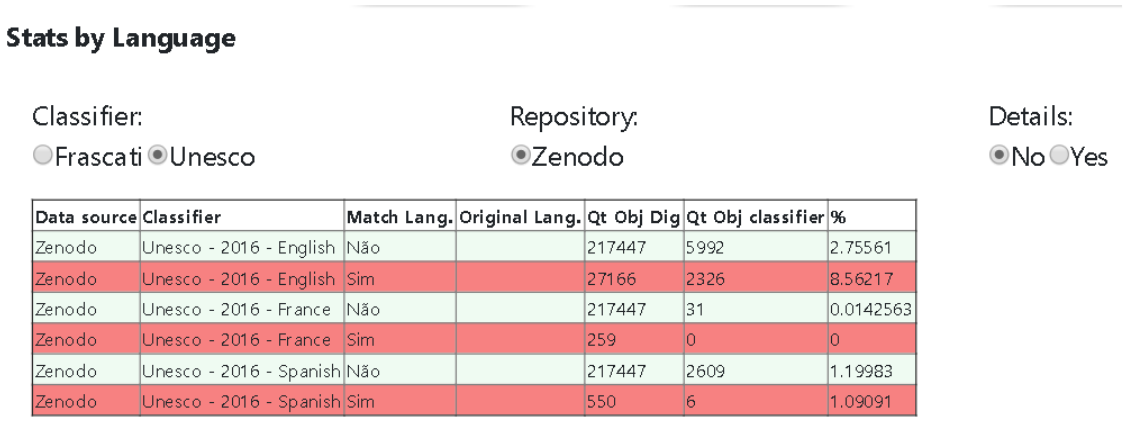
Repository: ◉Zenodo

Details: ○No ◉Yes

Execute

Data source	Classifier	Match Lang.	Original Lang.	Qt Obj Dig	Qt Obj classifier	%
Zenodo	Frascati - 2002 - English	Não		217447	2251	1.03519
Zenodo	Frascati - 2002 - English	Não	akh	217447	2	0.000919764
Zenodo	Frascati - 2002 - English	Não	ang	217447	2	0.000919764
Zenodo	Frascati - 2002 - English	Não	bah	217447	1	0.000459882
Zenodo	Frascati - 2002 - English	Não	bdl	217447	1	0.000459882
Zenodo	Frascati - 2002 - English	Não	deu	217447	21	0.00965753
Zenodo	Frascati - 2002 - English	Não	ell	217447	1	0.000459882
Zenodo	Frascati - 2002 - English	Não	eng	217447	708	0.325597
Zenodo	Frascati - 2002 - English	Não	est	217447	89	0.0409295
Zenodo	Frascati - 2002 - English	Não	fin	217447	1	0.000459882
Zenodo	Frascati - 2002 - English	Não	fra	217447	10	0.00459882
Zenodo	Frascati - 2002 - English	Não	ind	217447	7	0.00321918
Zenodo	Frascati - 2002 - English	Não	ita	217447	6	0.00275929
Zenodo	Frascati - 2002 - English	Não	jpn	217447	1	0.000459882
Zenodo	Frascati - 2002 - English	Não	lat	217447	23	0.0105773
Zenodo	Frascati - 2002 - English	Não	pol	217447	14	0.00643835
Zenodo	Frascati - 2002 - English	Não	por	217447	4	0.00183953
Zenodo	Frascati - 2002 - English	Não	pso	217447	1	0.000459882
Zenodo	Frascati - 2002 - English	Não	ron	217447	1	0.000459882
Zenodo	Frascati - 2002 - English	Não	rus	217447	33	0.0151761
Zenodo	Frascati - 2002 - English	Não	slk	217447	1	0.000459882
Zenodo	Frascati - 2002 - English	Não	spa	217447	9	0.00413894
Zenodo	Frascati - 2002 - English	Não	swe	217447	2	0.000919764
Zenodo	Frascati - 2002 - English	Não	tci	217447	44	0.0202348
Zenodo	Frascati - 2002 - English	Não	tur	217447	1	0.000459882
Zenodo	Frascati - 2002 - English	Não	ukr	217447	57	0.0262133
Zenodo	Frascati - 2002 - English	Não	und	217447	34	0.015636
Zenodo	Frascati - 2002 - English	Sim	eng	27166	708	2.6062
Zenodo	Frascati - 2015 - English	Não		217447	1494	0.687064
Zenodo	Frascati - 2015 - English	Não	akh	217447	1	0.000459882
Zenodo	Frascati - 2015 - English	Não	arb	217447	1	0.000459882
Zenodo	Frascati - 2015 - English	Não	bah	217447	1	0.000459882
Zenodo	Frascati - 2015 - English	Não	bdl	217447	1	0.000459882
Zenodo	Frascati - 2015 - English	Não	bnd	217447	1	0.000459882
Zenodo	Frascati - 2015 - English	Não	cat	217447	1	0.000459882
Zenodo	Frascati - 2015 - English	Não	deu	217447	20	0.00919764
Zenodo	Frascati - 2015 - English	Não	ell	217447	2	0.000919764
Zenodo	Frascati - 2015 - English	Não	eng	217447	837	0.384921
Zenodo	Frascati - 2015 - English	Não	fra	217447	3	0.00137965
Zenodo	Frascati - 2015 - English	Não	ind	217447	16	0.00735811
Zenodo	Frascati - 2015 - English	Não	ita	217447	6	0.00275929
Zenodo	Frascati - 2015 - English	Não	pol	217447	8	0.00367906
Zenodo	Frascati - 2015 - English	Não	por	217447	3	0.00137965
Zenodo	Frascati - 2015 - English	Não	ron	217447	4	0.00183953
Zenodo	Frascati - 2015 - English	Não	rus	217447	53	0.0243738
Zenodo	Frascati - 2015 - English	Não	slk	217447	1	0.000459882
Zenodo	Frascati - 2015 - English	Não	spa	217447	20	0.00919764
Zenodo	Frascati - 2015 - English	Não	tur	217447	1	0.000459882
Zenodo	Frascati - 2015 - English	Não	ukr	217447	147	0.0676027
Zenodo	Frascati - 2015 - English	Sim	eng	27166	837	3.08106

