U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

MESTRADO EM CIÊNCIA DA INFORMAÇÃO

Metadados para o uso de ferramentas de gestão de dados de investigação com investigadores do I3S

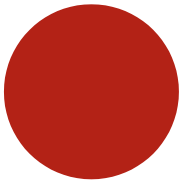
Marcelo Costa Sampaio

M

2019

UNIDADES ORGÂNICAS ENVOLVIDAS

FACULDADE DE ENGENHARIA
FACULDADE DE LETRAS



Marcelo Costa Sampaio

Metadados para o uso de ferramentas de gestão de dados de investigação com investigadores do I3S

Dissertação realizada no âmbito de
Mestrado em Ciência da Informação, orientada pela Professora Doutora Maria Cristina Alves Ribeiro e Coorientador João Daniel Aguiar Castro

Faculdade de Engenharia e Faculdade de Letras

Universidade do Porto

junho de 2019

Marcelo Costa Sampaio

Metadados para o uso de ferramentas de gestão de dados de investigação com investigadores do I3S

Dissertação realizada no âmbito de
Mestrado em Ciência da Informação, orientada pela Professora Doutora Maria Cristina Alves Ribeiro e Coorientador João Daniel Aguiar Castro

Membros do Júri

Presidente: Professor Doutor Gabriel David

Faculdade de Engenharia da Universidade do Porto

Arguente: Professora Doutora Irene Rodrigues

Universidade de Évora

Orientadora: Professor Doutora Cristina Ribeiro

Faculdade de Engenharia da Universidade do Porto

Agradecimentos

Na realização da presente dissertação, contei com o apoio de múltiplas pessoas e instituições às quais estou profundamente grato. Correndo o risco de injustamente não mencionar algum dos contributos quero deixar expresso os meus agradecimentos:

Ao orientador desta dissertação, a Professora Doutora Cristina Ribeiro, pela orientação prestada, pelo seu incentivo, disponibilidade e apoio que sempre demonstrou.

Ao coorientador João Castro pela disponibilidade, ajuda e grande apoio no desenvolvimento desta dissertação.

A todos os investigadores do I3S que disponibilizaram um pouco do seu tempo para colaborar na realização deste trabalho, em particular aos do grupo de Diversidade Genética por me terem arranjado um lugar para lá trabalhar.

A todos os meus colegas do InfoLab que contribuíram para tornar os dias mais alegres, em particular à Ana Ferreira por me ajudar no que fosse preciso neste projeto de dissertação e ao João Rocha pela ajuda incansável com o Dendro.

A todos os meus amigos, cada um com o seu jeito especial, que contribuíram para tornar estes últimos cinco anos numa experiência única. Para não correr o risco de não enumerar algum não vou identificar ninguém, aqueles a quem este agradecimento se dirige sabê-lo-ão.

Á minha mãe que sempre me apoiou nos momentos mais difíceis e de maior tensão, dando me carinho e ajuda, tornando esta batalha possível e conquistável.

A todos, um muito obrigado!

Este trabalho foi financiado pelo FEDER (Fundo Europeu de Desenvolvimento Regional) através do programa Operacional de Competividade e Internacionalização – COMPETE 2020 e por Fundos Nacionais através da FCT- Fundação para a Ciência e Tecnologia no projeto TAIL.

Resumo

Nas últimas décadas, a produção de dados de investigação tem vindo a crescer muito, principalmente devido ao desenvolvimento tecnológico que transformou todo o fluxo de trabalho dos investigadores. Esta situação cria desafios relativos às atividades de gestão dos dados de investigação, sobretudo ao nível da análise, armazenamento, preservação e partilha desses mesmos dados. A gestão de dados de investigação é essencial para a prática científica e existem bastantes intervenientes nas diferentes etapas deste processo – investigadores, agências de financiamento, universidades, curadores – que se preocupam com o valor dos dados produzidos. Torna-se também importante apoiar os investigadores com ferramentas que simplifiquem o trabalho necessário na gestão dos seus dados de investigação.

As ferramentas eletrónicas de gestão de dados de investigação são ferramentas importantes já que permitem aos investigadores cumprir os requisitos e criar uma ponte entre as diferentes etapas do fluxo da gestão de dados de investigação.

A adoção de uma ferramenta de gestão de dados de investigação pode contribuir para auxiliar a controlar o ciclo de vida dos dados já que é possível armazenar os dados e associar-lhes metadados de modo a torná-los FAIR – *Findable, Accessible, Interoperable, Reusable.* Além disso, a sua integração com repositórios de dados de investigação é também essencial na medida da indexação, preservação e disponibilização dos dados à comunidade científica.

Com o objetivo de apoiar os investigadores nas tarefas de gerir os seus dados de investigação, neste trabalho colabora-se com um grupo de investigadores do Instituto de Investigação e Inovação em Saúde (I3S) de modo a testar a plataforma Dendro, ferramenta desenvolvida na FEUP e INESC TEC, assim como para validar um modelo de metadados específico desenvolvido para os domínios dos investigadores.

Os resultados obtidos a partir do *feedback* dos investigadores demonstram que o modelo desenvolvido favorece um ponto de entrada fácil na descrição de dados, mas não impede os investigadores de apresentar limitações e identificar os seus requisitos específicos.

Palavras-Chave: gestão de dados de investigação, metadados, dados de investigação, LabTablet, Dendro, ontologias

Abstract

In the last decades, research data production has been growing consistently, mainly due to the technological development which has transformed the entire workflow of the researchers. This situation creates challenges regarding research data management activities, especially at the level of analysis, storage, preservation and sharing of research data. Research data management is essential to scientific practice and there are many stakeholders involved in the different stages of this process – researchers, funding agencies, universities, curators –who care about the value of produced research data. In this context, it is also important to support researchers with tools that simplify their work in managing data.

Electronic research management tools are important as they enable researchers to meet the RDM requirements and create a bridge between the different stages in the flow of research data management.

Their adoption can help control the data life cycle since it is possible to store the data and associate it with metadata in order to make it FAIR - Findable, Accessible, Interoperable, Reusable. In addition, its integration with research data repositories is also essential as it allows the indexing, preservation and availability of the data for the scientific community.

In order to support the researchers in their research data management tasks, in this this study we collaborate with a group of researchers from the Institute of Research and Innovation in Health (I3S) in order to test and evaluate Dendro platform, developed in FEUP and INESC TEC, as well as to validate a metadata descriptors model developed specifically for the researchers domain.

The results obtained from the researcher´s feedback demonstrate that the developed model favours an easy entry point in the data description tasks but does not prevent researchers from presenting its limitations and identifying their specific requirements.

Keywords: research data management, metadata, research data, LabTablet, Dendro, ontologies

Lista de figuras

Lista de tabelas

Lista de anexos

Siglas e abreviaturas

CERIF	Common European Research Format Ontology
CERN	European Organization for Nuclear Research
CKAN	Comprehensive Knowledge Archive Network
CODATA	Committee on Data for Science and Technology of the International Council for Science
DC	Dublin Core
DCMES	Dublin Core Metadata Element Set
DCMI	Dublin Core Metadata Initiative
DDI	Documentation Data Initiative
DwC	Darwin Core
EFO	Experimental Factor Ontology
ELN	Electronic Laboratory Notebook
EML	Ecology Metadata Language
FAIR	Findable, Accessible, Interoperable and Reusable
FEUP	Faculdade de Engenharia da Universidade do Porto
INFOLAB	Information Systems Research Group Laboratory
INESC TEC	Instituto de Engenharia de Sistemas e Computadores Tecnologia e Ciência
ISA-TAB	Investigation Study Assay Tabular
MESH	Medical Subject Headings
METS	Metadata Encoding and Transmission Standards
MIBBI	Minimum Information for Biological and Biomedical Investigations
MIBBIUP	Minimum Information for Biological and Biomedical Investigations – University of Porto
NCBO	National Center for Biomedical Ontologies
NISO	National Information Standards Organization
NSF	National Science Foundation
OAIS	Open Archival Information System
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OBI	Ontology for Biomedical Investigation
OBOE	The Extensible Observation Ontology
OECD	Organisation for Economic Co-operation and Development

PSI-MI CV	Proteomics Standards Initiative – Molecular Interactions Controlled Vocabulary
PSI-PAR CV	Proteomics Standards Initiative - Protein Affinity Reagent Controlled Vocabulary
RDA	Research Data Alliance
RDM	Research Data Management
SepCV	Sample Processing and Separations Controlled Vocabulary
XML	Extensible Markup Language
XSD	XML Schema Definition

Índice

1. Introdução

O contexto científico é cada vez mais impulsionado pela produção de dados de investigação de tal forma que a sua gestão está a tornar-se num requisito importante para todos os projetos de investigação. A ausência de boas práticas de gestão de dados de investigação pode levar a que os dados nunca realizem o seu potencial de reutilização e percam o seu valor ao longo do tempo. De modo a evitar este problema e garantir a transparência dos dados, as agências de financiamento têm exigido aos investigadores planos de gestão de dados no início dos seus projetos, onde é referida, por exemplo, a forma como os dados serão preservados e onde serão depositados e acessíveis após a finalização do projeto (Borgman 2012).

A gestão de dados de investigação é essencial para o desenvolvimento científico pelo que preocupa os atores envolvidos nas diferentes etapas do processo, desde os próprios investigadores, aos responsáveis por gestão de ciência, editores e agências de financiamento. Os investigadores têm interesse em ver os seus dados publicados, citados e reutilizados, os responsáveis por políticas de ciência pretendem garantir que os resultados obtidos com os seus planos de investimento têm o maior impacto possível. Além do mais, o investimento na gestão de dados de investigação e a consequente partilha e reutilização dos dados trazem bastantes vantagens ao evitar a duplicação de esforços, ao permitirem a verificabilidade dos dados, ao promover nova investigação, o que garante a inovação científica e contribui para a visão da "ciência aberta" (Van den Eynden et al. 2011).

O fluxo de trabalho na gestão de dados de investigação coloca questões tecnológicas e conceptuais (Ribeiro et. al. 2016). As questões tecnológicas dizem respeito ao armazenamento dos dados como, por exemplo, as interfaces usadas para descrever, organizar e depositar os dados. Por sua vez, os problemas conceptuais estão ligados à descrição dos dados. Os dados armazenados numa infraestrutura serão inúteis sem a organização e descrição que só os criadores podem fornecer e que permitirá a outros especialistas do domínio reutilizá-los. É neste sentido que se pretende apoiar os investigadores para serem uma parte ativa no fluxo da gestão dos seus dados de investigação, principalmente ao nível da descrição dos dados através da utilização de metadados.

No entanto, os investigadores nem sempre têm disponibilidade para estas tarefas pelo que a colaboração entre os vários atores envolvidos na gestão de dados de investigação é importante para o desenvolvimento de ferramentas ajustadas às necessidades dos domínios. É particularmente importante a colaboração entre os curadores de dados peritos na gestão da

informação e os investigadores com maior conhecimento dos domínios. Os curadores de dados estão mais conscientes das melhores práticas de metadados e podem ajudar os investigadores a promover a acessibilidade dos seus dados melhorando a qualidade dos metadados.

Nesta perspetiva, ao longo deste trabalho pretende-se colaborar com investigadores do instituto de Inovação e Investigação em Saúde (I3S) de modo a apoiá-los com a gestão dos seus dados de investigação. O principal foco deste trabalho é a descrição dos dados de investigação e por essa razão foi desenvolvido um modelo de metadados a partir de algumas normas já estabelecidas nos domínios das ciências biológicas. De seguida, pretendeu-se colocar os investigadores a descrever os seus dados a partir deste novo modelo, procurando obter feedback sobre os descritores. Simultaneamente, foi feita uma avaliação da usabilidade e utilidade da ferramenta Dendro como interface de organização e descrição de dados de investigação.

1.1 Motivação e definição de objetivos

As recentes diretrizes para a gestão de dados de investigação têm fomentado o surgimento de ferramentas orientadas aos investigadores. Desenvolvidas na Faculdade de Engenharia da Universidade do Porto e INESC TEC, as ferramentas Dendro e LabTablet são ferramentas de apoio à gestão de dados de investigação. No entanto, antes da generalização deste tipo de ferramentas, torna-se necessário avaliar a sua conformidade com as necessidades dos investigadores.

Neste trabalho pretende-se ter contacto com um grupo de investigadores do I3S de modo a perceber as suas necessidades e apoiá-los na gestão dos seus dados. Ao integrar as ferramentas de gestão de dados de investigação no fluxo de trabalho dos investigadores, sobretudo ao nível da descrição de dados a partir de metadados, é possível melhorar a qualidade, disponibilizar e preparar os conjuntos de dados de investigação para depósito e publicação.

Para atingir este objetivo central, as atividades a desenvolver são:

- **Identificação e análise de repositórios de dados de investigação e modelos de metadados utilizados nos domínios dos investigadores do I3S, sobretudo ao nível das ciências biológicas.** Para tal, recorreu-se a portais de repositórios como o re3data, de ontologias como o BioPortal e a diretórios de metadados como o do Research Data Aliance.
- **Seleção de um conjunto de descritores e respetiva representação através do desenvolvimento de uma ontologia** de modo a incluí-la no Dendro;
- **Ter contacto com um conjunto de investigadores de diferentes grupos de investigação do I3S** com o objetivo de observar o seu comportamento na descrição de

dados com as ferramentas de gestão de dados, assim como validar o conjunto de descritores e perceber possíveis necessidades e requisitos. Além disso, também se fizeram entrevistas relacionadas com as práticas de gestão de dados de investigação.

Simultaneamente, serão também anotadas as limitações e possíveis erros na interface que possam ocorrer durante a utilização do Dendro de modo a que sejam solucionados numa fase posterior.

Para uma visão mais pormenorizada, os objetivos deste trabalho foram organizados em níveis hierárquicos através da elaboração de uma árvore de objetivos (FIGURA 1). Assim, o sucesso de um objetivo e respetivas tarefas influenciam e contribuem para o sucesso dos seguintes de nível superior.

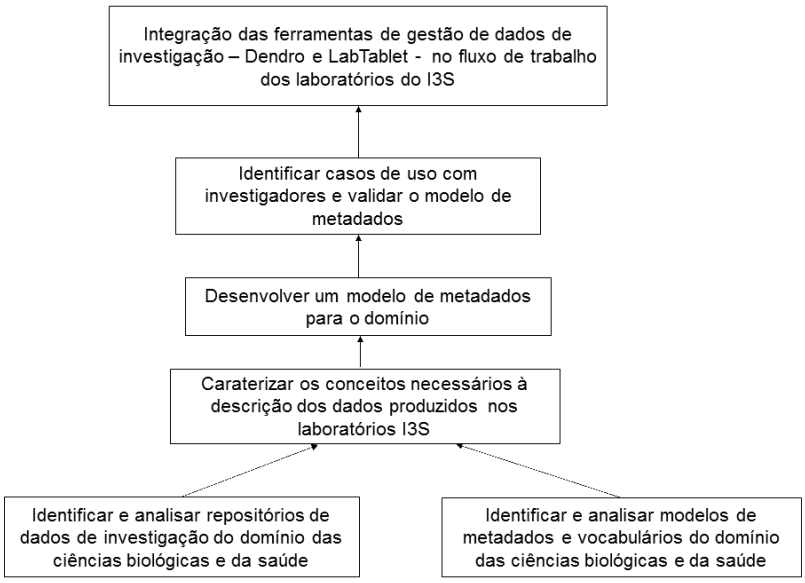


Figura 1 - Árvore de objetivos

1.2 Metodologia

A abordagem adotada para a realização desta dissertação foi o método qualitativo de investigação, nomeadamente a pesquisa exploratória sobre um caso de estudo. A pesquisa exploratória tem o intuito de familiarizar com o caso em estudo, promovendo uma investigação preliminar sobre os temas a abordar na dissertação para que seja possível refletir sobre estes e melhorar a sua compreensão. Numa primeira fase, realiza-se uma pesquisa de caráter exploratório, sobretudo ao nível dos repositórios, esquemas de metadados e ferramentas utilizadas neste âmbito. Visto este ser um caso de estudo sobre o Instituto de Investigação e

Inovação em Saúde (I3S), procura-se dar maior ênfase àquilo que já existe nestes domínios de investigação.

Relativamente à parte prática da dissertação, esta inicia-se com o desenvolvimento de uma ontologia, após uma seleção de conceitos considerados relevantes para a interpretação dos dados de investigação dos domínios de investigação do I3S. Esta ontologia será então testada e validada com um grupo de investigadores e modificada de acordo com os seus requisitos.

O processo de recolha de dados a partir das interações com os investigadores pode ser dividido em duas fases. A primeira correspondeu às primeiras semanas de presença no I3S nas quais se teve a oportunidade de conhecer os sete investigadores que colaboraram nas sessões. Estas sessões iniciais serviram para introduzir os investigadores ao conceito de metadados, assim como para apresentar os descritores selecionados a partir do modelo MIBBI. Nestas sessões foram também pedidos aos investigadores novos descritores considerados úteis para apoiar a interpretação dos seus dados.

Este projeto de dissertação foi realizado em simultâneo com outro projeto de dissertação denominado por "*Aplicação da ferramenta LabTablet em contexto laboratorial: Caso de estudo I3S*" por parte da Ana Luís Ferreira do Mestrado em Bioengenharia no ano letivo de 2018/2019. Por essa razão, nesta dissertação utilizaram-se os dados das respostas das entrevistas [ANEXO I] realizadas com os mesmos investigadores acerca das práticas de gestão de dados de investigação, sobretudo as respostas das perguntas relacionadas com a organização e descrição de dados. As entrevistas foram gravadas, transcritas e partilhadas e serviram como ponto de partida para conhecer o investigador.

Numa segunda fase e já com a ontologia importada para o Dendro, realizaram-se novas sessões com os investigadores de modo a que estes realizassem tarefas de descrição de dados de investigação através da plataforma Dendro. Nesta fase, apenas quatro investigadores tiveram disponibilidade para participar nas tarefas de descrição.

1.2.1 Participantes

Para a realização deste trabalho foi necessário estabelecer contactos e ter sessões presenciais com investigadores do Instituto de Inovação e Investigação em Saúde (I3S) de modo a atingir os objetivos propostos no início do trabalho. Após todos os contactos realizados, colaboraram, no total, sete investigadores de quatro grupos de investigação do I3S. Dentro deste grupo, realizaram-se sessões com três investigadores do grupo de Diversidade Genética, dois do grupo de Interações Epiteliais no Cancro, um investigador do grupo de Diferenciação e Cancro e, por fim, um investigador do grupo de Biologia Celular Glial.