Figura 4.12: Estatísticas por linguagem - *Frascati* (detalhada)



Stats by Language

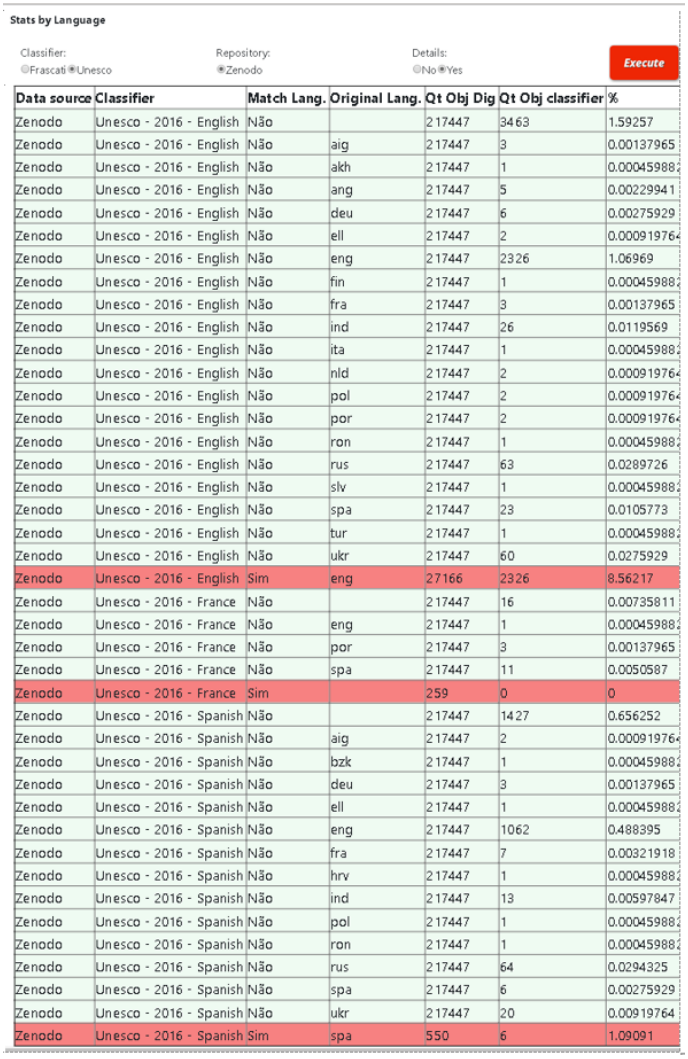
Classifier: ○Frascati ◉Unesco

Repository: ◉Zenodo

Details: ◉No ○Yes

Data source	Classifier	Match Lang.	Original Lang.	Qt Obj Dig	Qt Obj classifier	%
Zenodo	Unesco - 2016 - English	Não		217447	5992	2.75561
Zenodo	Unesco - 2016 - English	Sim		27166	2326	8.56217
Zenodo	Unesco - 2016 - France	Não		217447	31	0.0142563
Zenodo	Unesco - 2016 - France	Sim		259	0	0
Zenodo	Unesco - 2016 - Spanish	Não		217447	2609	1.19983
Zenodo	Unesco - 2016 - Spanish	Sim		550	6	1.09091

Figura 4.13: Estatísticas por linguagem - *Unesco*

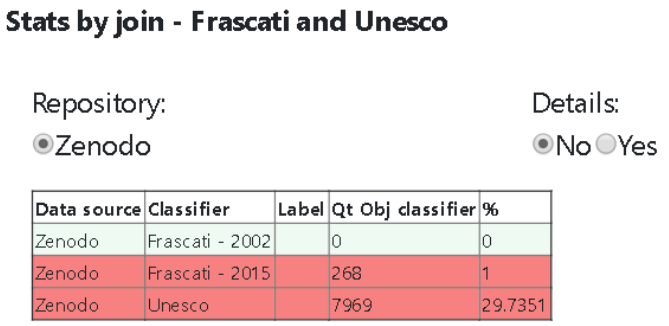


Stats by Language

Classifier: ○Frascati ◉Unesco | Repository: ◉Zenodo | Details: ○No ◉Yes | Execute