1.3 Estrutura da dissertação

Esta dissertação está organizada da seguinte forma:

No primeiro capítulo, destinado à introdução, é feito um resumo geral sobre as práticas de gestão de dados de investigação, assim como se referem os principais objetivos e tarefas a levar a cabo durante a realização deste trabalho. Ainda neste capítulo, são referidos os métodos utilizados durante o desenvolvimento do trabalho, assim como os participantes deste estudo.

O segundo, terceiro e quarto capítulo são referentes à revisão de literatura e são considerados importantes para a correta contextualização dos temas relacionados com a dissertação a realizar. No segundo capítulo, refere-se a questão da gestão dos dados de investigação, salientando-se, o seu ciclo, assim como as diferentes tipologias de dados existentes. Ainda neste capítulo, abordam-se as infraestruturas para gestão de dados de investigação, sobretudo, os repositórios de dados de investigação. Primeiramente, exemplificam-se alguns repositórios de dados multidisciplinares tais como o Figshare e o Zenodo, e numa segunda fase, são apresentados repositórios de dados disciplinares das ciências biológicas. Esta fase foi importante de modo a compreender os domínios dos investigadores do I3S, assim como perceber o tipo de dados produzidos e os metadados necessários para a sua correta contextualização.

No terceiro capítulo, aborda-se a questão dos metadados, nomeadamente ao nível da sua importância para os dados de investigação. Os metadados contribuem para dar contexto aos dados de investigação, permitindo uma maior facilidade de recuperação dos dados num determinado contexto e possibilitam a reutilização dos dados a nível futuro. Ainda neste capítulo, é efetuada a apresentação de alguns modelos de metadados desenvolvidos por comunidades de diferentes domínios de investigação para facilitar a documentação, troca, preservação e reutilização dos dados. São exemplo o Darwin Core para a biodiversidade, o Data Documentation Initiative para as ciências sociais e comportamentais e o Ecological Metadata Language para a ecologia. No entanto, existem também esquemas mais genéricos, isto é, que pretendem descrever qualquer objeto de informação, tais como o Dublin Core e o Metadata Encoding and Transmission Standards. Para além destes, são apresentados dois *standards* de metadados, definidos pela Research Data Alliance, utilizados no âmbito do domínio das ciências da vida, mais especificamente ao nível das disciplinas da biologia celular, genética, bioquímica, bioengenharia, etc. O Investigation Study Assay Tabular Format e o Minimum Information for Biological and Biomedical Investigations são normas que pretendem garantir que todos os dados de investigação destes domínios sejam interpretados a partir da utilização de metadados associados.

Dada a importância dos cadernos de laboratório como ferramentas de apoio ao trabalho dos investigadores, no último capítulo da revisão de literatura, são referidos exemplos de cadernos de laboratório procedendo-se a uma divisão entre aqueles considerados como aplicações genéricas de *note-taking* e aqueles com maior capacidade de personalização e maiores funcionalidades para a gestão de dados de investigação. Por fim, elabora-se uma tabela comparativa ao nível das caraterísticas entre os diferentes cadernos de laboratório da amostra selecionada.

O capítulo cinco é referente à ontologia desenvolvida para este caso de estudo. Assim é descrito o procedimento de desenvolvimento desde a seleção e análise de modelos de metadados inicial, passando pela sua formalização no Protégé até à sua exportação para a plataforma Dendro. Neste capítulo, também é feita uma apresentação sumária do Dendro, ferramenta utilizada para as tarefas de descrição com os investigadores participantes.

O capítulo seis é referente aos resultados obtidos das sessões presenciais com os investigadores participantes. Desta forma, ao longo deste capítulo é mostrado o feedback dos investigadores em relação ao conjunto de metadados apresentado numa primeira fase e de seguida, o resultado das descrições dos investigadores através da utilização da ferramenta Dendro. Ainda neste capítulo, discutem-se os resultados obtidos a partir da comparação dos resultados entre os investigadores.

Esta dissertação termina com uma reflexão global à experiência realizada, num capítulo dedicado às conclusões, limitações e investigação futura.

2. Gestão de dados de investigação

Com a crescente produção de dados de investigação em todas as áreas científicas, torna-se necessário ter uma gestão eficaz dos dados de investigação de modo a responder a todos os novos desafios. Whyte e Tedds (2011) definem a gestão de dados de investigação (designação internacional: Research Data Management, ou RDM) como um processo que se inicia com a organização dos dados desde a sua entrada no ciclo de investigação até à disseminação e preservação dos novos resultados, pretendendo-se assegurar uma verificação confiável dos resultados e que uma nova investigação seja construída sobre o que já existe.

O ciclo de vida dos dados engloba um conjunto de atividades que começa com um plano de gestão de dados de investigação; recolha dos dados como parte da investigação; pela identificação, descrição e processamento do conjunto de dados até à preservação e partilha dos dados num repositório de dados apropriado (Borgman 2012). O plano de gestão de dados de investigação, muitas vezes requerido pelas agências de financiamento, indica, por exemplo, onde e como os dados serão depositados, preservados e acessíveis após a conclusão do projeto. Por sua vez, a recolha de dados diz respeito à produção de dados durante toda a fase de investigação através da utilização de métodos científicos e instrumentos tecnológicos com o objetivo de recolher dados para posterior análise e tratamento. A identificação e descrição dos dados assumem uma natureza crítica no processo da gestão de dados de investigação já que permitem que outros investigadores localizem e reutilizem os dados após estes terem sido publicados. É por isso importante que os investigadores descrevam os seus dados durante o processo de investigação, evitando o adiamento para o final do projeto. No final, recomenda-se que os dados sejam partilhados e publicados nos repositórios adequados de modo a contribuir para a visão da "ciência aberta" (designação internacional: *"Open Science")*. A ciência aberta tem como objetivo tornar os dados e resultados e dados da investigação científica acessíveis à comunidade científica pelo que assenta também no princípio dos dados abertos ("*Open Data*").

Figura 2- Ciclo de vida dos dados de investigação (Universidade de Otava)

A ênfase na partilha dos dados de investigação contribui para debates científicos, promove a inovação e a transparência, a análise dos resultados, a validação dos métodos, evita a duplicação de esforços, promove a visibilidade e a colaboração entre os utilizadores e os produtores dos dados (Van den Eynden et al. 2011 e Borgman 2012). No entanto, de forma a que os dados possam ser partilhados e reutilizados é necessário que existam organizações e indivíduos responsáveis pela sua curadoria, já que por si só, os investigadores não são as pessoas indicadas para assegurar a preservação e o acesso contínuo aos dados que recolhem e produzem uma vez que em geral não são responsáveis por infraestruturas e dão prioridade à produção dos próprios dados.

São diversos os atores envolvidos no processo da gestão dos dados de investigação (Lyon 2007). Esses atores são os próprios investigadores, as instituições responsáveis pelo projeto, os curadores, as agências de financiamento e os utilizadores e cabe a cada um desempenhar o seu papel. É da responsabilidade dos investigadores, bem como da instituição e do curador de dados, trabalhar os dados resultantes de forma a garantir que podem ser usados por outros. Durante esta fase, os investigadores revelam-se fundamentais na descrição de dados, já que, mesmo não tendo muito conhecimento sobre práticas de gestão de dados, têm conhecimento sobre o domínio dos dados e são capazes de produzir metadados que permitem a sua interpretação pela comunidade. Simultaneamente, os curadores de dados, como peritos na gestão de dados, podem complementar o trabalho dos investigadores. Por outro lado, as instituições também têm interesse em ter os seus dados reconhecidos e preservados nos repositórios em conformidade com os requisitos das agências de financiamento pelo que devem incentivar os investigadores a preocuparem-se com a gestão dos dados.

2.1 Dados de investigação

A produção de dados é uma das caraterísticas da ciência moderna e a forma e o volume desses registos ou dados científicos foram naturalmente evoluindo, crescendo em dimensão e complexidade, de acordo com a própria evolução da investigação científica, dos seus objetos, metodologias e instrumentos.

Os dados de investigação são definidos como "registos factuais usados como fontes primárias na investigação científica, e que são geralmente aceites na comunidade científica como necessários para validar os resultados de investigação" (OECD 2007). Por outras palavras, dados de investigação, contrariamente a outros tipos de informação, podem ser recuperados, observados ou produzidos com o objetivo de análise e produção de resultados originais (Rice 2009). A Comissão Europeia (2017) define os dados de investigação como factos ou números, recolhidos com o objetivo de serem analisados e considerados como base para o desenvolvimento da fundamentação, discussão ou cálculo de investigação científica.

Os dados de investigação são heterogéneos já que podem assumir diversas formas dependendo da sua origem, do problema de investigação a ser tratado e do domínio e disciplina do investigador (Kennan & Markauskaite 2015). Em 2005, a National Science Foundation definiu uma classificação para os dados, dependendo da sua origem:

- **Observacionais:** Este tipo de dados é recolhido a partir da observação de um determinado comportamento ou atividade única em tempo real e por isso devem ser preservados indefinidamente (Lide 1981). Dados capturados a partir de sensores tais como de temperatura, humidade ou precipitação são exemplos de dados observacionais;

- **Experimentais:** dados produzidos a partir da intervenção ativa do investigador para produzir e medir mudanças ou criar diferenças quando uma variável é alterada. Podem ser reproduzidos, embora possam existir casos nos quais as condições experimentais ou variáveis sejam desconhecidas ou os custos bastante altos, influenciado assim o processo de validação e a sua caraterização como dados validados ou não validados (Qin, Ball, & Greenberg, 2012). Os resultados de experiências de laboratório dos domínios da biologia ou química podem ser classificados como dados experimentais;

- **Computacionais:** dados produzidos a partir da execução de modelos computacionais ou de simulações imitando uma operação de um processo ou sistema do mundo real. Este método é usado para estimar a evolução de sistemas sob certas condições. Estes dados computacionais podem ser reproduzidos e verificados desde que exista informação suficiente sobre o processo utilizado na sua produção (Lide 1981).

O glossário CODATA, referido por (Willis, Greenberg, and White 2012) distingue ainda os dados de acordo com o seu nível de processamento, entre dados em estado bruto, isto é, na sua forma original e dados processados que já foram sujeitos a um processo de manipulação.

2.2 Repositórios de dados de investigação

A gestão de dados de investigação tem vindo a tornar-se, cada vez mais, num problema e preocupação para os investigadores pelo que as instituições têm a necessidade de lhes fornecer plataformas para apoiar a organização e gestão dos dados e prepará-los para publicação. As principais preocupações dos investigadores resultam das cada vez maiores pressões que existem sobre as práticas de gestão e partilha dos dados. Estas pressões vêm, sobretudo, das agências de financiamento que buscam agregar valor a projetos científicos dispendiosos, dos editores que procuram responder a pedidos de transparência e reprodutibilidade dos objetos científicos, das organizações e instituições que procuram gerir dados valiosos e da comunidade científica e público que procuram aceder e reutilizar os dados para os seus próprios objetivos (Borgman, 2012).

Algumas destas instituições investem em repositórios institucionais como suporte para o depósito de dados, enquanto outras têm vindo a utilizar ferramentas em ambientes mais ricos a nível de descrição de dados (Amorim et al. 2017). Além do mais, os repositórios de dados de investigação são ferramentas eficientes para a organização, preservação e partilha de dados, no entanto necessitam de estar bem estabelecidos numa comunidade e dispor de ferramentas capazes de descrever e divulgar os dados de tal forma que estes possam ser acedidos e reutilizados (Walport e Brest 2011).

Devido à crescente produção dos dados de investigação, existem soluções desenvolvidas tanto por comunidades de código aberto, quer por organizações relacionadas diretamente com a gestão de dados. Consequentemente, a escolha de uma plataforma adequada às necessidades e requisitos de cada comunidade pode ser uma tarefa difícil devido à grande variedade de alternativas existentes. Algumas destas plataformas seguem duas abordagens distintas, nomeadamente na descrição dos objetos científicos e os esquemas de metadados, existindo a divisão entre esquemas de metadados específicos de domínios e esquemas de metadados genéricos, sobretudo para as publicações científicas.

Neste sentido, a descrição de dados tem de ser suficientemente detalhada para estes serem bem interpretados. Os seus requisitos podem variar de domínio para domínio levando a que os repositórios necessitem de estar preparados e serem flexíveis para representar estes objetos. Há alguns aspetos considerados relevantes e a levar em conta quando se comparam as diversas plataformas existentes, nomeadamente ao nível do seu licenciamento, controlo sobre os dados,

conformidade com o modelo OAIS e o OAI-PMH, assim como com o conceito da "ciência aberta".

Ao nível do licenciamento das plataformas, estas podem ser código aberto ou então serem de tipo proprietário. As plataformas de código aberto apresentam muitas vantagens, principalmente ao nível da manutenção, desenvolvimento e atualizações, assim como ao nível da compatibilidade com normativos de metadados e outros sistemas.

Para um sistema estar em conformidade com o modelo do protocolo OAIS, necessita de representar os dados em três tipos de pacotes – o de submissão (SIP) relativo à fase de ingestão e depósito no repositório dos dados e metadados, o de arquivo (AIP) dentro do repositório e o de disseminação (DIP) relativo à representação e acesso por parte dos utilizadores (Consultative Committee for Space Data Systems, 2012).

Por outro lado, o protocolo OAI-PMH[1] define um mecanismo para a troca e partilha de metadados entre repositórios com o objetivo de permitir a sua interoperabilidade. Este protocolo define alguns conceitos de modo a modelar o fluxo de informação, nomeadamente o de *harvester*, repositório, item, formato e identificadores. Tendo por base estas caraterísticas, pode fazer-se uma comparação entre as plataformas de repositórios de dados mais usadas:

- **Zenodo:** O Zenodo[2] é um repositório multidisciplinar *online*, lançado em 2013, desenvolvido pelo CERN através da iniciativa da União Europeia OpenAIREplus. Este repositório permite que investigadores, cientistas, projetos e instituições da União Europeia partilhem e contribuam com resultados de projetos científicos, quer sejam conjuntos de dados ou publicações. Além do mais, não impõe restrições de formatos, incluindo texto, áudio, vídeos e imagens, e aceita conteúdos de qualquer domínio de investigação. As suas principais vantagens são sobretudo a conformidade com o protocolo OAI-PMH, compatibilidade na exportação de metadados através de Dublin Core, MARC e MARCXML, assim como a não necessidade de manutenção e a capacidade de exportação de referências para o BibTeX, DataCite, DC, EndNote, NLM e RefWorks.

- **Figshare:** O FigShare[3] é um repositório multidisciplinar comercial *online,* lançado em 2011, no qual os investigadores podem partilhar os resultados dos seus projetos, incluindo imagens, conjunto de dados ou vídeos. O FigShare pretende que as

[1] http://www.openarchives.org/OAI/openarchivesprotocol.html
[2] https://zenodo.org/
[3] https://figshare.com/

publicações de todos os resultados de investigação sejam fáceis de partilhar, citar e recuperar. As suas principais vantagens são a capacidade de exportação de referências para programas tais como Mendeley e EndNote, a não necessidade de manutenção e a associação de créditos aos autores através de citações e referências.

- **CKAN:** O CKAN é uma plataforma de gestão de dados de código aberto baseado na *web* para o armazenamento e distribuição de dados abertos. Sendo desenvolvido desde 2014 pela Open Knowledge International, o CKAN é uma plataforma que funciona como um sistema de catalogação de dados e tem vindo a ser usado por bastantes instituições públicas que têm como objetivo publicar e partilhar os seus dados. A principal vantagem é a licença de código-aberto que permite a personalização das caraterísticas por parte de uma comunidade.

2.2.1 Repositórios de dados de investigação no domínio das ciências biológicas

Os repositórios de dados de investigação online caraterizam-se como sistemas configurados e especializados para a gestão, partilha, acesso e preservação de grandes conjuntos de dados de investigação. Estes repositórios podem, porém, ser especializados na gestão de dados de um domínio de investigação ou então ser multidisciplinares e agregar dados de diferentes disciplinas, como são exemplo o FigShare e o Zenodo.

De modo a encontrar repositórios por domínios, os portais ou diretórios de repositórios são fundamentais. O mais utilizado é o re3data[4], lançado em 2013, que funciona como um registo global de repositórios de dados de investigação das mais variadas disciplinas. É possível pesquisar por disciplina, tipo de conteúdo e país do repositório, sendo que posteriormente, também é possível restringir e filtrar os resultados de acordo com alguns parâmetros. Além deste, existem também o OpenDOAR[5] a nível global e o FAIRsharing[6], anteriormente BioSharing, como portal para as ciências da vida, abrangendo as ciências biológicas, ambientais e biomédicas. Na Tabela 1, são listados exemplos de alguns dos repositórios disciplinares das ciências biológicas com maior impacto atualmente. Para todos eles existem guiões para o depósito de dados. Alguns foram avaliados e são recomendados pela revista

[4] https://www.re3data.org/ Consultado a 14/12/2018
[5] http://v2.sherpa.ac.uk/opendoar/
[6] https://fairsharing.org/

Nature de acordo com os princípios e requisitos de acesso, preservação e estabilidade dos dados[7].

Tabela 1 - Repositórios de dados de investigação no domínio das ciências biológicas

Repositório	Disciplinas	Região	Fonte	Dimensão[8]
Gene Expression Omnibus (GEO)	Genética	Estados Unidos da América	https://www.ncbi.nlm.nih.gov/geo/	4348 datasets
ArrayExpress	Medicina, Microbiologia, Virologia e Imunologia, genética	União Europeia e Reino Unido	https://www.ebi.ac.uk/arrayexpress/	71647 datasets
PubChem	Bioquímica	Estados Unidos da América	https://pubchem.ncbi.nlm.nih.gov/	97 milhões de compostos 249 milhões de substâncias 1 milhão de bioensaios
BioModels Database	Investigação biológica e médica; Genética,	União Europeia	http://www.ebi.ac.uk/biomodels-main/	8000 modelos
NeuroMorpho.org	Neurobiologia, Medicina, Zoologia	Estados Unidos da América	www.NeuroMorpho.Org	101 503 reconstruções neurobiológicas
BIGG Database	Bioquímica, Genética	Estados Unidos da América	http://bigg.ucsd.edu/	85 modelos, 7339 metabolitos, 24311 reações químicas
Mass Spectrometry Virtual Interactive Environment (MassiVE)	Bioquímica	Estados Unidos da América	https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp	9004 datasets
Metabolights (MTBLS)	Medicina; Metabolismos, Bioquímica e Genética de Microrganismos	União Europeia e Estados Unidos da América	https://www.ebi.ac.uk/metabolights/	444 datasets
FlowRepository	Biologia Celular	Canadá	http://flowrepository.org/	700 datasets
Database of Genomic Variants Archive (DGVA)	Genética, Genómica, Biomédicas	União Europeia	https://www.ebi.ac.uk/dgva	206 datasets

[7] https://www.nature.com/sdata/policies/repositories#life
[8] Dados recolhidos no dia 4 de janeiro de 2019

3. Metadados para dados de investigação

É necessário estabelecer medidas de gestão e documentação dos dados de investigação de modo a que estes sejam facilmente acessíveis pela comunidade de interesse, reutilizados e preservados a longo prazo. Além disso, de modo a tornar os dados utilizáveis, é necessário preservar toda a documentação adequada relativa ao conteúdo, estrutura, contexto e fonte da recolha dos dados.

Definidos como "dados sobre dados" (NISO 2017), os metadados assumem-se como sendo bastante importantes para a representação e descrição de dados, assim como uma componente essencial para a comunicação científica atual (Willis, Greenberg, and White 2012). Deste modo, a utilização de metadados associados ao conjunto de dados permite uma maior facilidade de recuperação dos dados por parte dos agentes interessados, a sua contextualização, e consequentemente melhor interpretação, assim como a possibilidade da reutilização dos dados no futuro. Independentemente do uso a que se destinam, sejam eles sobre o conteúdo ou contexto, os metadados podem ser classificados em diferentes categorias (NISO, 2017):

- **Descritivos:** metadados utilizados na identificação e na descrição de objetos de informação, tendo como principal objetivo orientar o utilizador para a sua pesquisa e recuperação. Contêm informação como, por exemplo, o nome do autor, título, assunto ou palavras-chave.
- **Estruturais:** metadados que descrevem os objetos digitais ao nível da sua estrutura física e lógica, permitindo a navegação sobre eles e a sua apresentação. Informação sobre a paginação, índice e ordem dos capítulos são exemplo deste tipo de metadados.
- **Administrativos:** metadados que incluem informação necessária para gerir um objeto, incluindo, por exemplo, a forma como foi criado (metodologia ou *software* utilizado) ou quem tem direitos de acesso. Estes metadados podem ser divididos em três subcategorias:
    - **Direitos de acesso:** metadados relacionados com a documentação dos direitos de propriedade intelectual.
    - **Preservação:** metadados necessários para a preservação de um objeto digital de modo a que este continue acessível ao longo do tempo. Descrevem a informação de arquivo e os processos de preservação efetuados sobre o objeto (p.ex. migrações de formatos)

- **Técnicos:** metadados relacionados com a informação necessária para armazenar os dados (tipo e tamanho do ficheiro, data de criação).

3.1 Requisitos para modelo de metadados

Os requisitos na descrição de dados variam de domínio para domínio dependendo da cultura do grupo de investigação, competências na gestão de dados ou imposições das agências de financiamento. De acordo com Qin, Ball & Greenberg (2012), uma investigação levada a cabo num ambiente digital necessita que os dados de investigação tenham as propriedades da "E-Science", nomeadamente serem verificáveis, interoperáveis, analisáveis e interfuncionais. Além disso, os dados devem ser também seguir os princípios FAIR - *Findable, Accessible, Interoperable and Reusable (*Mark D. Wilkinson et al, 2016). As propriedades dos dados de investigação podem, então, ser transformadas em requisitos.

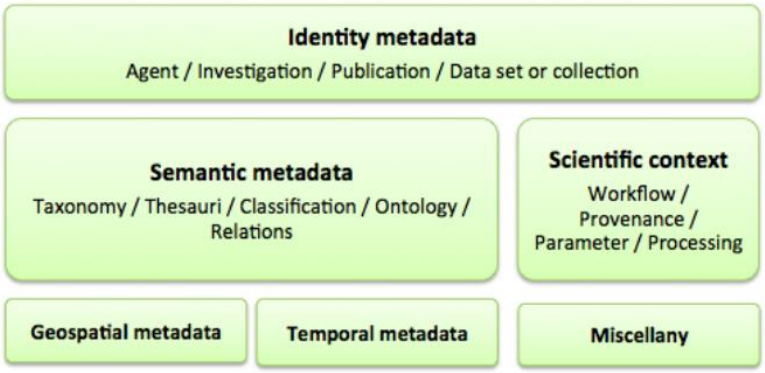


Figura 3 - Arquitetura dos requisitos de metadados (Qin, Ball & Greenberg, 2012)

A "visão arquitetónica" (Qin, Ball & Greenberg, 2012) dos metadados descreve os atributos dos metadados como blocos que formam uma representação abrangente dos dados ou objetos de informação.

O primeiro nível, no topo, inclui identificadores para as entidades ligadas à investigação ou estudo, tais como os investigadores, organizações, lugar, evento ou objeto. Cada uma destas entidades tem um conjunto de metadados a si associados, como por exemplo, um investigador tem um nome, laboratório, afiliação e informação de contacto e um conjunto de dados tem um nome, descrição, lugar, entre outros atributos. No segundo nível, os metadados semânticos são utilizados, principalmente, como identificadores de assunto para os dados ou então como mecanismo da ligação entre conjunto de dados com conteúdo e assuntos semelhantes. Para este efeito, ferramentas semânticas tais como as taxonomias, tesauros, classificações universais ou ontologias são fundamentais para permitir uma maior flexibilidade na

representação dos dados. Por fim, o contexto científico dos dados e os metadados temporais e geográficos ajudam a preencher os requisitos relativos à verificabilidade, confiabilidade e reprodutibilidade dos dados.

3.2 Esquemas de metadados

Existem diversos de esquemas de metadados distinguindo-se pelo tipo de dados e recursos que pretendem descrever. Um esquema de metadados é "*um plano que mostra as relações entre os descritores de metadados, normalmente através do estabelecimento de regras para o seu uso e gestão no que diz respeito à sua semântica, sintaxe e opcionalidade dos valores*[9]". De uma forma geral, isto significa que um esquema de metadados define como estes devem ser estruturados, por exemplo através da utilização de um XSD (define as regras de validação de documentos em formatos XML), e atribui a cada elemento um significado ou uma relação com outros elementos.

Os esquemas de metadados, dependendo do seu objetivo, podem ser classificados como multidisciplinares ou como específicos a um domínio de investigação. Os esquemas mais genéricos foram desenvolvidos, maioritariamente, para a descrição bibliográfica, e são facilmente compreendidos pelos utilizadores pois não necessitam de um conhecimento sobre um domínio de investigação. Por outro lado, os esquemas necessários para a descrição de dados de um determinado domínio de investigação têm de ser mais específicos com o objetivo de satisfazer as necessidades dessa comunidade (Silva 2016).

Willis, Greenberg e White (2012), num estudo de análise a uma amostra de esquemas de metadados de diferentes domínios e de descrição de diferentes tipos de dados científicos, concluíram que, independentemente dos dados a serem descritos ou do domínio de investigação, os esquemas devem obedecer a onze requisitos e objetivos fundamentais para a correta documentação dos dados. Os onze requisitos apresentados são:

- **Abstração do esquema:** A abstração de um esquema permite que as necessidades sejam capturadas de uma forma que suporta múltiplas apresentações ao longo do tempo.
- **Extensibilidade, flexibilidade e modularidade do esquema:** Estes requisitos assegurarão a longevidade do esquema de metadados, facilitando a adoção, modificação ao longo do tempo e extensão às necessidades a serem identificadas.

[9] ISO 23081.1 s3 Terms and Definitions

- **Abrangência e suficiência:** Necessidade de definir um conjunto de elementos (ou vocabulário) que seja abrangente, mas também identificar um conjunto mínimo de elementos essenciais para a documentação dentro de um domínio de investigação.
- **Simplicidade:** Ser simples e fácil de utilizar.
- **Troca dos dados:** Ter a capacidade de permitir a troca, partilha e comunicação dos dados entre os membros da comunidade.
- **Recuperação dos dados:** O esquema de metadados deve facilitar a pesquisa, descoberta e recuperação dos dados, levando em consideração os caminhos de acesso específicos do domínio.
- **Preservação dos dados:** Facilitar a preservação e documentação dos dados.
- **Publicação dos dados:** Facilitar e apoiar a publicação dos dados em revistas científicas ou repositórios.

Em suma, os metadados para a representação e descrição de dados de investigação são uma componente essencial na gestão de dados de investigação. Nas últimas décadas, comunidades de diferentes domínios de investigação têm desenvolvido esquemas de metadados para facilitar a documentação, troca, preservação e reutilização dos dados sendo que muitos deles estão associados a repositórios de dados disciplinares. Posto isto, a compreensão dos diferentes domínios e dos tipos de dados revela-se fundamental para a pesquisa dos esquemas de metadados desenvolvidos.

3.2.1 Dublin Core (DC)

Desenvolvido nos meados da década de 1990 como resultado de uma colaboração internacional, o modelo de metadados Dublin Core é um exemplo de um esquema de metadados genérico, sendo normalizado pela ISO 15836:2009[10], mais tarde revisto pela ISO 15836-1:2017[11]. O Dublin Core Metadata Element Set (DCMES)[12] é um vocabulário com quinze propriedades a serem usadas na descrição de um objeto. O nome Dublin deve-se à sua origem em Dublin, Ohio nos Estado Unidos da América e "*Core*" ao facto de os elementos serem amplos e genéricos, podendo ser utilizados na descrição num vasto conjunto de recursos, sejam eles em suporte físico ou digital. No seu formato mais simples, os quinze elementos base denominados por DCMES são fáceis de compreender e implementar, sendo eles:

- **Title:** nome dado a um recurso;

[10] https://www.iso.org/standard/52142.html
[11] https://www.iso.org/standard/71339.html
[12] http://dublincore.org/documents/dces/ Consultado a 20/12/2018

- **Creator:** Entidade responsável pela produção do recurso;
- **Date:** Data da publicação do recurso;
- **Contributor:** Entidade responsável por contribuir para a produção do recurso;
- **Subject:** Tópico ou palavras-chave do recurso;
- **Language:** Idioma padrão do recurso;
- **Type:** Natureza do recurso;
- **Source:** Recurso do qual o recurso descrito é derivado;
- **Description:** Descrição do recurso;
- **Format:** Formato digital ou físico do recurso;
- **Identifier:** Referência inequívoca do recurso num determinado contexto;
- **Coverage:** localização espacial, temporal ou de jurisdição;
- **Publisher:** Entidade responsável pela disponibilidade do recurso;
- **Relation:** Relações com outros recursos;
- **Rights:** descrição sobre os direitos mantidos sobre o recurso;

A simplicidade de utilização deste esquema pode, simultaneamente, ser a sua maior riqueza e fraqueza, já que alguns descritores são ambíguos e os metadados que podem ser representados através deste esquema são limitados, nomeadamente ao nível de descrição de *datasets* (Silva, 2016). De modo a evitar a falta de pormenor, a DCMI criou uma versão de elementos chamada DCTERMS[13], composta pelos clássicos quinze elementos do DCMES mais quarenta novos elementos. Para estes elementos, a DCMI definiu também classes de elementos que podem ser usadas como domínios e valores (ex. is*PartOf* ou *HasFormat* que refina o termo *Relation)*.

3.2.2 Metadata Encoding and Transmission Standards (METS)

O esquema METS[14] é um esquema normalizado definido pela Biblioteca do Congresso dos Estados Unidos para a codificação de metadados descritivos, administrativos e estruturais referentes a objetos de uma biblioteca digital. O esquema permite relacionar tipos de metadados e diferentes elementos presentes noutros esquemas de metadados, como por exemplo, do Dublin Core ou do Metadata Object Description Schema (MODS).

[13] http://dublincore.org/documents/dcmi-terms/ Consultado 20/12/2018
[14] http://www.loc.gov/standards/mets/mets.xsd Consultado a 26/12/2018

3.2.3 Darwin Core (DwC)

O esquema Darwin Core (DwC)[15] é um exemplo de um esquema desenvolvido para um domínio de investigação específico. Este esquema inclui um glossário de termos destinados a facilitar a troca de informação acerca da biodiversidade através de identificadores, etiquetas e definições. O Darwin Core adota uma abordagem semelhante à do esquema Dublin Core ao domínio da biodiversidade e baseia-se, principalmente em ocorrências na natureza conforme documentado através de observações, espécimes, amostras e outra informação relacionada.

3.2.4 Data Documentation Initiative (DDI)

O esquema DDI[16] é um esquema internacional normalizado baseado em XML para a descrição, apresentação e troca de dados produzidos através de estudos e outros métodos de observação nas ciências sociais, comportamentais, económicas e da saúde, permitindo de igual forma documentar e gerir diferentes fases do ciclo de vida dos dados de investigação.
O esquema aborda as restrições relacionadas com a distribuição de metadados nas ciências sociais, procurando colmatar a falta de diretrizes e formatos estabelecidos neste domínio.

3.2.5 Ecological Metadata Language (EML)

O esquema EML[17] é um esquema normalizado de metadados desenvolvido pela Ecological Society of America para a documentação de dados do domínio da ecologia. O EML é implementado através de um conjunto de documentos XML que permitem descrever os dados a vários níveis, seja de forma modular com as estruturas definidas ou de modo extensível possibilitando a introdução de novos metadados. Assim, o seu maior propósito é fornecer à comunidade do domínio da ecologia um esquema extensível e flexível para o uso na análise e tratamento de dados.

3.3 Esquemas de metadados utilizados no domínio das ciências biológicas

Os esquemas de metadados padrão mais utilizados no domínio da biologia, assim como nos seus diversos campos tais como a bioquímica, bioengenharia ou biologia celular, são o ISA-TAB e as listas do projeto MIBBI.

---

[15] https://dwc.tdwg.org/terms/ Consultado a 26/12/2018
[16] http://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/ Consultado a 26/12/2018
[17] https://knb.ecoinformatics.org/external//emlparser/docs/eml-2.1.1/index.html Consultado a 26/12/2018

3.3.1 Investigation Study Assay Tabular (ISA-TAB)

A primeira versão do esquema ISA-TAB[18] foi desenvolvida por uma equipa da Universidade de Oxford e lançado em 2008, sendo a versão mais recente de 2016. O ISA-TAB é um esquema de metadados desenvolvido de modo a possibilitar a gestão de um conjunto de experiências no âmbito das ciências da vida, ambientais e biomédicas. Foi construído em torno dos conceitos de *Investigation (*o contexto do projeto), *Study* (unidade de investigação) e *Assay (*medições de análise) e permite a descrição rica dos dados, através de metadados experimentais (caraterísticas da amostra, tecnologia e tipos de medição, entre outras) com o objetivo de permitir que estes e as descobertas resultantes possam ser reproduzíveis e reutilizáveis.

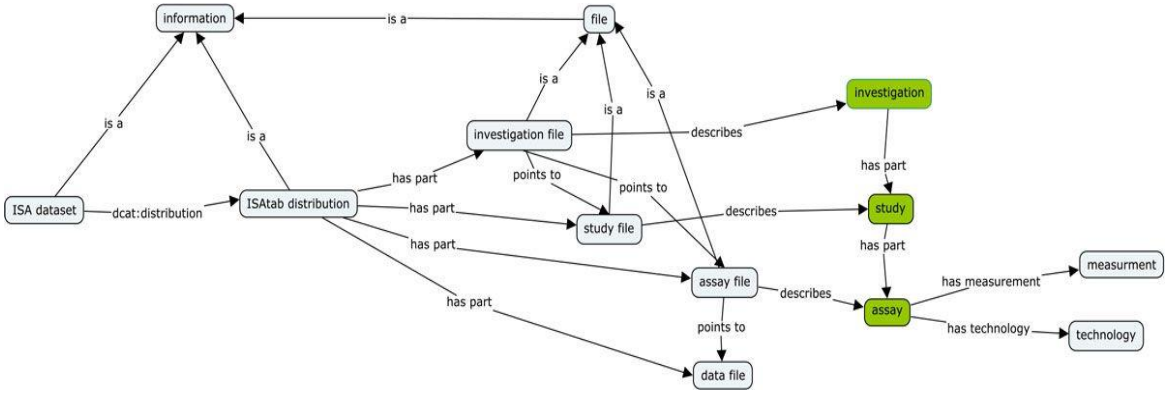


Figura 4 - Formato da terminologia ISA-TAB

Autoria: González-Beltrán et al (2014)

Tal como se pode observar na Figura 4, o modelo ISA-TAB consiste nas três entidades mencionadas, na qual a investigação contém toda a informação necessária para compreender os objetivos e métodos utilizados numa experiência, enquanto que os passos experimentais são descritos no estudo e no ensaio. Alguns dos elementos fundamentais da entidade de investigação são a sua descrição, título, identificador, assim como as publicações e contactos ligados à investigação. No estudo e no ensaio, devem ser registados, com base em ontologias desenvolvidas na área, o tipo de experiência, o nome e a classificação do fator experimental utilizado, o tipo de medição e a tecnologia utilizada.

[18] https://isa-specs.readthedocs.io/en/latest/index.html Consultado a 28/12/2018

3.3.2 Minimum Information for Biological and Biomedical Investigations (MIBBI)

O MIBBI[19] é um projeto que fornece recursos e descreve listas de verificação de informação mínimas que os dados e metadados que descrevem uma determinada experiência do domínio das ciências biológicas devem cumprir. O projeto MIBBI funciona como um portal para um grupo de quase 40 listas de verificação de informação mínima para a documentação de várias disciplinas biológicas. Estas listas têm vindo a ser desenvolvidas nas últimas décadas e a ser incorporadas e registadas na coleção da MIBBI. Na seguinte tabela, estão representados alguns nomes de esquemas de requisitos mínimos de metadados para ciências biológicas, assim como o domínio e subdomínio a que pertencem, os vocabulários ou ontologias que utilizam, a sua fonte e casos de uso onde são implementados.