Data source	Classifier	Match Lang.	Original Lang.	Qt Obj Dig	Qt Obj classifier	%
Zenodo	Unesco - 2016 - English	Não		217447	3463	1.59257
Zenodo	Unesco - 2016 - English	Não	aig	217447	3	0.00137965
Zenodo	Unesco - 2016 - English	Não	akh	217447	1	0.00045988
Zenodo	Unesco - 2016 - English	Não	ang	217447	5	0.00229941
Zenodo	Unesco - 2016 - English	Não	deu	217447	6	0.00275929
Zenodo	Unesco - 2016 - English	Não	ell	217447	2	0.00091976
Zenodo	Unesco - 2016 - English	Não	eng	217447	2326	1.06969
Zenodo	Unesco - 2016 - English	Não	fin	217447	1	0.00045988
Zenodo	Unesco - 2016 - English	Não	fra	217447	3	0.00137965
Zenodo	Unesco - 2016 - English	Não	ind	217447	26	0.0119569
Zenodo	Unesco - 2016 - English	Não	ita	217447	1	0.00045988
Zenodo	Unesco - 2016 - English	Não	nld	217447	2	0.00091976
Zenodo	Unesco - 2016 - English	Não	pol	217447	2	0.00091976
Zenodo	Unesco - 2016 - English	Não	por	217447	2	0.00091976
Zenodo	Unesco - 2016 - English	Não	ron	217447	1	0.00045988
Zenodo	Unesco - 2016 - English	Não	rus	217447	63	0.0289726
Zenodo	Unesco - 2016 - English	Não	slv	217447	1	0.00045988
Zenodo	Unesco - 2016 - English	Não	spa	217447	23	0.0105773
Zenodo	Unesco - 2016 - English	Não	tur	217447	1	0.00045988
Zenodo	Unesco - 2016 - English	Não	ukr	217447	60	0.0275929
Zenodo	Unesco - 2016 - English	Sim	eng	27166	2326	8.56217
Zenodo	Unesco - 2016 - France	Não		217447	16	0.00735811
Zenodo	Unesco - 2016 - France	Não	eng	217447	1	0.00045988
Zenodo	Unesco - 2016 - France	Não	por	217447	3	0.00137965
Zenodo	Unesco - 2016 - France	Não	spa	217447	11	0.0050587
Zenodo	Unesco - 2016 - France	Sim		259	0	0
Zenodo	Unesco - 2016 - Spanish	Não		217447	1427	0.656252
Zenodo	Unesco - 2016 - Spanish	Não	aig	217447	2	0.00091976
Zenodo	Unesco - 2016 - Spanish	Não	bzk	217447	1	0.00045988
Zenodo	Unesco - 2016 - Spanish	Não	deu	217447	3	0.00137965
Zenodo	Unesco - 2016 - Spanish	Não	ell	217447	1	0.00045988
Zenodo	Unesco - 2016 - Spanish	Não	eng	217447	1062	0.488395
Zenodo	Unesco - 2016 - Spanish	Não	fra	217447	7	0.00321918
Zenodo	Unesco - 2016 - Spanish	Não	hrv	217447	1	0.00045988
Zenodo	Unesco - 2016 - Spanish	Não	ind	217447	13	0.00597847
Zenodo	Unesco - 2016 - Spanish	Não	pol	217447	1	0.00045988
Zenodo	Unesco - 2016 - Spanish	Não	ron	217447	1	0.00045988
Zenodo	Unesco - 2016 - Spanish	Não	rus	217447	64	0.0294325
Zenodo	Unesco - 2016 - Spanish	Não	spa	217447	6	0.00275929
Zenodo	Unesco - 2016 - Spanish	Não	ukr	217447	20	0.00919764
Zenodo	Unesco - 2016 - Spanish	Sim	spa	550	6	1.09091

Figura 4.14: Estatísticas por linguagem - *Unesco* (detalhada)



Stats by join - Frascati and Unesco

Repository: ◉Zenodo | Details: ◉No ○Yes

Data source	Classifier	Label	Qt Obj classifier	%
Zenodo	Frascati - 2002		0	0
Zenodo	Frascati - 2015		268	1
Zenodo	Unesco		7969	29.7351

Figura 4.15: Estatísticas por *join - Unesco*

Stats by join - Frascati and Unesco

Repository: Details:

◉ Zenodo ○ No ◉ Yes **Execute**

Data source	Classifier	Label	Qt Obj classifier	%
Zenodo	Frascati - 2002	Mathematics and computer sciences	0	0
Zenodo	Frascati - 2015	Mathematics	268	1
Zenodo	Unesco	Extrapolation (Extrapolación or Extrapolation)	26	0.097014
Zenodo	Unesco	Interpolation (Interpolación or Interpolation)	94	0.350746
Zenodo	Unesco	Statistical inference (Inferencia estadística or Inférence statistique)	9	0.033582
Zenodo	Unesco	Time series (Series temporales or Série temporelle)	259	0.966417
Zenodo	Unesco	Probability theory (Teoría de las probabilidades or Théorie des probabilités)	6	0.022388
Zenodo	Unesco	Geometry (Geometría or Géométrie)	386	1.4403
Zenodo	Unesco	Mathematics (Matemáticas or Mathématiques)	275	1.02612
Zenodo	Unesco	Statistics (Estadística or Statistique)	946	3.52985
Zenodo	Unesco	Mathematical logic (Lógica matemática or Logique mathématique)	3	0.011194
Zenodo	Unesco	Mathematical models (Modelo matemático or Modèle mathématique)	70	0.261194
Zenodo	Unesco	Data visualization (Visualización de datos or Visualisation de données)	31	0.115671
Zenodo	Unesco	Algebra (Álgebra or Algèbre)	169	0.630597
Zenodo	Unesco	Number theory (Teoría de los números or Théorie des nombres)	14	0.052238
Zenodo	Unesco	Algorithms (Algoritmo or Algorithme)	1146	4.27612
Zenodo	Unesco	Regression analysis (Análisis de regresión or Analyse de régression)	271	1.01119
Zenodo	Unesco	Variance analysis (Análisis de variancia or Analyse de variance)	22	0.082089
Zenodo	Unesco	Statistical analysis (Análisis estadístico or Analyse statistique)	302	1.12687
Zenodo	Unesco	Factor analysis (Análisis factorial or Analyse factorielle)	101	0.376865
Zenodo	Unesco	Functional analysis (Análisis funcional or Analyse fonctionnelle)	13	0.048507
Zenodo	Unesco	Mathematical analysis (Análisis matemático or Analyse mathématique)	13	0.048507
Zenodo	Unesco	Multivariate analysis (Análisis multivariado or Analyse multivariée)	41	0.152985
Zenodo	Unesco	Numerical analysis (Análisis numérico or Analyse numérique)	61	0.227611
Zenodo	Unesco	Arithmetic (Aritmética or Arithmétique)	111	0.414179
Zenodo	Unesco	Calculus (Cálculo or Calcul)	2418	9.02239
Zenodo	Unesco	Graph theory (Teoría de los gráficos or Théorie des graphes)	46	0.171641
Zenodo	Unesco	Set theory (Teoría de los conjuntos or Théorie des ensembles)	49	0.182835
Zenodo	Unesco	Topology (Topología or Topologie)	322	1.20149
Zenodo	Unesco	Statistical data (Datos estadísticos or Donnée statistique)	68	0.253731
Zenodo	Unesco	Correlation (Correlación or Corrélation)	1410	5.26119
Zenodo	Unesco	Random processes (Proceso aleatorio or Processus aléatoire)	1	0.003731
Zenodo	Unesco	Equations (Ecuación or Équation)	669	2.49627

Figura 4.16: Estatísticas por *join - Unesco* (detalhada)

Capítulo 5

Conclusões

O trabalho permeou questões relativas à classificação de objetos digitais de uma maneira prática. Não se pode furtar que devido à abordagem abrangente requer ainda muito estudo. O que se pode afirmar é que foram abertas novas possibilidades ao tema tratado.

A solução apresentada procurou unificar os estudos de duas áreas complementares: Ciência da Informação e Ciência da Computação. Poderia ser comparado a montagem de um *puzzle*. Reuniu "peças"da Ciência da Informação a partir de estruturas já consolidadas naquele campo e selecionou técnicas da Ciência da Computação para produção de uma ferramenta de recuperação de informação.

O capítulo anterior, dedicado à exposição dos resultados obtidos contém detalhes das experiências realizadas agrupadas na secção "Estatísticas", das quais derivam boa parte das conclusões obtidas.

As conclusões são apresentadas sob o título de "Visão prática", pois este pode ser o principal contributo do trabalho. Sem fugir da base teórica demonstrou experiências reais e factíveis.

5.1 Visão prática

A figura 5.1 expõe os resultados alcançados que serão explicados em seguida.

A solução desenvolvida apresentou três insumos, isto é, utilizou a combinação de três recursos triviais. Primeiramente, a utilização de metadados extremamente simples como é o caso do título, descrição dos objetos e palavras-chave, ou seja, metadados textuais, comuns a praticamente qualquer documento. Em segundo lugar, o uso de sistemas de classificação e por fim, os tesauros.

A conclusão relativa ao desempenho dos recursos citados pode ser realizada de duas maneiras a depender da associação entre eles. A primeira abordagem, da simples utilização do sistema de classificação foi refutada de acordo com as estatísticas referidas anteriormente. Mediante a técnica utilizada, a aplicação do *Frascati* 2002 e 2015 demonstrou de forma inequívoca que o simples uso de um sistema de classificação não é suficiente para um classificador eficiente.

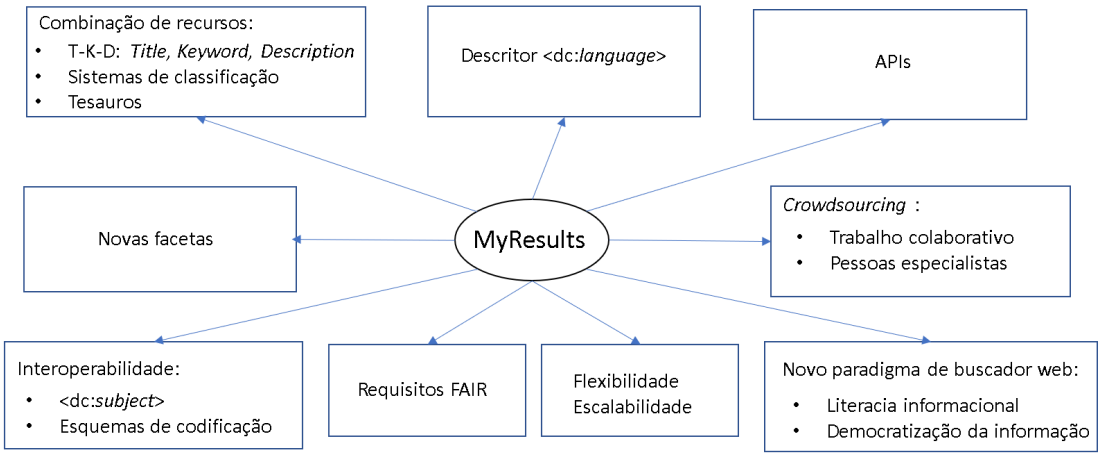


Figura 5.1: Conclusões

Todavia, as estatísticas também evidenciaram que a utilização de termos dos tesauros em combinação com os sistemas de classificação tendem a atingir um melhor resultado no classificador em estudo.

Outra importante conclusão refere-se ao metadado linguagem, ou seja, ao descritor `<dc:language>`. Várias experiências demonstradas na secção "Estatísticas", supra citada, sugerem a importância desse metadado. A atribuição de uma etiqueta classificadora em virtude de se usar o sistema de classificação ou tesauro em conformidade com a linguagem do recurso aumenta a quantidade de objetos classificados quando se faz o *match* de linguagem.

Uma exceção constatada porém, evidencia uma interessante questão que pode ser atribuída ao possível preenchimento errado do referido campo. Foi constatado o decréscimo na quantidade de objetos classificados quando foi utilizado o idioma francês em uma das experiências.

Convém referir a questão dos dados analisados relativamente a cada *API* oferecida pelo *Zenodo*. Embora a princípio tenham sido utilizados somente os metadados considerados simples, nomeadamente título, palavras-chave e descrição dos objetos, disponíveis nos ficheiros *XML*, o estudo dos metadados adicionais encontrados nos ficheiros *JSON* possibilitaram uma visão mais alargada do conteúdo de cada objeto digital. Acabaram de forma indireta, sendo importantes para perceção de atributos complementares e significativos na composição do classificador, capazes de explicar relações implícitas nos dados, além de apontar possíveis novas facetas ao classificador.

Portanto, embora a *API* que disponibiliza os ficheiros *JSON* não implique a interoperabilidade preconizada pelo protocolo *OAI-PMH*, os descritores fornecidos naquela solução contribuíram efetivamente para compreensão dos relacionamentos entre os diversos metadados.