[19] https://fairsharing.org/collection/MIBBI Consultado a 28/12/2018

Tabela 2 - Amostra de esquemas do projeto MIBBI

Esquema	Domínio	Uso de Vocabulários Controlados e/ou Ontologias	Fonte	Casos de uso
Minimum Information About a Microarray Experiment (MIAME)	Ciências da vida - Genética	OBI	http://www.fged.org/projects/miame/	ArrayExpress; GermOnline; ImmPort; GEO
Minimum Information about a Molecular Interaction Experiment (MIMIx)	Ciências da vida- Biologia molecular	PSI-MI CV	http://www.psidev.info/mimix	IntAct, MINT, MatrixDB
Minimum Information about a Proteomics Experiment (MIAPE)	Ciências da vida- Proteómica	PSI-MS e sepCV	http://www.psidev.info/miape	ProteoRed; ProteomeXChange; World 2D-PAGE; PRIDE
Minimum Information about a Simulation Experiment (MIASE)	Ciências da vida- Bioquímica	KiSAO	http://co.mbine.org/standards/miase	BioModels
Minimum Information About a RNAi Experiment (MIARE)	Ciências da vida- Biologia molecular	Nenhum	http://miare.sourceforge.net/HomePage	PubChem
Minimum Information About a Bioactive Entity (MIABE)	Ciências da vida - Farmacologia	PSI-MI CV	http://mibbi.sourceforge.net/projects/MIABE.shtml	canSAR
Minimum Information About a Cellular Assay (MIACA)	Ciências da vida – Biologia celular	Nenhum	http://miaca.sourceforge.net/	
Minimum information about a Neuroscience Investigation (MINI)	Ciências da vida- Neurociência	Nenhum	http://www.carmen.org.uk/?page_id=245	

4. Cadernos de Laboratório

As ferramentas disponíveis para os investigadores são decisivas na sua motivação para a gestão de dados. Em geral, espera-se que as ferramentas que simplificam o trabalho necessário na gestão de dados e produzam resultados claros sejam mais facilmente adotadas (Ribeiro et al. 2015). Em ambiente de investigação, os cadernos de laboratório são ferramentas fundamentais para dar suporte ao trabalho em laboratório, a fim de dar contexto aos dados e documentar o processo, procedimentos e descobertas.

No entanto, apesar da constante digitalização da atividade científica, existem ainda algumas áreas com práticas que não passaram para o digital como por exemplo a utilização dos cadernos de laboratório em suporte de papel durante a investigação. Apesar de os investigadores se sentirem confortáveis com a sua utilização, existem algumas desvantagens ao nível da preservação e uso da informação, como por exemplo, a dificuldade que colocam ao trabalho em colaboração. De modo a apoiar as atividades diárias dos investigadores com ferramentas como computadores e *tablets* e diminuir a dependência do papel, começaram a ser desenvolvidas aplicações – os cadernos de laboratório eletrónicos - de modo a ajudar os investigadores com as suas observações e anotações.

Estes cadernos de laboratório eletrónicos podem ser mais direcionados para uma abordagem geral nas tarefas de anotação ou serem mais específicos de um domínio de investigação, permitindo uma maior personalização. Uma abordagem generalizada nas atividades de anotação consiste no uso de *software* desenvolvido para responder às necessidades básicas relacionadas com as anotações (por exemplo lembretes de eventos futuros), sendo que os cadernos de domínios específicos permitem responder a necessidades mais específicas. As maiores diferenças de um caderno eletrónico comparativamente a um caderno analógico são (Amorim 2016):

- **Confiabilidade:** Apesar da preocupação sobre a preservação dos dados, existe uma maior facilidade de armazenamento e migração de dados em suporte digital;
- **Redundância:** Capacidade de evitar a perda dos dados em caso de acidentes ou ações indesejadas através de histórico de versões;
- **Colaboração:** capacidade de permitir a colaboração entre os diferentes membros de uma comunidade;
- **Acesso multiplataforma:** a sincronização com dispositivos portáteis torna os dados facilmente disponíveis em diferentes contextos e plataformas;

4.1 Aplicações genéricas de *note-taking*

Uma abordagem multiuso nas tarefas de anotação consiste numa aplicação genérica desenvolvida para responder às necessidades básicas dos utilizadores de armazenar as notas em meio digital e de colaborar em grupo.

4.1.1 Evernote

O Evernote é um *software* desenvolvido pela *Evernote Corporation,* permitindo aos seus utilizadores criar notas em formato de texto, imagem ou de áudio. Para além disso, permite a organização de notas em pastas, colocação de *tag*s e comentários e ainda a partilha de notas entre utilizadores. Esta é uma aplicação desenvolvida para plataformas diversas, podendo ser executada numa versão *online*¸ numa versão *desktop* tanto em Windows, como macOS, assim como em dispositivos portáteis Android e iOS. Apesar de contar com muitos utilizadores em todo o mundo, esta aplicação é considerada genérica e multidisciplinar já que não tem caraterísticas próprias para um domínio específico (Kanza 2017).

4.1.2 OneNote

O *software* OneNote desenvolvido pela Microsoft foi lançado em 2003. Este é um programa que permite a colaboração entre utilizadores, a organização de conteúdos em blocos de notas, secções e páginas, a colocação de etiquetas e a capacidade de criar notas através de formatos multimédia (imagens, texto, áudio e vídeos). As notas de cada utilizador podem ser partilhadas com outros utilizadores através da Internet ou rede pessoal. O OneNote em versão aplicação está disponível para sistemas Windows e macOS, assim como dispositivos iOS, Windows Phone e Android. Existe também uma versão *web* do OneNote através do OneDrive ou OfficeOnline.

4.1.3 Google Environment

Nos últimos anos, verificou-se um investimento cada vez maior da Google em plataformas colaborativas. De entre estas, é possível destacar a ferramenta Google Docs que permite aos seus utilizadores editar e rever documentos dentro de um grupo. O Google Drive permite o armazenamento de ficheiros e gerir o seu acesso a utilizadores e o Google Keep destina-se guardar notas. Os primeiros, Google Docs e Google Drive, são oferecidos e estão disponíveis como uma plataforma baseada na *web,* sendo acessíveis através de *browsers,* assim como em dispositivos móveis Android, iOS e Windows Phone. O Google Keep é uma aplicação especializada em *note-taking* e também está disponível na *web* e em dispositivos portáteis.

4.2 Mercado de cadernos de laboratório eletrónicos

O mercado de cadernos de laboratório eletrónicos pode ser dividido em duas categorias: os cadernos de laboratório específicos de um domínio de investigação que contêm aplicações específicas para instrumentação científica e tipos de dados próprios ou os cadernos de laboratório genéricos ou multidisciplinares que foram desenvolvidos de modo a suportar todos os dados e informação que devem ser registados num caderno de laboratório. Além do mais, estes cadernos de laboratório podem ser baseados em ambiente *web* ou serem software proprietário. Ao nível do seu licenciamento, estes podem ser comerciais, comerciais uma versão de teste gratuita e em código-aberto sujeito a vários tipos de licença (Kanza et al. 2017).

4.2.1 LabArchives

O LabArchives[20], lançado no mercado em 2009, é uma ferramenta colaborativa baseada num ambiente *cloud* que foi especificamente projetada para o armazenamento, organização, partilha e publicação de dados de investigação. O LabArchives fornece uma ferramenta simples para ser utilizada por investigadores para a gestão dos seus dados de investigação. Para além de um simples ELN, o LabArchives fornece também uma plataforma flexível e extensível que pode ser personalizada, de modo a combinar com o fluxo de trabalho existente nos laboratórios de investigação. Algumas das suas principais caraterísticas e funções são a capacidade de conectar dados de laboratório e ficheiros de imagens a observações e anotações, facilidade de importação e acesso a dados experimentais digitais capturados pelos instrumentos de laboratório e produzidos através de *software ou hardware,* simplicidade de *upload* de imagens ou vídeos durante a realização de experiências, e por fim a possibilidade de publicação e partilha dos dados de investigação para indivíduos específicos ou público em geral. Ao nível das plataformas, o LabArchives pode ser utilizado através de sistemas Windows e macOS, assim como em dispositivos portáteis Android e iOS.

4.2.2 Benchling

O Benchling é uma empresa, fundada em 2012, baseada em San Francisco nos Estados Unidos que se foca na criação de ferramentas de *software* baseadas na *cloud* e que permitem a edição digital de sequências de ADN, registo de experiências de laboratório e a análise e partilha de dados de investigação. Um *software* comercializado é o caderno de laboratório eletrónico[21] com o mesmo nome que permite aos investigadores documentar as suas

[20] https://www.labarchives.com/ Consultado a 05/01/2019
[21] https://benchling.com/enterprise/lab-notebook Consultado a 05/01/2019

experiências, sobretudo no domínio das ciências biológicas e químicas. Este caderno permite a integração com outro *software* comercializado pela empresa, tal como o Bioregistry (sistema de registo de organismos biológicos), Sample Tracking e o Molecular Biology Suite (análise de moléculas).

4.2.3 LabCollector

O LabCollector[22] é um caderno de laboratório eletrónico desenvolvido pela AgileBio e lançado no mercado em 2002. É facilmente adaptado a qualquer situação ou necessidade devido à sua incorporação de módulos de dados pré-definidos, totalmente personalizáveis com bastantes campos, filtros de pesquisa e ferramentas de análise específicas. Além disso, os módulos personalizados podem ser definidos para armazenar qualquer tipo de dados e informação sobre células, sequências biológicas, anticorpos, moléculas, animais, entre outras. Para além disso, permite a colaboração entre investigadores do mesmo laboratório, tem a possibilidade de os dados serem acedidos por qualquer dispositivo ou computador ligado à rede e exporta dados em diversos formatos normalizados.

4.2.4 LabTablet

O LabTablet é um caderno de laboratório eletrónico desenvolvido e projetado para a facilidade de uso e integração contínua com o fluxo de trabalho de investigação, ajudando os investigadores a produzir melhores descrições para os seus dados, aproveitando os recursos de um dispositivo móvel tais como sensores, câmara e microfone. Pode ser personalizado de modo a responder melhor às necessidades de descrição de dados de investigadores de diferentes domínios, incluindo diferentes conjuntos de descritores baseados em ontologias. De seguida, estes podem ser usados para criar registos de metadados que podem ser partilhados dentro de um grupo por meio de uma plataforma colaborativa de gestão de dados baseada na *web*, o Dendro. O Dendro é uma plataforma intermédia baseada em ontologias com o objetivo de permitir que os investigadores descrevam e giram dados com metadados específicos de diferentes domínios. Por fim, a plataforma Dendro está ligada e integrada com repositórios, como o CKAN de modo a que os dados e metadados sejam exportados e preservados a longo prazo e sejam mais facilmente acessíveis pela comunidade. Todo este fluxo de trabalho a partir do LabTablet está representado na Figura 5.

[22] https://labcollector.com/ Consultado a 05/01/2019

Figura 5 – Exemplo de workflow de gestão dos dados de investigação a partir do LabTablet

4.2.5 Comparação de cadernos de laboratório

O desenvolvimento de cadernos de laboratório eletrónicos contribui para a revolução digital na ciência, sendo ferramentas importantes na gestão de dados de investigação. Comparando os diferentes cadernos de laboratório eletrónicos, há algumas preocupações a ter em conta, nomeadamente relacionadas com aspetos de portabilidade, segurança e partilha dos dados, assim como a capacidade de colaboração entre os investigadores a trabalhar num projeto. Todas estas caraterísticas influenciam e têm impacto no sucesso de implementação de um caderno eletrónico num laboratório de investigação. A Tabela 3 apresenta uma comparação entre quatro cadernos de laboratório de acordo com as caraterísticas de compatibilidade com sistemas operativos e dispositivos móveis, armazenamento e partilha de dados, colaboração entre membros, integração direta com repositórios e acessibilidade de uso.

Tabela 3 - Comparação entre cadernos de laboratório eletrónicos

ELN	Compatibilidade	Versão dispositivo móvel	Armazenamento	Partilha de dados	Colaboração	Integração com repositório de dados	Acessibilidade
LabArchives	Windows, MacOS, Linux, Android e iOS	Sim	Cloud	Sim	Sim	Sim	Versões gratuitas e pagas
Benchling	Windows, MacOS, Linux	Não	Cloud	Sim	Sim	Sim	Versões gratuitas e pagas
LabCollector	Android, MacOS, Windows e Linux	Sim	Cloud ou local	Sim	Sim	Sim	Versões gratuitas e pagas
LabTablet	Android	Sim	Através do Dendro	Sim	Através do Dendro	Não	Gratuito

5. Ontologia para as ciências biomédicas

Ao longo do seguinte capítulo é apresentado o processo seguido tendo em vista o desenvolvimento da ontologia para os domínios de investigação dos investigadores do I3S. Numa primeira fase, fez-se uma seleção e análise de modelos de metadados destes domínios de modo a recuperar conceitos relevantes para serem inseridos como *data properties* na ontologia. De seguida, desenvolve-se a ontologia através da ferramenta Protégé sob o formato de OWL (Ontology Web Language). Por fim, exportou-se esta ontologia para a plataforma Dendro de modo a que esta possa ser utilizada nas tarefas de descrição de dados com os investigadores do I3S.

5.1 Procedimentos

5.1.1 Seleção e análise de modelos de metadados

A maioria das soluções atuais para a gestão de dados de investigação baseia-se num conjunto fixo de descritores (por ex. Dublin Core) para a descrição dos recursos. Apesar de serem fáceis de usar e compreender, a sua semântica é limitada a conceitos mais gerais, deixando de parte metadados específicos dos domínios. Apesar de os descritores normalizados serem úteis para assegurar a interoperabilidade, cada domínio de investigação necessita de descritores específicos de modo a garantir a precisão dos metadados. Os modelos de metadados devem ser abrangentes de modo a fornecerem todos os descritores necessários para os investigadores descreverem o seu conjunto de dados, serem simples de utilizar, e devem ainda ajudar a promover a troca de dados dentro do grupo de investigação e a sua descoberta e recuperação para posterior reutilização (Willis et. al. 2012). É neste sentido que neste trabalho se pretendeu validar um vocabulário para a correta descrição dos dados de investigação dos grupos de investigação do I3S, dando ainda liberdade de sugestão de descritores aos investigadores.

Em trabalhos anteriores elaborados no InfoLab, o processo de recolha de dados começou com uma entrevista conduzida pelo curador aos investigadores (Castro et al. 2017). Esta entrevista é útil na medida que fornece ao curador uma visão mais enriquecedora sobre o domínio de investigação, e sobre, as metodologias de trabalho adotadas pelos investigadores. Porém, face à indisponibilidade dos investigadores de colaborarem continuamente ao longo de um período alargado devido aos projetos em que estão envolvidos, a metodologia utilizada neste trabalho para a criação do modelo de metadados foi adaptada. Numa primeira fase optou-se por uma pesquisa por repositórios da área das ciências biológicas no portal re3data de modo a fazer uma cobertura das plataformas existentes no domínio (2.2.1 Repositórios de

dados de investigação no domínio das ciências biológicas). Durante esta fase analisaram-se os repositórios de uma perspetiva da organização e tipologia dos dados de investigação existentes e metadados utilizados para a pesquisa. Esta fase inicial foi importante já que, não havendo um contacto *a priori* com os investigadores, serviu para compreender melhor os domínios de investigação, além de ajudar a propor algo mais abrangente nos domínios dos investigadores I3S.

Numa fase seguinte, pesquisaram-se também modelos de metadados padrão utilizados nestes domínios. Utilizando os sítios *web* do Digital Curation Centre (DCC) e da Research Data Alliance (RDA), procuraram-se modelos utilizados nas diversas ciências biológicas (bioquímica, genética, biologia celular, biodiversidade, neurociência, etc.). Para além destes, utilizou-se também o portal FAIRsharing (www.fairsharing.org) visto ser um portal curado e pesquisável com *standards* e repositórios de dados das ciências da vida abrangendo as ciências biológicas, ambientais e biomédicas. A equipa deste portal organiza a informação sobre os *standards* utilizados na identificação, citação e descrição de dados e metadados a partir da sua divisão em quatro tipos (Sansone et al., 2019).

Em primeiro lugar, as diretrizes mínimas, também conhecidas como listas de verificação mínimas, delineiam as informações necessárias e suficientes para contextualizar e compreender um objeto digital. Em segundo lugar, as terminologias, variando desde vocabulários controlados a ontologias, fornecem definições inequívocas para conceitos e objetos. Além destes, os modelos e formatos definem a estrutura e relação da informação para um modelo concetual e incluem formatos para a partilha e transmissão de dados de modo a facilitar a sua troca entre diferentes sistemas. Por fim, os esquemas de identificação são sistemas formais de recursos e outros objetos digitais que permitem a sua identificação única.

Na fase seguinte, foi necessário estabelecer um número mínimo de descritores a serem utilizados para a descrição de dados nestes domínios, assim como definir os parâmetros utilizados para a sua seleção. O principal objetivo foi assegurar a comunicação e a partilha de metadados experimentais de modo a garantir que os conjuntos de dados de investigação sejam compreensíveis, reproduzíveis, comparáveis e reutilizáveis (Willis, Greenberg, and White 2012).

Para tal, procedeu-se a uma análise de uma amostra dos diferentes normativos como o Investigation Study Assay Tabular Data (ISA-TAB), algumas listas de diretrizes do projeto Minimum Information Study for Biological and Biomedical Investigations (MIBBI) (**Tabela *2*** - Amostra de esquemas do projeto MIBBI) (Taylor et al. 2009) e algumas ontologias

desenvolvidas para estes domínios (ex. Ontology for Biomedical Investigations (OBI)[23] e Experimental Factor Ontology (EFO)[24].

Devido ao grande número de descritores que podem ser utilizados para a descrição das experiências destes domínios, tornou-se necessário definir um número mínimo para representar na ontologia proposta e na plataforma Dendro. A seleção dos descritores, com base no estado de arte, seguiu os seguintes critérios:

- Generalidade dos descritores de modo a que possam ser utilizados em qualquer estudo, dando liberdade de sugestão de descritores a cada investigador;
- Representação das caraterísticas das amostras utilizadas na investigação (Age; Organism; Organism Part; Sex, Sample Type);
- Representação dos procedimentos e passos realizados ao longo do estudo e respetivos ensaios (Methods, Protocols);
- Representação dos materiais utilizados ao longo do estudo e respetivos ensaios (Material, Instruments, Software);
- Frequência de utilização como metadado em repositórios destes domínios (Kevin Read, 2015).

No final da seleção, definiram-se e representaram-se 30 descritores na ontologia proposta para este caso. Estes descritores passaram, antes das sessões com os investigadores, pela validação da Ana Luís Ferreira[25].