Relativamente à questão das facetas, conforme destacado no capitulo dedicado ao portal, cumpre enfatizar o breve estudo relativo a uma possível faceta em relação a origem dos dados. Foram propostos quatro grandes ramos: dados bibliográficos, *web sites*, *wikis* e *database* no primeiro nível da estrutura. No caso de serem do tipo bibliográficos poderiam à partida, serem do tipo

repositórios, museus ou bibliotecas. A estrutura foi desenhada a permitir outras categorizações possíveis.

É bem possível que esta componente dos dados indique não somente uma nova forma de etiquetar os dados, mas que permita estabelecer a relação entre o tipo de fonte de dados e os melhores sistemas de classificação e vocabulários controlados pertinentes.

A utilização de ferramentas de *crowdsourcing* planeada para a fase de validação está alinhada com técnicas relacionadas à própria evolução do meio digital. Normalmente, serviços de *crowdsourcing* são pagos mas pode haver lugar a trabalho voluntário. Como o trabalho proposto tem como filosofia a questão democrática e a busca da literacia da informação pode-se pensar no trabalho colaborativo para sua execução, ou seja, uma motivação social.

Tão importante como a questão colaborativa é o fato da necessidade de respostas baseadas em experiência profissional que possam atender com qualidade a validação proposta. A participação de pessoas especialistas no tratamento da informação é fundamental. A contribuição corroborada por autoridades técnicas envolvidas no processo de indexação e classificação pode conferir a chancela necessária ao processo.

Relativamente à questão da interoperabilidade, a ideia apresentada não exige a criação de novos metadados, uma vez que é necessário somente incluir uma entrada no descritor <*dc:subject*> do *Dublin Core*, desenvolvido justamente para representar entre outros assuntos, os códigos de classificação. É importante ainda enfatizar que várias referências podem ser incluídas, especialmente dentro do mesmo domínio, como é o caso por exemplo de classificações bibliográficas. Um documento de um repositório ou de uma biblioteca podem fazer referências às classificações *FOS, CDU, LCC, UDC, DDC* e outras, em simultâneo.

A iniciativa *DCMI* recomenda a utilização de esquemas de codificação específicos ao descritor <*dc:subject*> nomeadamente em relação aos códigos de classificação, o que pode facilitar ainda mais a interoperabilidade. Mediante a possibilidade de consignação dos resultados no formato de esquemas de codificação do *DCMI* pode-se falar que o classificador satisfaz aos quatro princípios *FAIR*, uma vez que contribui para que os dados sejam localizáveis, acessíveis, interoperáveis e reutilizáveis.

De acordo com as 14 métricas para quantificar os níveis de *FAIR*, a classificação automática apresentada alcança 6 pontos, destacados a seguir:

- **F3:** Os (meta) dados são registados ou indexados em um recurso pesquisável;
- **A1:** Os (meta) dados são recuperáveis pelo seu identificador usando um protocolo de comunicação padronizado;
- **A2:** Os metadados estão acessíveis, mesmo quando os dados não estão mais disponíveis.
- **I1:** Os (meta) dados usam uma linguagem formal, acessível, compartilhada e amplamente aplicável para a representação do conhecimento.
- **R1:** Os meta (dados) tem uma pluralidade de atributos precisos e relevantes.

- **R1.3:** Os (meta) dados atendem aos padrões da comunidade relevantes ao domínio.

Uma conclusão importante refere-se à flexibilidade e escalabilidade da solução uma vez que a escolha do sistema de classificação é uma liberalidade do curador dos dados em conformidade com o domínio de cada repositório. O mesmo acontece com a escolha do tesauro, que de maneira óbvia, deve priorizar aqueles específicos em conformidade com o tema tratado. Ao classificar documentos do domínio da Medicina por exemplo, buscar-se-ão tesauros projetados para o ambiente médico.

É possível também utilizar mais de um sistema de classificação para cada repositório. Em virtude da interoperabilidade pode-se escolher mais de um sistema de classificação simultaneamente.

Além da flexibilidade relativa ao sistema de classificação, a solução foi projetada para trabalhar com outros repositórios. O estudo limitou-se a utilizar os dados do *Zenodo*, mas o classificador está preparado para trabalhar com várias fontes de dados.

Por fim, cabe referir à possibilidade de uma nova alternativa ao paradigma dos buscadores *web* tradicionais. O agrupamento dos resultados, isto é, a exibição dos resultados obtidos com o classificador serem apresentados de acordo com o domínio a que pertencem, pode contribuir para uma melhor escolha do resultado por parte do utente. Em outras palavras, os resultados não são apresentados de acordo com o perfil ou capacidade de pesquisa do utilizador. Essa configuração pode permitir ao utilizador verificar de uma forma abrangente as possibilidades para sua consulta e a partir de seu interesse selecionar o conteúdo desejado. Essa "fuga"das "bolhas de filtro"pode proporcionar um acesso mais democrático à informação e também maior literacia aos utentes.

O carácter exploratório proporcionado pela investigação abre oportunidade a diversos estudos futuros que serão tratados na próxima sessão.

5.2 Trabalhos futuros

Para explicar a secção relativa aos trabalhos futuros as sugestões foram agrupadas na figura 5.2 cujos detalhes são abordados seguidamente.

Embora o estudo esteja assente maioritariamente na fase da classificação propriamente dita, a depender da validação do resultado e da fase final nomeada como "Apresentação dos resultados", a proposição de trabalhos futuros do atual projeto é analisada de acordo com as três fases referidas.

5.2.1 Classificação automática de documentos

A proposta de trabalhos futuros relativamente à primeira fase do estudo, ou seja, referente à classificação automática de documentos está subdividida em três sugestões: "Novos recursos", "Estudo de metadados"e "Ferramentas externas".

- **Novos recursos**

  Conforme referido na secção anterior, os três recursos que dão sustentação ao projeto, nomeadamente a fonte de dados representada pelos objetos a serem classificados, o sistema

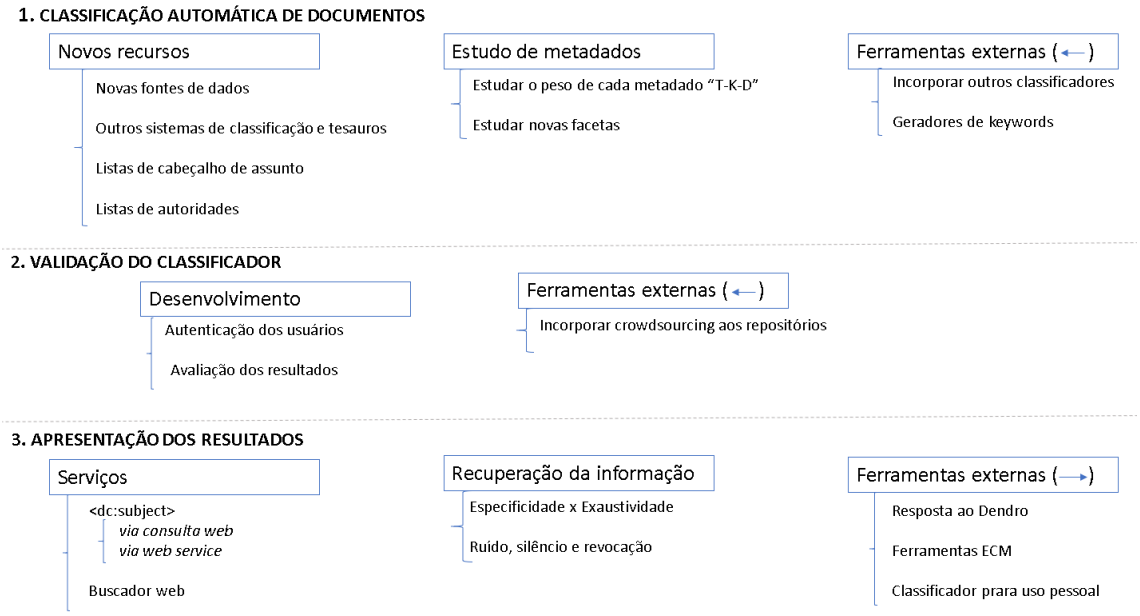


Figura 5.2: Trabalhos futuros

de classificação e o tesauro adotado são totalmente parametrizáveis. Para cada conjunto de dados a serem classificados não há que se preocupar com a escolha de um único sistema de classificação e nem tampouco o tesauro.

Portanto, há imenso trabalho no estudo relativo às diversas possibilidades de sistemas de classificação e tesauro de acordo com as diversas fontes de dados. A título de exemplo, pode-se estudar as características de certos repositórios temáticos a fim de que sejam indicados os sistemas de classificação, assim como o tesauros que melhor representem o tema de cada fonte de dados.

Os tesauros são designados como vocabulários controlados, importante ferramenta utilizada na classificação, indexação e recuperação de documentos. Além dos tesauros há as listas de cabeçalho de assunto e as listas de autoridades. O trabalho apresentado explorou de forma bem limitada o uso dos tesauros, mas como trabalho futuro, a utilização das listas de cabeçalho de assunto e as listas de autoridades são instrumentos que podem auxiliar o processo de classificação de objetos digitais.

Ainda em relação aos tesauros, foi planeado o desenvolvimento de um mecanismo de termos excludentes durante o processo classificatório de forma a assegurar a remoção de possíveis classificações erróneas o qual enseja um trabalho importante a ser desenvolvido.

Outro trabalho paralelo poderia verificar a flexibilização dos sistemas de classificação. O objeto de estudo nesse caso seria a especificação do sistema *FOS* no âmbito do Sistema

Estatístico Nacional de Portugal[1]. Poderá ser interessante, estudar a necessidade e a capacidade do sistema automático de classificação ora proposto em considerar especificidades de determinadas organizações. Por um lado podem trazer melhoria na recuperação da informação ao estender os níveis e denominações mas podem demarcar uma incompatibilidade de versões e ao longo prazo impactar negativamente a recuperação da informação.

- **Estudo de metadados**

  No decorrer do trabalho foi abordada a questão relativa à atribuição de um *score* a cada metadado *Title - Description - Keyword (T-K-D)*. O estudo a ser realizado prende-se à análise do "acerto"de cada descritor separadamente e também em conjunto. Esta análise permite verificar a potencialidade dos descritores de maneira geral ou específica.

  Pode-se por exemplo, verificar a assertividade do classificador em relação às editoras de livros, podendo sugerir que o melhor preenchimento do descritor *description* assegura maior acerto no processo de classificação da respetiva editora.

  A classificação facetada sempre pautou os trabalhos na busca de variáveis que pudessem auxiliar o processo classificatório sendo parte do escopo da presente investigação. Assim, algumas componentes analisadas na fase de preparação dos dados podem fornecer subsídios para novos enquadramentos e estudos futuros. Algumas possibilidades versam sobre os metadados referentes a autor, afiliação, tipo de recurso e comunidades.

  Além da descoberta de novas facetas poderá ser ampliado o estudo relativo à faceta origem dos dados conforme relatado anteriormente.

- **Ferramentas externas**

  O portal é a ferramenta visual para o sistema classificador desenvolvido. Este último foi concebido como uma estrutura bastante sólida capaz de atender aos mais diversos requisitos de funcionalidade além da capacidade analítica e sintética de avaliação de resultados. Por isso mesmo, nada impede que classificadores externos possam ser incorporados à solução.