Tabela 4 - Lista dos descritores selecionados

Descritor	Fonte	Descrição	Exemplo de Valor
Age	EFO	Physical age of the sampled organism(s).	32 years old / 3 weeks
Assay Type	OBI, EFO	A planned process with the objective to produce information about the material entity that is the evaluant, by physically examining it or its proxies. (e.g. in live cell assay, cytometry assay, in vivo intervention, etc).	Sequencing assay

[23] https://bioportal.bioontology.org/ontologies/OBI
[24] https://bioportal.bioontology.org/ontologies/EFO
[25] Trabalho realizado em conjunto com o mesmo grupo de investigadores e que tem formação na área de Bioengenharia

Cell Line	MIACA, OBI, MESH	A cultured cell population that represents a genetically stable and homogenous population of cultured cells that shares a common propagation history (i.e. has been successively passaged together in culture).	MKN28
Cell Type	MIACA, OBI	A cell type is a distinct morphological or functional form of cell. Examples are epithelial, glial etc.	Epithelial cells
Collection Date	MixS	Date on which the sample was collected.	17/05/2017
Developmental Stage		Developmental stage of the organism.	adult
Disease	EFO, MESH	A disease is a disposition that describes states of disease associated with a sample and/or organism. Disease being studied.	Alzheimer´s disease
Drug Usage		Any drug used in the study and the frequency of usage; can include multiple drugs used. Register the respective dose.	
Environmental Factor	EFO	The external elements and conditions which surround, influence, and affect the life and development of an organism or population (e.g. atmospheric, hydrostatic pressure, light, etc).	Free text
Ethnicity		Ethnicity of the subject.	asian
Experimental Factor	EFO	The name of factor used in the experiment. A factor corresponds to an independent variable manipulated by the experimentalist with the intention to affect biological systems in a way that can be measured by an assay. (e.g. chemical substance, temperature, biological replicate, etc).	Infection
Instrument Name	MIAME, MIAPE, OBI, EFO	An instrument is a device which provides a mechanical or electronic function.	Illumina 610K
Instrument Type	MIAME, MIAPE, OBI, EFO	Type of the instrument used (e.g. liquid handling robot, centrifuge, FACS, Plate Reader, etc).	Sequencing technology
Material	-	Description of all the materials used during the experiment.	ethanol

Measurement	EFO	A measurement is an information entity that is a recording of the output of a measurement such as produced by an instrument.	Free text
Method	MIAME, MIAPE	A description of the methods followed during the experiment.	Free text
Molecule		Any polyatomic entity that is an electrically neutral entity consisting of more than one atom.	Carbon dioxide
Organism	OBI, MIAME	Identifier or name of the organism from which the biomaterial was derived.	Homo sapiens
Organism Part	OBI	The part of organism's anatomy or substance arising from an organism from which the biomaterial was derived, excludes cells. E.g. tissue, organ, system, sperm, blood or body location (arm).	Stomach
Reagent	OBI	Description of reagents (media, supplements, transfection reagents) used during the experiment.	sulphur
Sample Collection Protocol	EFO	Describes the procedure whereby biological samples for an experiment are sourced.	Free text
Sample Size	MESH	The number of units (persons, animals, patients, specified circumstances, etc.) in a population to be studied. The sample size should be big enough to have a high likelihood of detecting a true difference between two groups.	466 samples
Sample Type		Description and characteristics of the samples gathered or used in the experiment.	genomic
Sex	EFO, MESH	Physical sex of sampled organism(s).	female
Software	MIAME, MIAPE, MIATA	The software applications used in the gathering or processing procedures carried out in Experiment.	Bead Studio
Study Design	OBI	A plan specification comprised of protocols (which may specify how and what kinds of data will be gathered) that are executed as part of an investigation and is realized during a study design execution. (e.g. in vitro design, clinical study design, cellular process design, etc).	in vitro design

Study Domain	-	A domain of an experiment is a field of study in which an experiment is designed to discover new knowledge.	Disease susceptibility
Temperature		Temperature registered during some Experiment/Assay.	18 Cº
Tissue	OBI, MIACA	Type of tissue the sample was taken from.	Muscle tissue
Treatment Protocol	EFO	A protocol in which the aim is to treat a sample, collection of samples, organism or group of organisms for some experimental analysis of outcome.	Free text

5.1.2 Desenvolvimento da ontologia

Uma ontologia (*onto + logia)* é definida, na ciência da informação e computação, como um conjunto estruturado de termos e conceitos que representa um conhecimento sobre o mundo[26]. Por esta razão, as ontologias são reconhecidas na comunidade como ferramentas essenciais para a descrição de recursos na *web* semântica uma vez que possibilitam representar o significado de cada descritor usado como metadado numa forma facilmente processada por uma máquina ou sistema (Berners-Lee, Hendler & Lassila 2001). A adoção de uma ontologia no âmbito da gestão de dados de investigação é conveniente numa perspetiva que as ontologias podem promover o entendimento acerca do vocabulário necessário dentro dos diferentes domínios de investigação (Castro et al. 2014).

Diversas comunidades científicas, tendo conhecimento das vantagens de adotar ontologias para a descrição de recursos, têm trabalhado de modo a desenvolver ontologias orientadas para os seus domínios. Dois exemplos são a OBOE (The Extensible Observation Ontology) desenvolvida para os domínios da biodiversidade, ecologia e ciências naturais e a CERIF (The Common European Research Format Ontology) que tem sido adotada para a partilha e descrição de projetos e conjuntos de dados de investigação. No domínio das ciências biológicas e biomédicas, existe também um esforço por parte da comunidade, a NCBO (National Center for Biomedical Ontologies), no desenvolvimento de ontologias de modo a permitir que os cientistas criem, disseminem e giram informação e conhecimento de uma forma facilmente processada por uma máquina. Esta comunidade é responsável pela manutenção do portal BioPortal, o maior repositório de ontologias na área das biomédicas. Neste portal, é possível

[26] https://dicionario.priberam.org/ontologia Consultado a 25/04/2019

pesquisar pelo nome das ontologias e por classes na página inicial [Figura 6], assim como filtrar ontologias através de parâmetros tais como a sua categoria e domínio, o seu grupo desenvolvedor, a data de submissão e o seu formato.

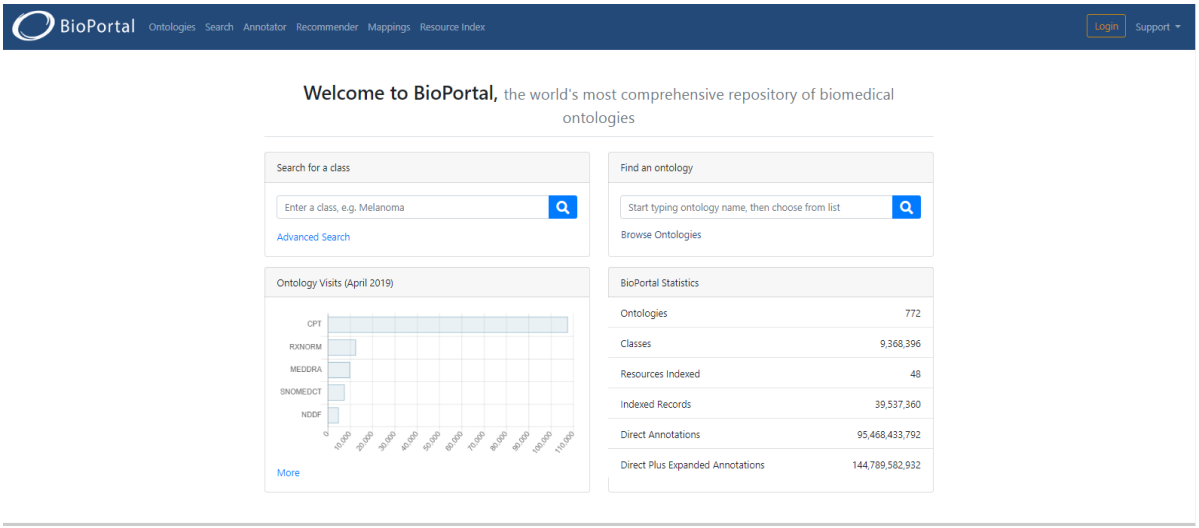


Figura 6 - Página inicial do BioPortal

As ontologias são a base do modelo de funcionamento do Dendro, oferecendo aos curadores de dados a possibilidade de adicionarem descritores considerados importantes para a descrição de dados de investigação. Este processo é relativamente simples já que os curadores, mesmo com capacidades de programação limitadas podem desenvolver ontologias e importá-las para o Dendro para serem combinadas com outras ontologias já disponíveis (Silva 2016). Neste sentido, após a seleção dos descritores, foi necessário representá-los sobre a forma de uma ontologia, sendo utilizado *software* Protégé[27] para esta tarefa. Todos os descritores foram formalizados como "*data properties*" através do desenvolvimento de uma *"lightweight ontology"* e para cada propriedade atribuíram-se annotation *properties* especificando as suas *rdfs:labels* (representa o descritor em linguagem natural) e a *rdfs:comments* (significado de um descritor) . Esta descrição em linguagem natural é adequada para facilitar a interpretação por humanos e, principalmente porque a plataforma Dendro usa as *annotation properties* das ontologias na sua interface. Além disso, uma *"lightweight ontology"* é conveniente devido à sua simplicidade (poucas classes e restrições fracas) facilitando o processamento por sistemas como o Dendro.

Numa primeira fase, criaram-se todas as 30 *data properties* na nova ontologia que se denominou por MIBBIUP (MIBBI – University of Porto) visto ser uma versão mais resumida

[27] https://protege.stanford.edu/

e genérica baseada nas listas de informação mínimas do projeto MIBBI para as experiências dentro destes domínios.

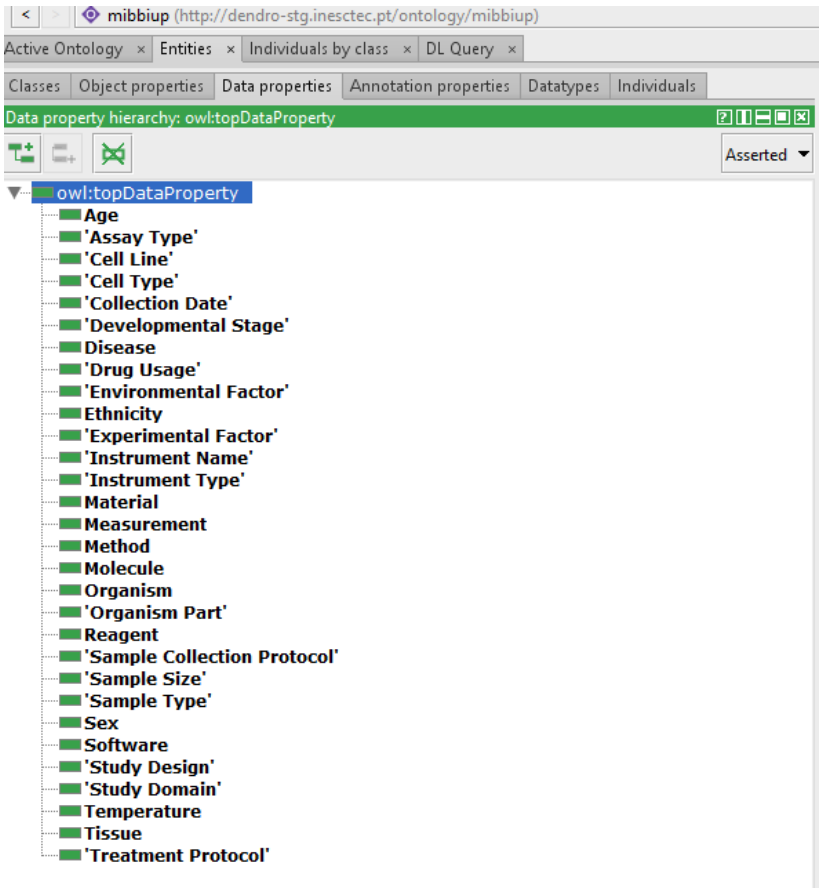


Figura 7 - Data properties da ontologia MIBBUP

Durante este processo de criação das *datas properties,* foi necessário associar-lhes informação adicional sob a forma de *annotation properties*. Desta forma, foram inseridas a *annotation property rdfs:label* em formato *string* para definir como o descritor deve ser apresentado ao utilizador, a propriedade *comments* onde se apresenta uma definição do descritor e por vezes um exemplo de um valor, e a *IsDefinedBy* para os casos em que um exemplo de um descritor siga uma norma específica. Na propriedade *comments*, a definição do descritor foi retirada de outras ontologias.

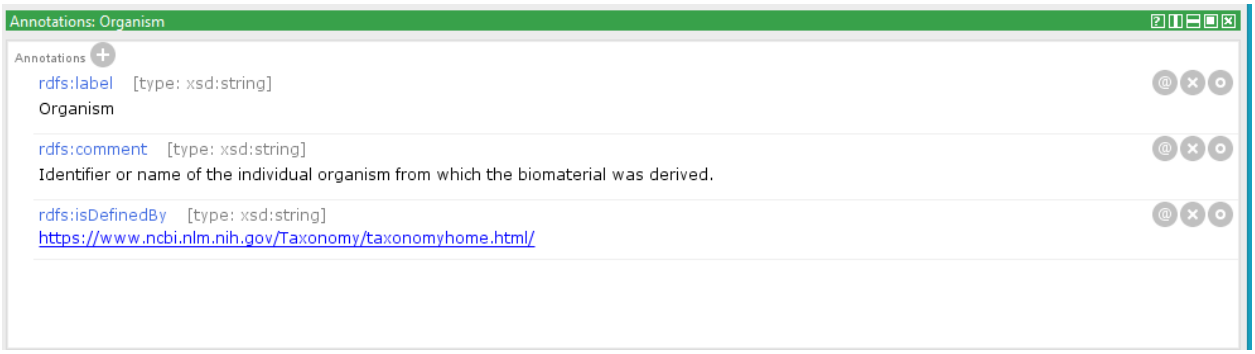


Figura 8 - Annotation properties na ontologia

Ainda antes da exportação da ontologia para o Dendro, foi necessário criar um documento de texto para definir o formato da caixa de texto que cada descritor apresentaria no Dendro.

```
/**
* Elements of the mibbiup
*/
Elements.mibbiup =
{
        Age :
        {
                type: Dbconnection.string,
                control: Config.controls.input_box
        },
        Assay_Type :
        {
                type: Dbconnection.string,
                control: Config.controls.input_box
        },
```

Figura 9- Definição do formato da caixa de texto dos descritores da ontologia no Dendro

5.1.3 A plataforma Dendro

A plataforma Dendro[28], desenvolvida na FEUP e INESC TEC, é uma plataforma para a organização e descrição de dados de investigação, facilitando a criação de metadados e a sua exportação como Linked Open Data (Silva 2014). A sua interface é semelhante a outras plataformas de armazenamento remoto, tais como a Dropbox[29]. O principal objetivo do Dendro é envolver os investigadores na descrição dos seus dados e prepará-los para

[28] http://dendro-stg.inesctec.pt/
[29] https://www.dropbox.com/

publicação num repositório de dados integrado com a plataforma caso os investigadores pretendam fazê-lo.

O Dendro apresenta uma interface de preenchimento de campos cujo principal objetivo é guiar o investigador na descrição dos seus dados. Cada elemento que se adiciona à folha de descrição (ex. Creator) define um determinado contexto e está associado a um determinado ficheiro ou pasta inserida na plataforma. A plataforma, fruto da sua adaptabilidade, ainda é capaz de criar um perfil para cada investigador com base nos dados recolhidos ao longo das diversas interações sendo que, dentro de cada pasta, os investigadores podem copiar a folha de descrição e replicá-la para outro local do projeto, alterando apenas as descrições necessárias.

Na sua utilização, os investigadores podem selecionar livremente descritores de várias ontologias se considerarem que a sua inclusão é relevante para apoiar a interpretação dos seus dados. Esta plataforma apresenta já bastantes ontologias genéricas (ex. Friend of a Friend e uma versão normalizada de Dublin Core), assim como ontologias desenvolvidas, após sessões com investigadores, especificamente para determinados domínios de investigação (ex. Double Cantilever Beam, Hydrogen Generation e Biological Oceanography).

Relativamente ao *design* do Dendro, destaca-se a hierarquia de pastas e ficheiros que tem no topo os diversos projetos criados pelos utilizadores e nos seguintes níveis as diferentes pastas criadas que podem conter no seu interior os mais diversos ficheiros em diferentes formatos, assim como os metadados a eles associados. Na Figura 9 é possível observar a interface de descrição da plataforma Dendro. Após a criação do projeto, os investigadores podem criar diversas pastas e depositar dentro delas os diversos ficheiros (conjunto de dados em Excel, artigos, outros). À direita encontram-se as listas de descritores das quais os investigadores podem escolher e combinar os descritores livremente. No centro encontra-se a área de descrição na qual aparecem os descritores escolhidos e preenchidos.

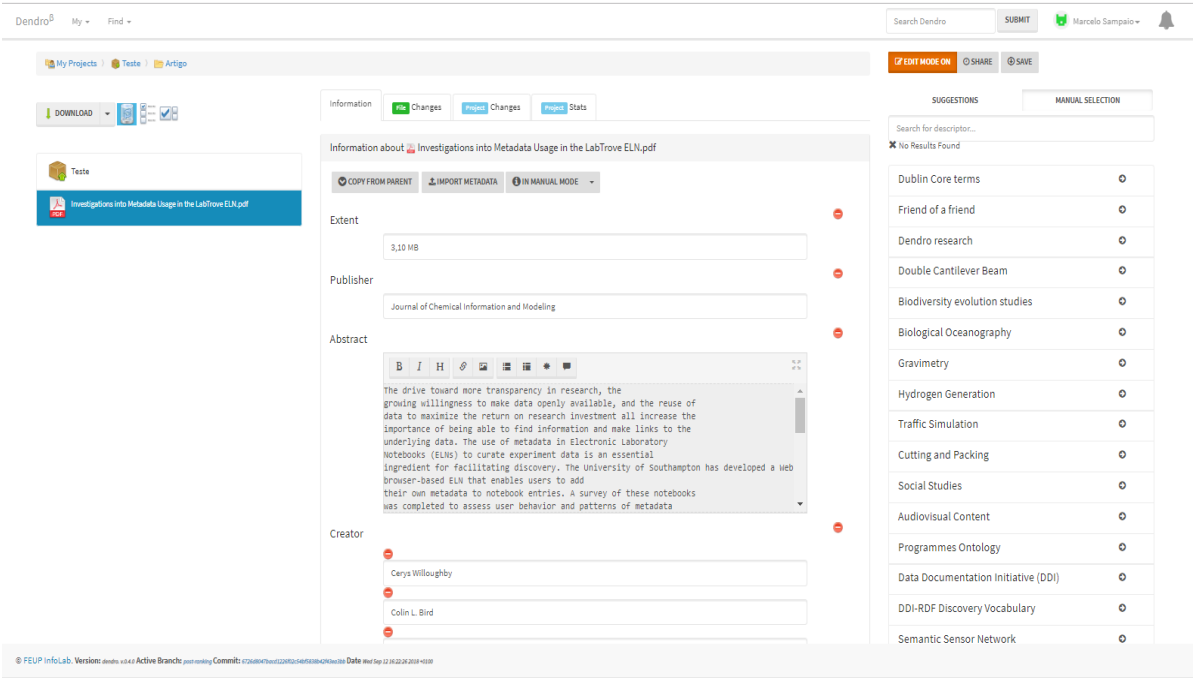


Figura 10 - Interface da plataforma Dendro

A ontologia "Minimum Information for Biological and Biomedical Investigations" foi também introduzida na plataforma Dendro, contribuindo para a lista total de modelos e oferecendo descritores dos domínios das ciências biológicas e biomédicas.