  Em virtude da possível capacidade de atender aos mais diversos domínios poderão ser utilizadas outras técnicas, como por exemplo o uso de *data mining* e ferramentas de inteligência artificial para classificação de documentos. Em outras palavras, podem ser usados classificadores já desenvolvidos em outros projetos ou mesmo, serem desenvolvidos novos classificadores a atenderem determinados cenários.

  Ferramentas complementares podem ser de grande valia ao classificador, especialmente aquelas voltadas à semântica do conteúdo dos objetos digitais. Assim, serviços que possam ser incorporados ao classificador, como os "Geradores de keywords"podem aprimorar significativamente o processo classificatório estudado.

[1] http://www.dgeec.mec.pt/np4/28/%7B\protect\T1\textdollarclientServletPath%7D/?newsId=26&fileName=Classificacao_FOS_VersaoPortuguesa.pdf

5.2.2 Validação do classificador

A sugestão de trabalhos futuros no que tange à validação do classificador foi subdividida em duas sugestões: "Desenvolvimento"e "Ferramentas externas".

- **Desenvolvimento**

  A etapa de validação do classificador foi projetada para ser executada por meio de *crowd-sourcing*. Essa funcionalidade foi desenvolvida uma vez que os dados utilizados para o classificador não permitiam um mecanismo automático de validação, como por exemplo, dados etiquetados, ou seja, dados com as classificações já pré-determinadas.

  A opção conhecida como anotação de dados, ou seja, fazer a classificação manual para uma possível comparação com o resultado apurado também não se mostrou oportuna.

  Portanto, há que se referir a necessidade de continuidade no desenvolvimento dessa ferramenta para validação dos resultados. Aqui cabe lugar ao processo de autenticação dos utentes e análise dos resultados obtidos.

- **Ferramentas externas**

  Uma possível oportunidade de fomentar a utilização desse mecanismo de validação pode ser sua inclusão dentro dos próprios repositórios. Ao invés de uma componente exclusiva dentro do próprio portal, pode vir a fazer parte de ferramentas que lidam com a gestão da informação, especificamente aquelas que tratam de metadados.

5.2.3 Apresentação dos resultados

Nesse tópico estão inseridas três sugestões de apresentação dos resultados do classificador que podem ser transformadas em trabalhos futuros: "Serviços", "Recuperação da Informação"e "Ferramentas externas".

- **Serviços**

  O desdobramento da apresentação dos resultados pode ser um interessante caso de estudo isto é, aprimorar as ferramentas de pesquisa nos três formatos propostos que se enquadram em dois serviços.

  Ao primeiro serviço é suposto atribuir uma tag <*dc:subject*> com o código classificador. No alinhamento dessa proposta é preciso disponibilizar a *tag* através de uma consulta na página do sistema classificador a qualquer utente que queira utilizar o classificador. Esse serviço precisa também estar disponível através de um protocolo de comunicação de dados, interpretável por máquinas.

  O segundo serviço demanda ainda mais trabalho em virtude dos requisitos de privacidade, democracia e literacia a serem atendidos na construção de um buscador dos resultados do classificador.

- **Recuperação da informação**

  Alguns problemas relativos à indexação de documentos, problemas clássicos da Ciência da Informação podem ser estudados em continuidade ao presente projeto. Entre eles, destacam-se os problemas relativos à recuperação da informação, como especificidade e exaustividade, pontos relacionados respetivamente a uma abordagem mais qualitativa ou a aspetos quantitativos dos objetos classificados.

  Podem também ser estudadas ainda, as questões relativas à recuperação da informação no que tange a problemas de ruído e silêncio que implicam em última análise, a aspetos relacionados ao interesse da comunidade de utentes e a imparcialidade da classificação. Aqui cabem estudos relativos à revocação significando em termos simples, à capacidade de recuperação de documentos relevantes e à precisão, colocada de uma maneira bem descomplicada, como a capacidade de impedir a recuperação de documentos não relevantes.

- **Ferramentas externas** Não pode deixar de ser referenciado como trabalho futuro, o aporte que deu mote à presente dissertação, ou seja, a indicação de metadados específicos de acordo com cada domínio. Parece ser um trabalho que demanda tempo, mas é completamente exequível e com excelentes perspectivas de concretização.

  A classificação de documentos pode ser aplicada a repositórios, a exemplificar seu uso em grandes redes, mas pode ser estendida a redes menores como por exemplo a redes corporativas, como uma ferramenta *Enterprise Content Management - ECM*. Pode inclusive ser aplicada às redes privadas até mesmo para classificar documentos em um portátil pessoal.

5.3 Considerações finais

O presente trabalho tinha três objetivos principais, dos quais pretendia-se abordar os dois primeiros, sendo o terceiro desde o princípio atribuído à trabalhos futuros.

O primeiro e principal objetivo que enseja o título da dissertação, "Classificação de conjuntos de dados de investigação com base em seus registos de metadados"foi atingido.

O segundo objetivo que seria a indicação de metadados específicos para o Repositório Dendro não foi completamente alcançado em virtude da ampla abordagem ao principal objetivo do trabalho proposto. Entretanto, pode-se considerar que foi parcialmente atingido em virtude das questões de interoperabilidade estudadas.

Sem dúvida, o tempo despendido a tarefa mais importante do projeto em detrimento do cumprimento do objetivo subsidiário, a indicar os metadados específicos ao Dendro, foi uma decisão acertada. O esforço foi concentrado na busca de uma solução abrangente e dinâmica e por isso mesmo, mais consolidada a dar suporte ao Dendro.

Além do cariz eminentemente prático da solução apresentada, merecem destaque a flexibilidade e escalabilidade incorporadas, que podem justificar a continuidade do projeto.

Essas componentes, flexibilidade e escalabilidade são o resultado da preparação dos objetos do banco de dados, o qual foi projetado de modo a permitir uma gama de possibilidades de execução e atendimento de vários cenários estendendo a classificação de documentos a vários ambientes.

Importante ressaltar ainda, como pode ser observado na figura anterior, o destaque às ferramentas externas, ou seja, a questão voltada à capacidade interoperável do classificador. A interoperabilidade está planeada a ser realizada nos dois sentidos, de entrada e saída de dados. Em outras palavras, a estrutura está construída de forma a receber contribuições de outras ferramentas e também a fornecer serviço a outras ferramentas através de protocolos de comunicação já implementados na *web*.

Importante ainda é destacar a questão da subjetividade no processo de classificação, conforme demonstrado no capítulo 3, em que o mesmo livro, dentro da mesma instituição e a usar o mesmo sistema de classificação obteve dois códigos de classificação diferentes. É incontestável portanto, a dificuldade imposta pela subjetividade de quem está a classificar. Portanto faz-se necessária a utilização de ferramentas que possam tornar o processo mais universal nos dois cenários: ambiente físico e digital.

Em última análise ficou evidente a importância da classificação de objetos digitais. A conclusão final pode ser resumida a um trabalho que cumpriu o principal objetivo e abriu novas possibilidades para a classificação de documentos digitais a partir de experiências realizadas de forma prática e voltadas às premissas de democracia e literacia da informação.

Referências

Alves, R. C. V. U. (2010). Metadados como elementos do processo de catalogação. *Aleph*, 132 f. : il.

Araya, E. R. M. and S. A. B. G. Vidotti (2010). *Criação, proteção e uso legal de informação em ambientes da World Wide Web*. UNESP, São Paulo.

Baca, M. (2016). *Intro to Metadata* (3rd ed.). Getty Research Institute, Los Angeles.

Baptista, A. A. (2017). Desafios à comunidade ibero-americana de metadados em repositórios digitais para maximização da interoperabilidade. pp. 193–204.

Bartling, S. and S. Friesike (2014). Opening Science. *Opening Science*, 213–224.

Campos, M. L. d. A. and H. E. Gomes (2003). Organização de domínio de conhecimento e os princípios ranganathianos. *Perspect. cienc. inf., Belo Horizonte 8*(2), 150–163.

Chapman, J. W., D. Reynolds, and S. A. Shreeves (2009). Repository metadata: Approaches and challenges. *Cataloging and Classification Quarterly 47*(3-4), 309–325.

Cocco, A. P. (2012). Repositórios Institucionais de Acesso Aberto: Análise do Cenário nos Países Ibero-Americanos. *Revista Eletrônica de Biblioteconomia e Ciência da Informação 17*(35).

Decker, S., S. Melnik, F. van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann, and I. Horrocks (2000). The Semantic Web: the roles of XML and RDF. *IEEE Internet Computing 4*(5), 63–73.

Duval, E., W. Hodgins, S. Sutton, and S. L. Weibel (2002). Metadata Principles and Practicalities. *D-Lib Magazine 8*(4).

Formenton, D., F. F. de Castro, L. D. S. Gracioso, A. C. M. Furnival, and M. D. G. d. M. Simões (2018, jan). Os padrões de metadados como recursos tecnológicos para a garantia da preservação digital. *Biblios: Journal of Librarianship and Information Science* (68), 82–95.

Furtado, F., P. Príncipe, and J. Carvalho (2017). Kit sobre dados de investigação RCAAP. pp. 1–34.

Hey, T. (2002). The UK e-Science Program and the Grid. pp. 6–6.

Hey, T. and J. Hey (2006, oct). e-Science and its implications for the library community. *Library Hi Tech 24*(4), 515–528.

Higgins, S. (2007). Using Metadata Standards.

Higgins, S. (2008). The DCC Curation Lifecycle Model. *The Internationalo Journal of Digital Curation 3*(1).

IEEE - Institute of Electrical and Electronics Engineers (2014). IEEE International Conference on eScience | a web site for the conference series.

Ltd, K. P. (2010). Data dimensions: disciplinary differences in research data sharing, reuse and long term viability. *Synthesis* (January), 1–36.

Maculan, B. C. M. dos Santos, and E. C. Aganette (2017). Desambiguação de relações em tesauros e o seu reúso em ontologias. *Ciência da Informação 46*.

Maria, A. and D. Tristão (2004). Sistema de classificação facetada e tesauros : instrumentos para organização do conhecimento . *Ciência da Informação*, 161–171.

Medeiros, F. (2014). *A historiografia medieval portuguesa na viragem do milénio: análise bibliométrica (2000-2010) e representação taxonómica*. Ph. D. thesis.

Rodrigues, E., R. Saraiva, C. Ribeiro, and E. M. Fernandes (2010). Os Repositórios De Dados Científicos: Estado da Arte.

Seiji, I. and I. I. Bittencourt (2015). *Dados abertos conectados*. Novatec, São Paulo.

Shakeri, S. and K. Gracy (2014, 01). A model for data curation research in small sciences: A model for data curation research in small sciences. *Proceedings of the American Society for Information Science and Technology 51*.

Simões, M. d. G. (2010). *A representação de Etnia e a sua evolução na Classificação Decimal Universal*. Ph. D. thesis, Coimbra.

Sobral, R. M. and C. A. C. M. dos Santos (2017). Repositórios institucionais digitais de informação científica: implementação com o software Dspace como solução técnica. *Prisma 35*, 152–184.

Souza, J. A. and O. Pestana (2017, 09). Elos interdisciplinares para estruturação semântica com vistas à organização da informação. *Páginas ab: arquivos e bibliotecas* (2017), 245–257.

Souza, R. F. D. (2006). ORGANIZAÇÃO E REPRESENTAÇÃO DE ÁREAS DO CONHECIMENTO EM CIÊNCIA E TECNOLOGIA : princípios de agregação em grandes áreas segundo diferentes contextos de produção e uso de informação ORGANIZATION AND REPRESENTATION OF KNOWLEDGE AREAS IN SCIENCE AND TECHNOLOGY : principles of aggregation in great areas. pp. 27–41.