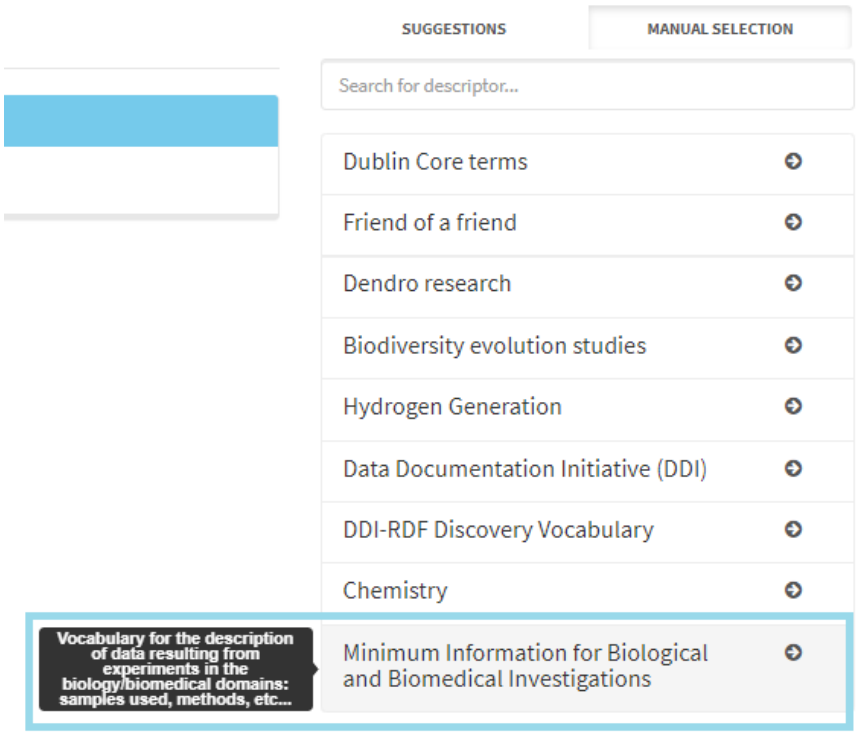


Figura 11 - Ontologia MIBBIUP na lista de ontologias do Dendro

Após a seleção da ontologia, todos os trinta descritores ficam representados por ordem alfabética e apresentam a descrição inserida na ontologia através das *annotation properties*. Para usar um descritor, basta carregar sobre este no painel da direita e preenchê-lo na área de descrição.

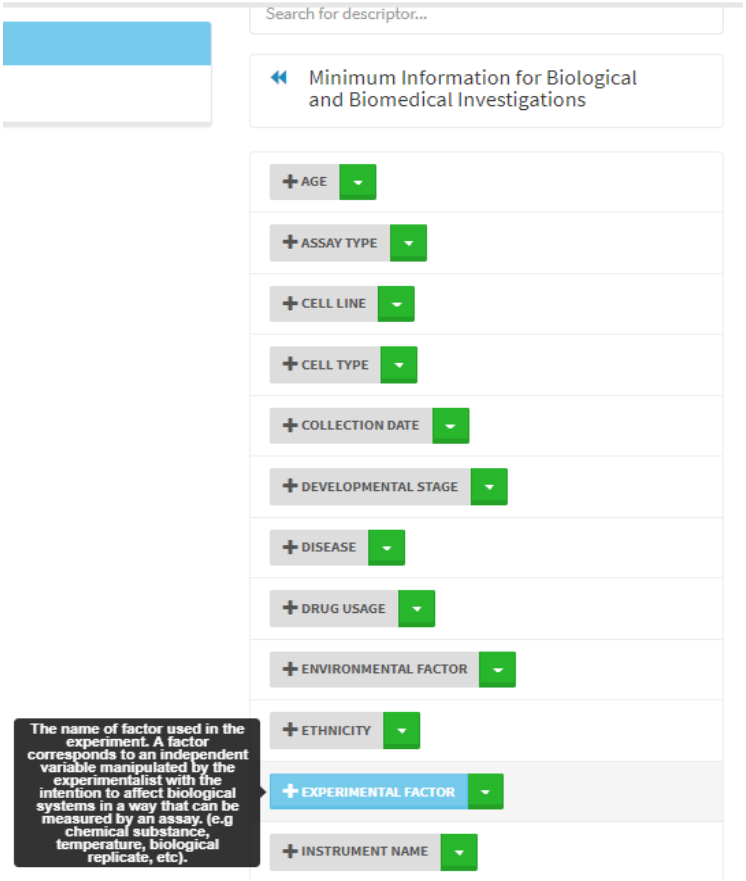


Figura 12 - Exemplo de descritores da ontologia MIBBIUP no Dendro

Numa fase final e se o investigador assim o entender, os dados e metadados podem ser transferidos para um repositório de dados integrado com a plataforma, tais como DSpace, Figshare, Zenodo. Aí o utilizador deve preencher o título, o endereço do repositório e nalguns casos o nome de utilizador, palavra-passe e *token* de acesso.

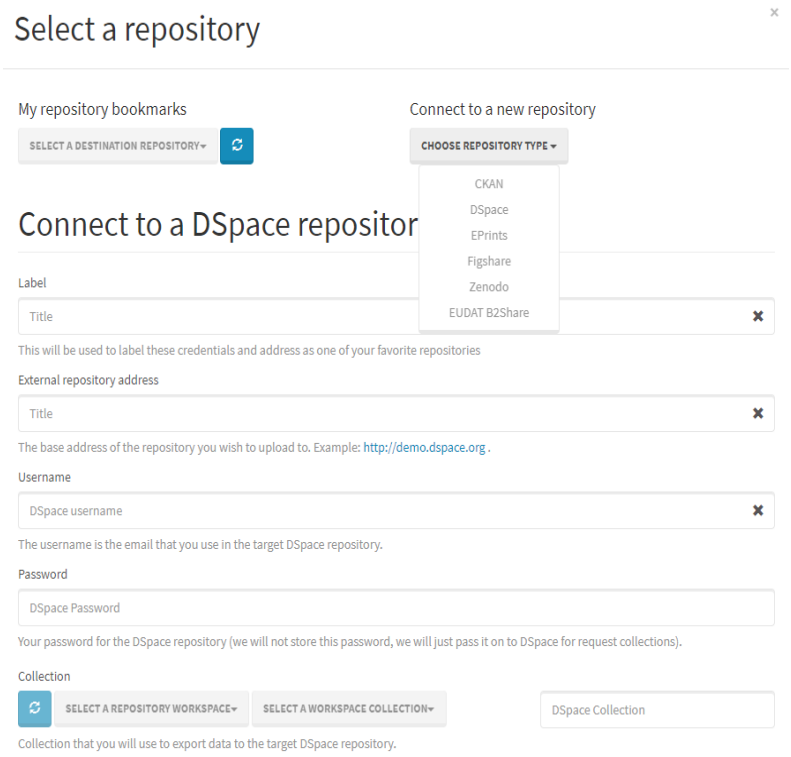


Figura 13 - Repositórios integrados com a plataforma Dendro

6. Resultados

Ao longo do seguinte capítulo, proceder-se-á à discussão e análise dos resultados obtidos nas sessões realizadas com os investigadores participantes do Instituto de Inovação e Investigação em Saúde. De modo a recolher os dados, foram feitas sessões presenciais com sete investigadores de quatro grupos de investigação das quais se procurou obter *feedback* relativamente ao modelo de metadados desenvolvido em formato de ontologia e mais tarde em colocar os investigadores a interagir com a plataforma Dendro, sobretudo ao nível da descrição de dados. Com estas sessões, espera-se contribuir para orientar os investigadores para a utilização de ferramentas eletrónicas de gestão de dados de investigação e importância da descrição de dados.

Utilizaram-se também as respostas das entrevistas feitas com os investigadores por parte da Ana Luís Ferreira[30] principalmente aquelas relacionadas com a organização e descrição de dados de investigação. Com estas entrevistas procurou-se conhecer melhor o grupo e projeto de investigação do investigador e as suas práticas relacionadas com a gestão de dados de investigação.

6.1 Feedback dos investigadores

Os investigadores devem ser considerados uma parte importante no processo de descrição de dados de investigação já que como produtores de dados e tendo maior conhecimento sobre os domínios de investigação têm a capacidade de criar registos de metadados com mais qualidade. Neste sentido, os curadores de dados devem procurar colaborar com os investigadores com o objetivo de desenvolver um vocabulário de metadados que apoie as atividades de descrição de dados.

Após a seleção de conceitos e a formalização como *data properties* no desenvolvimento da ontologia, deu-se início aos primeiros contactos com os investigadores do I3S. Durante esta fase, procurou-se obter investigadores de diferentes grupos de investigação de modo a ter variedade no que diz respeito aos domínios de investigação representados. Com esta opção, pretendeu-se observar se existiam diferenças significativas ao nível da gestão de dados de investigação por grupo de investigação, assim como avaliar se o modelo de metadados proposto podia ser utilizado nas tarefas de descrição de dados, independentemente do subdomínio de investigação. No entanto, deu-se também a liberdade a cada investigador de

[30] Trabalho realizado em conjunto com o mesmo grupo de investigadores

sugerir novos descritores que ajudem a apoiar a interpretação dos seus dados de investigação, bem como de comentar os descritores apresentados.

Entre março e maio, realizaram-se sessões com um investigador do grupo de Diferenciação e Cancro, um investigador do grupo de Biologia Celular Glial, dois do grupo de Interações Epiteliais no Cancro e três investigadores do grupo de Diversidade Genética.

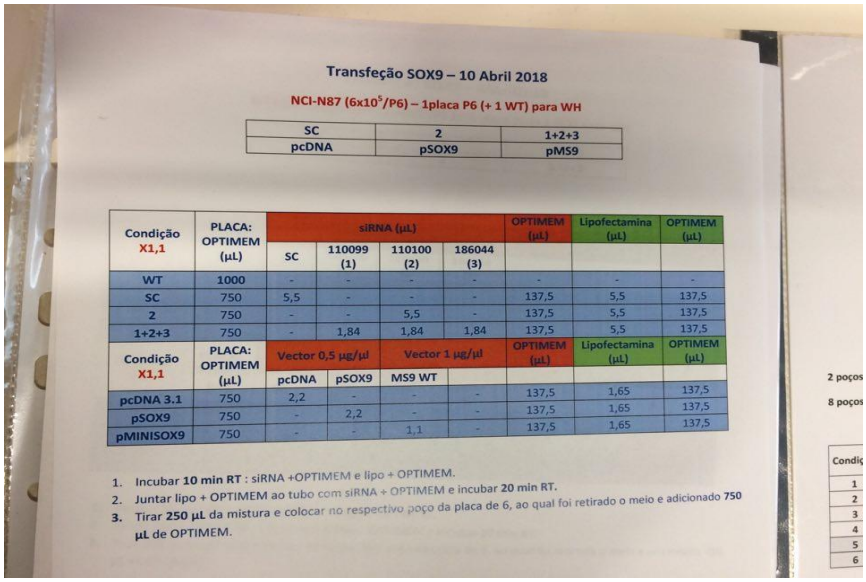
6.1.1 Diferenciação e Cancro

O grupo de investigação de Diferenciação e Cancro (em inglês Differentiation & Cancer) pretende compreender o impacto dos mecanismos regulatórios, transcricionais e pós transcricionais na regulação e progressão do cancro. Os investigadores focam-se principalmente nos mecanismos que influenciam a diferenciação (comparação entre tumores) e constituem o ponto de partida de um número significativo de processos carcinogénicos, nomeadamente no trato gastrointestinal. Para tal, adotam uma abordagem integrativa usando amostras clínicas, sistemas celulares ou modelos de animais de modo a identificar os principais eventos moleculares e biomarcadores que podem contribuir para entender a biologia do cancro.

A **investigadora A** é uma investigadora auxiliar post-doc neste grupo sendo que no momento em que a entrevista foi feita estava envolvida em mais que um projeto. Um dos projetos está relacionado com o estudo do cancro gástrico e intestinal. De uma forma resumida, o projeto consiste no estudo das regiões reguladoras em células isoladas do cancro gástrico e intestinal para ver a heterogeneidade do cancro e ver que fatores de transcrição estarão envolvidos na regulação das diferentes células para tentar arranjar forma de melhorar o diagnóstico e a terapêutica.

Relativamente à organização e descrição de dados, a investigadora utiliza o computador para organizar os dados dividindo-os por projeto e técnicas utilizadas. Por vezes, imprime também os dados já que se sente mais confortável observando os dados e técnicas utilizadas através do papel. Quando questionada sobre o conceito de metadados, esta respondeu não o conhecer, mas após explicação reconheceu-o e afirma fazer anotações sobre os seus dados, sobretudo ao nível das amostras e técnicas utilizadas para produzir os dados. Além disso, reconheceu a importância dos metadados para a pesquisa e recuperação dos dados.

Na Figura 14, estão representados os métodos e as quantidades dos materiais utilizados durante uma experiência que a investigadora realizou num determinado dia. Quando procura produzir novos dados, a investigadora apenas altera os campos da tabela e realiza novamente a experiência.



Transfeção SOX9 – 10 Abril 2018

NCI-N87 (6x10^5/P6) – 1placa P6 (+ 1 WT) para WH

SC	2	1+2+3
pcDNA	pSOX9	pMS9

Condição X1,1	PLACA: OPTIMEM (µL)	siRNA (µL)				OPTIMEM (µL)	Lipofectamina (µL)	OPTIMEM (µL)
		SC	110099 (1)	110100 (2)	186044 (3)			
WT	1000	-	-	-	-	-	-	-
SC	750	5,5	-	-	-	137,5	5,5	137,5
2	750	-	-	5,5	-	137,5	5,5	137,5
1+2+3	750	-	1,84	1,84	1,84	137,5	5,5	137,5

Condição X1,1	PLACA: OPTIMEM (µL)	Vector 0,5 µg/µl		Vector 1 µg/µl		OPTIMEM (µL)	Lipofectamina (µL)	OPTIMEM (µL)
		pcDNA	pSOX9	MS9 WT				
pcDNA 3.1	750	2,2	-	-	-	137,5	1,65	137,5
pSOX9	750	-	2,2	-	-	137,5	1,65	137,5
pMINISOX9	750	-	-	1,1	-	137,5	1,65	137,5

1. Incubar **10 min RT** : siRNA +OPTIMEM e lipo + OPTIMEM.
2. Juntar lipo + OPTIMEM ao tubo com siRNA + OPTIMEM e incubar **20 min RT.**
3. Tirar **250 µL** da mistura e colocar no respectivo poço da placa de 6, ao qual foi retirado o meio e adicionado **750 µL** de OPTIMEM.

Figura 14 - Método de Transfeção

Autoria: Investigadora A

Após a entrevista, a lista dos descritores selecionados foi mostrada à investigadora de modo a validar aqueles que considerasse úteis para descrever os seus dados de investigação e sugerir novos descritores. Dos trinta descritores apresentados, a investigadora validou treze descritores (43%), fazendo sugestões de melhoria às anotações de três descritores.

A investigadora necessitaria de mais descritores dependendo da experiência em que estaria a trabalhar, no entanto compreendeu a generalidade dos descritores e demonstrou ter preferência por um menor número de descritores de modo a evitar perder tempo a preenchê-los todos. Desta forma, a investigadora optaria pelo descritor Method para uma descrição detalhada dos passos realizados durante uma determinada experiência, enquanto os restantes serviriam para ajudar à interpretação dos resultados. Os descritores sobre as amostras parecem-lhe válidos já que no seu laboratório lidam com amostras clínicas de pessoas, assim como com sistemas celulares.

6.1.2 Biologia Celular Glial

O grupo de investigação de Biologia Celular Glial (em inglês Glial Cell Biology) concentra-se no processo de mielinização no qual os axônios são cobertos pela mielina (uma membrana biológica rica em lípidos que forma invólucros multimelares em espiral ao redor dos axónios). Utilizando abordagens transgênicas em ratos juntamente com sistemas de cultura de células apropriados, o grupo investiga como a integrina e as proteínas *rhoGTPases* regulam as diferentes fases de desenvolvimento e mielinização das células de Schwann (SC) e

oligodendrócitos (OL). Através da proteómica e sequenciação, o grupo pretende identificar novas moléculas candidatas a regular as interações neurónio-gliais que também podem ser importantes para melhorar a remielinização.

A **investigadora B** é uma aluna de doutoramento dentro deste grupo estando envolvida no estudo de um modelo de taupatia e drosófila, procurando descobrir uma fosfatasse que atue sobre a proteína tau. Para o armazenamento, a investigadora utiliza o disco interno do grupo pelo que organiza os dados de acordo com a experiência utilizada, por datas ou ainda por genótipos. Relativamente à familiarização com o conceito de metadados, a investigadora afirma já ter ouvido falar, mas mostrou dificuldades em explicar o termo. Após a explicação e demonstração de exemplos de metadados em repositórios de dados do domínio, percebeu a importância dos metadados afirmando já fazer anotações sobre os dados, assim como utilizá-los nas pesquisas de dados de investigação, sobretudo no repositório FlyBase.

Face aos descritores apresentados, a investigadora validou 18 de 30 descritores (60 %), realizando comentários de melhoria a três deles.

A investigadora B não demonstrou ter grande conhecimento acerca de metadados e apenas se sentiu mais confortável após ter pedido para aceder ao FlyBase durante a sessão presencial.

6.1.3 Interações Epiteliais no Cancro

O principal objetivo do grupo de investigação de Interações Epiteliais no Cancro (em inglês Epithelial Interactions in Cancer) é descobrir como as junções de células epiteliais, assim como o seu microambiente circundante podem influenciar a progressão do cancro. Mais especificamente e baseado nos três cancros mais comuns (gástrico, da mama e colo-rectais), o grupo visa compreender a contribuição de moléculas de adesão (caderinas E e P), infeções (Helicobacter pylori e microbiota) e componentes neoplásicos do tecido tumoral para o desenvolvimento do cancro. Os investigadores deste grupo dividem-se em três equipas de trabalho e têm experiência em moléculas de adesão ao cancro e conhecimentos complementares em genética, biologia molecular e celular, microbiologia, patologia e oncologia.

Dentro dos elementos deste grupo, foram feitas entrevistas com duas investigadoras alunas de doutoramento. A **investigadora C** está envolvida no estudo do cancro da mama e cancro gástrico, pretendendo perceber quais os mecanismos que a bactéria *Helicobacter pylori* utiliza após infetar o estômago humano e de que forma estes contribuem para o desenvolvimento do cancro. Nesta fase, a investigadora estuda as diferentes moléculas das células, as alterações que a bactéria introduz nessas moléculas e os mecanismos genéticos e moleculares associados.

A **investigadora D,** por sua vez, procura perceber de que forma a matriz extracelular, um dos componentes do microambiente do tumor, influencia a seleção de células estaminais e de que forma estas podem ser responsáveis pelo início do tumor.

Ao nível do armazenamento e organização de dados, ambas as investigadoras utilizam os computadores dividindo os dados por pastas e projetos, no entanto a investigadora D sente-se mais confortável com a utilização do caderno de laboratório em papel contrariamente à investigadora C que tem preferência pelo digital.

Relativamente à familiarização com o conceito de metadados, a investigadora C afirmou não ser fácil entendê-lo, no entanto compreendeu o objetivo dos metadados, nomeadamente aqueles associados às imagens. Por sua vez, a investigadora D mostrou dúvidas e confundiu o conceito com base de dados.

Quanto à apresentação da tabela dos descritores, a investigadora C validou 17 dos 30 descritores (57%), dando sugestões de melhoria a 4 descritores. Além disso, sugeriu também 5 novos descritores que considera importantes para ajudar à interpretação dos seus dados de investigação. A investigadora D validou 16 descritores (53%) sendo que não fez comentários nem sugeriu novos descritores.

A investigadora C recomendou 5 descritores relacionados com ensaios clínicos já que é algo que a investigadora realiza frequentemente. Assim, sugeriu "Collection Site" referindo-se ao local de recolha de amostras clínicas dos pacientes, neste caso os nomes dos hospitais, e ainda "Clinical Trial Title", "Clinical Trial Phase" e "Clinical Trial Type" referindo-se ao título, fase e tipo de ensaio clínico, respetivamente. Além destes recomendou também uma "Intervention Description" para descrever o ensaio clínico, caso este seja de tipo interventivo.

6.1.4 Diversidade Genética

O grupo de investigação de Diversidade Genética tem como objetivo prioritário estabelecer uma ponte entre hipóteses, métodos e resultados a partir de uma abordagem teórica da diversidade da genética humana através do tempo e espaço, integrando essas informações e dados em contextos clínicos e forenses. O grupo tem vindo a examinar chips e sequências de todo o genoma humano em amostras de populações e em grupos de controlo de caso. Estes permitem avaliar a nível geral a evolução global humana, assim como os genes que contribuem para a suscetibilidade a doenças complexas.

No total, três investigadores deste grupo tiveram disponibilidade de colaborar nas entrevistas e nas sessões de apresentação do modelo de descritores de metadados. Os três destes investigadores **E, F** e **G** são alunos de doutoramento que, apesar de pertencerem ao

mesmo grupo de investigação, estão inseridos em diferentes projetos. A investigadora E tem vindo a analisar o exoma, microbioma e metaboloma de amostras africanas de Angola e Moçambique, tendo como principal meta comparar essas amostras com amostras de angolanos e moçambicanos a viver em Portugal, bem como com portugueses. O projeto do investigador F é focado no estudo do cancro gástrico e divide-se em duas fases. A primeira corresponde à exploração das diferenças no microbioma gástrico de tecidos normais e cancerígenos e a segunda à compreensão dos efeitos da infeção da bactéria *Helicobacter pylori* em diferentes linhas celulares de diferentes ancestralidades e a sua possível relação com o desenvolvimento do cancro gástrico. A investigadora G tem vindo a estudar os efeitos das infeções virais em genes de diferentes populações já que se descobriu, noutros estudos, que existem genes resistentes a alguns vírus. O principal objetivo é compreender os mecanismos da dengue, a fim de melhorar o seu diagnóstico e tratamento.

Todos os investigadores deste grupo armazenam os dados de investigação nos computadores ou em discos externos, no entanto como lidam com uma grande quantidade de dados têm dificuldades em armazená-los num único sítio. Por vezes, transcrevem ou imprimem aquilo que consideram mais relevante para o caderno de laboratório individual. A nível de organização dos dados, os investigadores tendem a dividi-los por pastas e projetos.

Quando questionados sobre o conceito de metadados, o investigador soube defini-lo corretamente como "dados sobre os dados" e informação descritiva associada a "documentos ou ficheiros". A investigadora E mostrou algumas dúvidas sobre o conceito de metadados, mas após explicação afirmou conhecê-lo apenas como "sample info ou assay info". Por sua vez, a investigadora G não definiu corretamente o conceito, confundindo-o com meta análise.

Relativamente à tabela dos descritores, a investigadora E compreendeu 21 descritores (70 %) realizando comentários a quatro deles, o investigador F 21 descritores (70%) com uma sugestão e a investigadora G 13 descritores (43%).

A Tabela 5 apresenta um sumário dos descritores validados pelos 7 investigadores.

Tabela 5 - Sumário geral dos descritores validados pelos investigadores

Investigadores	A	B	C	D	E	F	G
Descritores							
Age		X	X		X	X	X
Assay Type		X		X		X	
Cell Type	X					X	
Cell Line	X	X	X			X	
Collection Date		X	X	X	X	X	
Developmental Stage		X					
Disease	X	X	X	X	X	X	X
Drug Usage							X
Environmental Factor							
Ethnicity			X		X	X	X
Experimental Factor		X	X	X	X	X	X
Instrument Name	X	X	X	X	X	X	X
Instrument Type	X	X	X	X	X	X	
Material	X	X	X	X	X	X	X
Measurement		X			X	X	
Method	X	X	X	X	X	X	X
Molecule		X					
Organism	X	X	X	X	X	X	X
Organism Part	X	X	X	X	X	X	X
Reagent		Dentro dos Material			Dentro dos Material		
Sample Collection Protocol		X	X	X	Pode estar dentro de Method		
Sample Size	X			X	X	X	X
Sample Type			X	X	X		
Sex		X	X (Substituir por Gender)		X (Substituir por Gender)	X	X
Software	X		X	X		X	
Study Design				X	X	X	
Study Domain	X		X	X	X	X	X
Temperature							
Tissue	Dentro de Organism Part		Dentro de Organism Part	Dentro de Organism Part	Dentro de Organism Part	Dentro de Organism Part	
Treatment Protocol					Pode estar dentro de Method		

6.2 Validação de resultados

As ferramentas de gestão de dados de investigação são muito importantes na motivação dos investigadores para as tarefas de gestão de dados. Em geral, espera-se que as ferramentas que simplificam o trabalho necessário na gestão de dados e produzem resultados claros sejam mais facilmente adotadas. De forma a promover a plataforma Dendro como uma ferramenta adequada à gestão de dados de investigação, principalmente ao nível da descrição de dados, algumas experiências têm vindo a ser feitas com investigadores de diferentes domínios de investigação. Estas experiências pretendem avaliar a ferramenta ao nível da sua utilidade e usabilidade, assim como validar a ontologia desenvolvida para o domínio.

Após a ontologia MIBBIUP ser importada e carregada na plataforma Dendro, foram marcadas novas sessões com os investigadores de modo a que estes realizassem uma simulação de uma experiência de descrição de dados. Neste contexto, optou-se por não solicitar o depósito de dados no Dendro devido à utilização de uma máquina virtual existente num computador pessoal que obrigaria à transferência prévia dos dados. Nesta fase, quatro investigadores dos sete participantes, **C, D, E e F** tiveram disponibilidade em participar nas sessões de descrição de dados.

As sessões de experiências de interação com a plataforma Dendro seguiram um guião elaborado para o efeito (ANEXO II). Numa primeira etapa, foi criado um projeto exemplo com pastas e descritores preenchidos de modo a mostrar o resultado das tarefas de descrição aos investigadores. De seguida, foram criados projetos individuais durante as sessões com os investigadores.

No final questionou-se aos investigadores o que acharam da plataforma Dendro, as maiores dificuldades na sua utilização e as suas sugestões, quer para o Dendro, quer para a ontologia MIBBIUP.

Sessão 1 – 18/04/2019

A primeira sessão com a plataforma Dendro foi realizada com o investigador F do grupo de Diversidade Genética. Após observar a demonstração do funcionamento da plataforma Dendro, o investigador foi convidado a preencher descritores da ontologia MIBBIUP. Para esta tarefa, foi pedido ao investigador que fosse o mais realista possível a partir do projeto atual ou outro projeto em que já trabalhou.

Ao longo da interação com a plataforma, o investigador não mostrou dificuldades em selecionar ou preencher os descritores. Mostrou também compreender a divisão entre a descrição ao nível de pastas e ficheiros, apesar de não lhe ser solicitado descrever qualquer ficheiro. A sessão demorou cerca de 25 minutos. Relativamente à sessão prévia de validação dos descritores, o investigador preencheu todos os descritores que considerava válidos e úteis para o seu domínio de investigação. Por uma questão de organização, o investigador preferiu fazer uma divisão entre duas pastas, uma para as amostras e outra para os ensaios de laboratório.

Relativamente às perguntas feitas no final da experiência de interação com a plataforma Dendro, o investigador caraterizou o Dendro como "uma ferramenta que poderia ser útil para armazenar e partilhar dados". A nível de sugestões, este gostaria que fosse possível exportar os metadados em formato de tabela já que lhe facilitaria o processo de submeter os metadados e dados nos repositórios de dados utilizados pelo grupo de investigação, nomeadamente o European Genome Archive (EGA). Ao nível do vocabulário MIBBIUP, o investigador sugeriu quatros novos descritores, mais um que na primeira sessão, sendo estes o "Replicate Count", "Replicate Type", "Country of Origin" e "Sample Caractheristics".

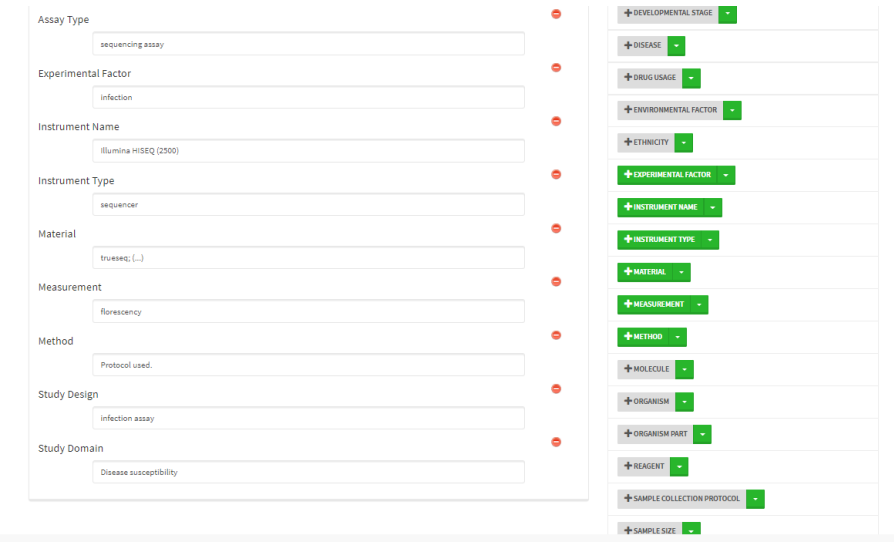


Figura 15- Exemplo de descrição do Investigador F

Sessão 2 – 9/05/2019

A segunda sessão com a plataforma Dendro foi realizada com a investigadora C do grupo de Interações Epiteliais no Cancro. Esta sessão demorou cerca de 30 minutos, no entanto o tempo total foi influenciado pela má conexão à rede que obrigou o investigador a preencher o mesmo

descritor duas vezes devido à falha no processo de gravar. Motivada por este contratempo, a investigadora sugeriu ter a opção de gravação automática sempre que se altera e adiciona um novo descritor.

Tal como o investigador F, a investigadora C também preferiu a divisão da descrição por pastas já que se sentia mais confortável em dividir os descritores entre as caraterísticas das amostras e os dados resultantes dos ensaios realizados em laboratório.

Excetuando o problema da ligação à rede, as tarefas de descrição ocorreram sem problema e a investigadora não demonstrou qualquer dificuldade em selecionar e preencher os descritores. Durante esta fase, a investigadora procurou usar alguns descritores mais que uma vez de forma a demonstrar que o mesmo descritor pode ser utilizado em duas situações diferentes. (exemplo: Software com "GraphPad v8 (Statistical Analysis)" e "IDEAS software v3" (Imaging Analysis).

Relativamente à primeira sessão de validação de conceitos, a investigadora preencheu todos os descritores que considerou úteis na primeira sessão, assim como sugeriu e preencheu os mesmos descritores que tinha sugerido na sessão prévia de apresentação dos descritores. Estes descritores sugeridos foram "Clinical Trial Description"; "Clinical Trial Phase", "Clinical Trial Type", "Time", "Collection Site" e "Region". Apesar de não sugerir o descritor à parte, foi possível inferir o descritor "Protocol Reference" como uma sugestão já que a investigadora referiu ser importante registar a referência de um determinado protocolo dentro de outro descritor, neste caso o "Sample Collection Protocol".

Relativamente às perguntas feitas no final da experiência de interação com a plataforma Dendro, a investigadora sugeriu mudar a cor da interface do Dendro de modo a que este se tornasse mais apelativo. Adicionalmente, também gostaria que o Dendro tivesse a opcionalidade de exportar os metadados em formato de documento de texto (.pdf ou .docx) de modo a poder imprimir e colar no caderno de laboratório em papel já que é algo que a investigadora já faz no seu dia-a-dia.

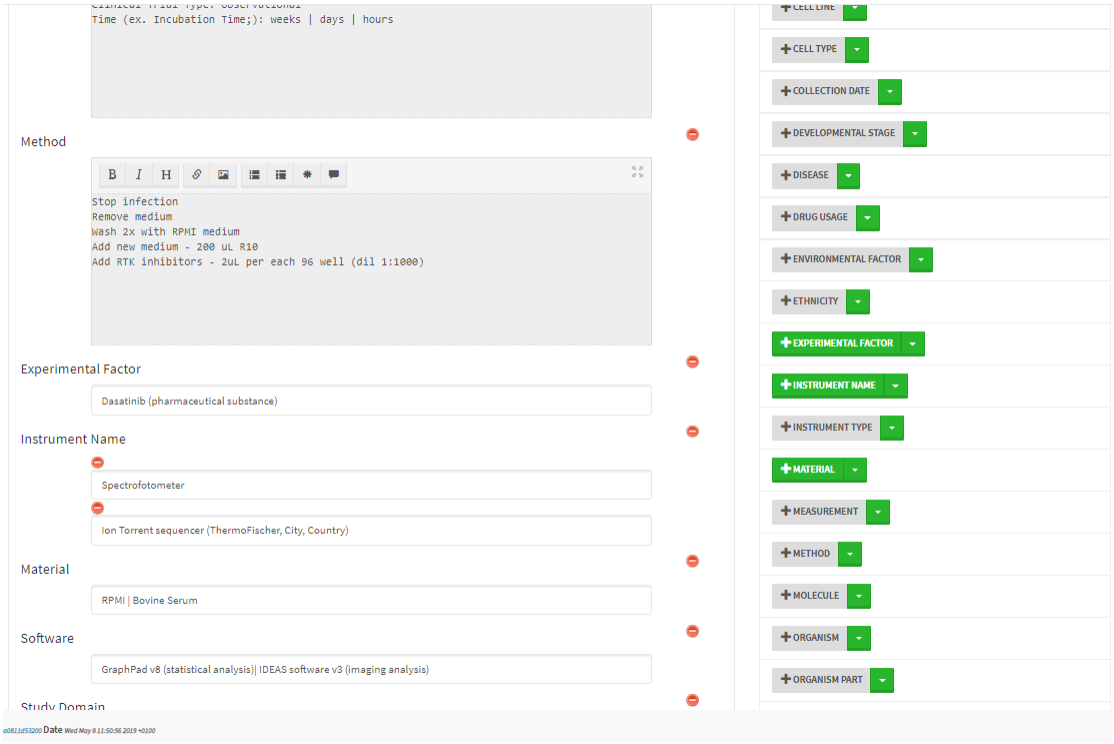


Figura 16 - Exemplo de descrição da investigadora C

Sessão 3 e 4 – 16/05/2019

A terceira e quarta sessão foram realizadas com a investigadora D do grupo Interações Epiteliais no Cancro e a Investigadora E do grupo de Diversidade Genética, respetivamente. Estas sessões ocorreram no mesmo dia, com um intervalo de tempo de uma hora.

Depois de explicado o processo de adicionar descritores, a investigadora D foi convidada a realizar uma experiência de descrição a partir da ontologia MIBBIUP. A sessão demorou aproximadamente 27 minutos com 21 descritores a terem sido preenchidos. Relativamente à primeira sessão de apresentação dos descritores, foram preenchidos mais 5 descritores do que aqueles considerados úteis numa primeira fase. Isto pode significar que a investigadora, no momento das tarefas de descrição, conseguiu compreender melhor o âmbito dos descritores.

É importante salientar que alguns descritores relativos aos instrumentos, materiais e tipos de ensaio foram utilizados mais que uma vez. Nesta fase, a investigadora tentou passar a ideia de que o mesmo descritor com valores diferentes pode ser usado para descrever diferentes ficheiros ou pastas.

No final a investigadora admitiu ter tido dificuldades na seleção dos descritores, principalmente aqueles que usou mais que uma vez que a obrigaram a percorrer a lista total de descritores novamente. Assim, sugeriu que o comando "*Add*" ao lado do descritor que pretende

repetir se tornasse operacional. Além disso, sugeriu também ter a liberdade de adicionar descritores ao longo das interações com a plataforma. Por fim, classificou o Dendro como uma ferramenta *user-friendly* e com uma boa interface para o armazenamento e organização de dados.

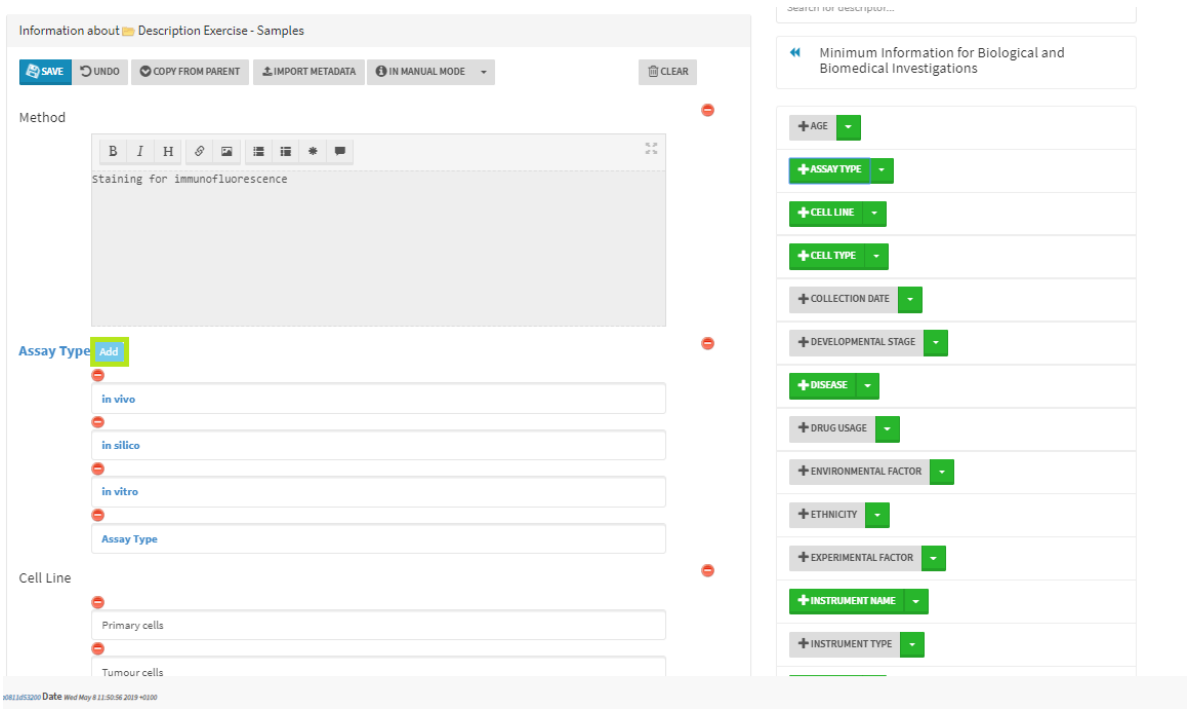


Figura 17 - Exemplo de descrição da Investigadora D

A última sessão de interação com a plataforma Dendro foi feita com a investigadora E, demorando cerca de 12 minutos com 18 descritores preenchidos. Estes descritores preenchidos foram os mesmos que foram validados pela investigadora na primeira sessão de apresentação dos descritores. A sessão correu sem problemas, porém a investigadora mencionou que a navegação pelos descritores não era muito prática e sugeriu ser o próprio investigador a definir livremente os seus próprios descritores de acordo com a experiência realizada.

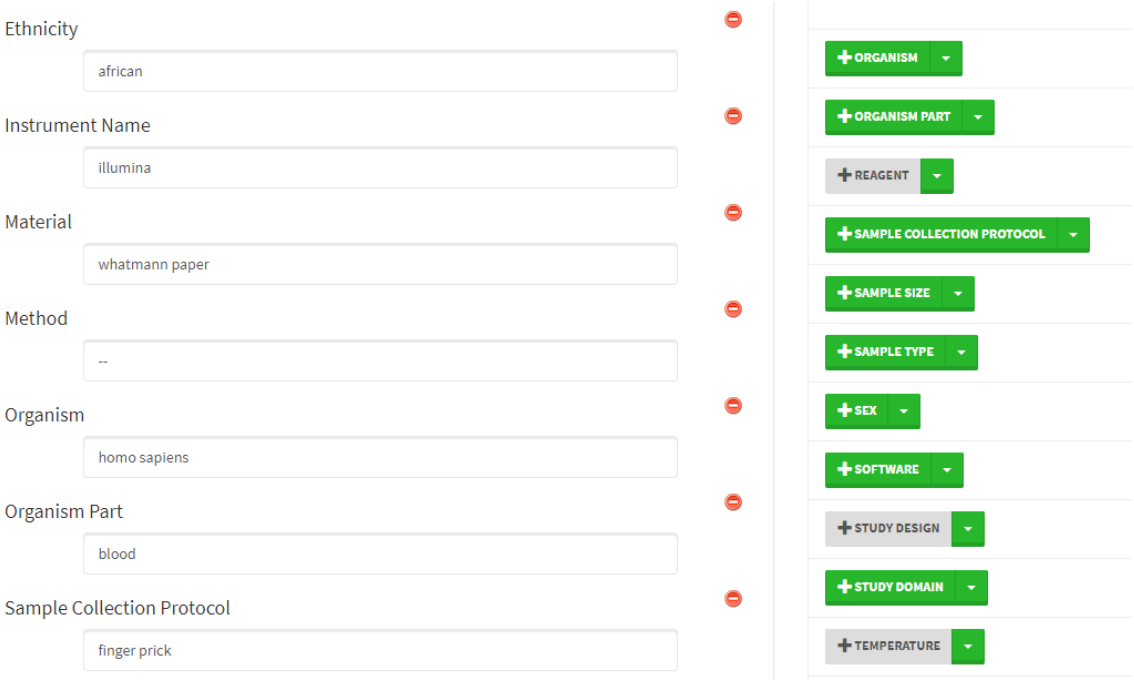


Figura 18- Exemplo de descrição do investigador E

6.2.1 Comparação da utilização de descritores e valores por investigador

Neste capítulo é exemplificado, a partir de tabelas, a comparação entre os diferentes investigadores ao nível dos descritores de metadados utilizados e os valores introduzidos. Com esta comparação pretende-se observar se existiram grandes diferenças entre os valores e descritores utilizados pelos investigadores do mesmo grupo de investigação e de seguida entre os grupos de investigação.

Nas tabelas 6, 7 e 8 encontram-se os valores dos descritores utilizados por investigadores sendo que se procedeu a uma divisão dos 30 descritores por categoria de modo a facilitar a leitura e a discussão.

Tabela 6 - Comparação dos valores dos descritores da categoria Amostra

Categoria	Descritores	Investigador F	Investigador E	Investigador C	Investigador D
Amostra	**Age**	28 years old	50	37 years old | 1 week	X
	Cell Line	MKN 28	X	ATCC ® CRL - 1593.2	Primary Cells
	Cell Type	stomach	human		stem cells
	Developmental Stage		adult		
	Disease	gastric carcinoma	hypertension	gastric cancer	cancer
	Ethnicity	asian	african	caucasian	
	Molecule				E- cadherin
	Organism	homo sapiens	homo sapiens	homo sapiens	human
	Organism Part	stomach	blood	stomach	gut
	Sample Type		DNA		cell lysaes
	Sample Size	321	50	N=50	1 cm2
	Sex	male	female	female	
	Tissue				Epithelial tissue

Dos 13 descritores da categoria da Amostra, todos foram preenchidos pelo menos uma vez, com 3 descritores a terem sido preenchidos por todos os investigadores e 5 pelo menos três vezes.

Os descritores “Organism”, “Organism Part” e “Sample Size” foram preenchidos por todos os investigadores. Relativamente aos valores do descritor “Organism”, foi observado que existiram diferenças na escrita entre os investigadores. O investigador D escreveu o termo do “Organism” em inglês, contrariamente aos outros investigadores que seguiram a taxonomia NCBI e escreveram-no em latim. O descritor “Sample Size” também provocou alguma confusão, uma vez que o investigador D escreveu o tamanho de um indivíduo de uma amostra (neste caso uma célula), enquanto os outros investigadores escreveram o número total de indivíduos de uma amostra. Neste sentido, Count e Size seriam dois descritores diferentes. No caso do descritor “Disease”, existe uma maior especificação no valor dos investigadores F e C em relação a E.

Os investigadores F e E, ambos do grupo de Diversidade Genética, preencheram todos os atributos relativos aos organismos da amostra através de “Age”, “Ethnicity” e “Sex”. Ambos os investigadores recomendaram ainda mais descritores para a caraterização do organismo da amostra com “Sample Identifier” “Country of Origin” e “Sample Characteristics”. O

investigador E referiu ainda que o descritor "Sample Type" seria genérico, mas suficiente para incluir tipo de moléculas como DNA.

Tabela 7 - Comparação dos valores dos descritores das categorias de Métodos e Materiais

Categoria	Descritores	Investigador F	Investigador E	Investigador C	Investigador D
Métodos e Materiais	**Assay Type**	Sequencing assay	sequenciação		in vitro assay
	Collection Date	05/08/2018	16.11.2018	27/02/2019	
	Drug Usage				
	Material	trueseq	whatmann paper	RPMI | Bovine Serum	collection tubes
	Measurement	florescency			cell viability
	Method	Protocol REF	Protocol REF	Stop infection Remove medium Wash 2x with RPMI medium Add new medium - 200 uL R10 Add RTK inhibitors - 2uL per each 96 well (dil 1:1000)	Staining for immunofluorescence
	Reagent				
	Sample Collection Protocol		finger prick	Protocol REF Identifier name or number	biopsy
	Study Design	infection assay			Prospective study
	Temperature				
	Treatment Protocol				Incubation with antibodies

Ao nível da categoria dos métodos e materiais, três descritores que não foram utilizados por nenhum investigador, no entanto, segundo os investigadores, os reagentes e drogas utilizados são considerados materiais e por isso podem ser referidos no descritor "Material". Os descritores mais genéricos "Method" e "Material" foram preenchidos por todos os investigadores, apesar de os valores terem diferenças entre si. No descritor "Method", os investigadores F e E referiram que escrever um protocolo seria algo moroso, por isso afirmaram que bastaria referi-lo no caso de ter uma referência *online*, ou então anexá-lo como ficheiro junto aos dados. Por sua vez, o investigador C escreveu uma parte de um protocolo de um ensaio, enquanto o investigador E foi mais genérico e apenas referiu o título do método. O descritor "Material" causou, igualmente, algumas diferenças entre os investigadores já que três investigadores descreveram o material utilizado como ferramentas auxiliares (tubos de ensaio

e papel *whatmann*), enquanto o investigador C descreveu-o com exemplos de reagentes químicos utilizados nos protocolos, não referindo, porém, a quantidade nem unidade utilizada.

Relativamente ao descritor "Assay Type" e "Study Design", os investigadores que o utilizaram preencheram-no com termos já existentes em ontologias desenvolvidas para o domínio, sobretudo da OBI.

Tabela 8 - Comparação entre os valores dos descritores das categorias de Tecnologia e Outros

Categoria	Descritores	Investigador F	Investigador E	Investigador C	Investigador D
Tecnologia e Outros	**Experimental Factor**	infection		Dasatinib (pharmaceutical substance)	
	Environmental Factor				
	Instrument Name	Illumina HISEQ (2500)	Illumina	Ion Torrent sequencer (ThermoFischer, City, Country)	Flow cytometer, microscope
	Instrument Type	sequencer	sequencer		
	Software	GraphPad	Sequencher	GraphPad v8 (statistical analysis) | IDEAS software v3 (imaging analysis)	FlowJo, AxioVision, GraphPad
	Study Domain	Disease susceptibility	Genetic Diversity	Oncology	Stem cells and cancer

Os descritores relativos à instrumentação utilizada para produzir ou analisar os dados foram preenchidos, apesar de apresentarem diferenças ao nível da sua escrita. Por exemplo, ao referir o nome do instrumento utilizado, os investigadores F e E nomearam o mesmo instrumento, mas o investigador F optou por registar a sua versão. Por outro lado, o investigador C apontou que era importante registar o fabricante e local de origem do instrumento, enquanto o investigador D referiu apenas o tipo de instrumento utilizado. A nível de *software* utilizado, apenas o investigador C referiu a sua versão e o objetivo para o qual foi utilizado. Todos os investigadores referiram o domínio do seu estudo, mas com maior especificidade nos casos F e D.

O descritor "Experimental Factor" causou dúvidas aos investigadores, mas, após explicação, compreenderam-no como variável de um estudo. Desta forma, os investigadores C e F referiram 2 valores que se enquadram no descritor.

No domínio das ciências biológicas e biomédicas, muito dos esquemas normalizados de metadados e repositórios requerem a utilização de termos controlados de ontologias nos valores dos metadados. Neste estudo optou-se por não criar vocabulários controlados para os valores dos descritores introduzidos na ontologia MIBBIUP por duas razões – 1) grande número de valores possíveis por descritor 2) deixar o investigador preencher livremente - de modo a ser possível comparar os diferentes valores entre investigadores e depois com os valores existentes nos repositórios.

De uma forma geral, os valores introduzidos pelos quatro investigadores não variam muito entre si, havendo casos em que os valores dos descritores “Study Design”, “Organism” e “Assay Type” correspondem aos termos controlados de taxonomias e ontologias.

7. Conclusões

As tecnologias digitais e a forma como elas evoluem levantam problemas para a gestão de dados de investigação, mas também trazem novas oportunidades para a comunidade científica. O estado atual da ciência é amplamente influenciado pela disponibilidade dos dados de investigação sendo que impõe bastantes desafios para os diferentes intervenientes envolvidos na gestão desses dados, no entanto é importante referir as vantagens resultantes do acesso e partilha dos dados de investigação a todos os envolvidos neste processo, principalmente aos investigadores.

De modo a atingir o objetivo da partilha e reutilização dos dados, a utilização dos metadados associados aos dados revela-se fundamental. A descrição de dados é benéfica para registar que dados foram recolhidos e estão disponíveis, mas também pode fornecer informação sobre o contexto dos dados e referir problemas que surgiram durante a etapa de investigação.

Ao longo deste trabalho colaborou-se com um grupo de investigadores com o objetivo de os integrar e consciencializar para as tarefas de gestão de dados de investigação, focando principalmente na questão da descrição dos dados. Foi desenvolvida uma ontologia para o domínio das ciências biomédicas com descritores provenientes de modelos de metadados MIBBI de modo a que os investigadores participantes neste estudo realizassem experiências de descrição de dados a partir da plataforma Dendro. Uma visão geral sobre os resultados obtidos mostra que os investigadores compreenderam mais de metade dos descritores e estão dispostos a descrever os seus dados de investigação se tiverem todas as ferramentas necessárias para o fazer. É possível afirmar que a realização deste trabalho e os contactos presenciais contribuíram para a consciencialização dos investigadores para a necessidade de descrever os dados por eles produzidos, assim como para compreender o seu potencial de reutilização.

7.1 Limitações do estudo

Existiram alguns fatores limitativos à realização deste estudo. Primeiramente, o tempo para a realização deste projeto de dissertação foi de apenas 6 meses pelo que os contactos presenciais com os investigadores apenas puderam ocorrer entre março e maio, período no qual os investigadores tiveram mais disponibilidade de participar nas sessões. De modo a agilizar o processo de exportação para o Dendro, a ontologia foi desenvolvida no período anterior aos contactos com os investigadores e por essa razão foi necessário um estudo autónomo dos domínios dos investigadores através dos repositórios de dados de investigação e esquemas de metadados existentes.

Para além disso, este estudo teve de levar em conta 1) a diversidade de requisitos dos investigadores do I3S e 2) que os investigadores não estão conscientes acerca de metadados e que a descrição de dados de investigação ainda não é uma prioridade para a maioria. Consequentemente, optou-se por criar um modelo de metadados considerado genérico com pouca especificidade e com um pequeno número de descritores de modo a limitar a participação dos investigadores. A maior dificuldade, neste período, passou por compreender e entender o quão genérico seria um descritor neste domínio.

Já após os contactos com os investigadores, alguns referiram que, mesmo dentro de um grupo de investigação, realizam diferentes tipos de experiências e por isso necessitam de diferentes metadados de acordo com a experiência realizada. O projeto de esquemas de metadados MIBBI tem como objetivo fornecer listas de metadados mínimos para as diversas experiências neste domínio, porém, estas são extensivas e em grande número pelo que obrigaria a apresentar um número bastante elevado de descritores aos investigadores.

Outras limitações ao estudo foram os problemas técnicos que o sistema Dendro apresentou ao longo do tempo do projeto de dissertação que, por sua vez, influenciou a importação da ontologia e as experiências de descrição de dados com os investigadores.

7.2 Investigação futura

O principal objetivo desta linha de trabalho em gestão de dados de investigação é aumentar o conhecimento dos investigadores dos mais variados domínios de investigação e tipo de dados produzidos. Assim, é importante continuar a colaborar com investigadores de diferentes domínios e grupos de investigação de modo a tornar a gestão de dados de investigação num tema cada vez mais discutido.

Como perspetiva de trabalho futuro, será importante utilizar as experiências realizadas na plataforma Dendro e o feedback dos investigadores como uma oportunidade de se realizar alterações à ontologia e Dendro de modo a que esta vá de encontro às necessidades dos investigadores. Além disso, a possibilidade de continuar a testar a ontologia com outros investigadores da área das ciências biomédicas permitirá continuar a levantar requisitos, assim como refinar a ontologia de modo a que se enquadre cada vez mais na descrição de dados de investigação destes domínios. Outro caminho alternativo a tomar seria dividir a ontologia e torná-la mais especializada para uma determinada experiência.

Durante este período do projeto de dissertação, foi também escrito um artigo científico[31] cuja parte teórica aborda os metadados utilizados nas ciências biológicas e biomédicas,

[31] Artigo com o título "Training biomedical researchers in metadata with a MIBBI-based ontology" submetido para publicação

enquanto a parte prática aborda o desenvolvimento da ontologia para o Dendro e os resultados das entrevistas e experiências de descrição com os investigadores. O principal contributo do artigo é demonstrar que é possível motivar os investigadores para as tarefas de descrição de dados de investigação desde que estes tenham as ferramentas necessárias para isso.

Referências bibliográficas

Amorim, Ricardo Carvalho. 2014. "LabTablet: A Multi-Domain Laboratory Book," Dissertação de mestrado, Universidade do Porto. 106. https://repositorioaberto.up.pt/bitstream/10216/73291/2/31847.pdf

Amorim, Ricardo Carvalho, João Aguiar Castro, João Rocha da Silva, e Cristina Ribeiro. 2014: "LabTablet: semantic metadata collection on a multi-domain laboratory notebook". In: Springer Communications in Computer and Information Science, vol. 478, 193–205

Amorim, Ricardo Carvalho, João Aguiar Castro, João Rocha da Silva, e Cristina; Ribeiro. 2015. "Engaging Researchers in Data Management with LabTablet, an Electronic Laboratory Notebook." *Languages, Applications and Technologies*, 216–21. https://doi.org/10.1007/978-3-319-27653-3_21 .

Amorim, Ricardo Carvalho, João Aguiar Castro, João Rocha da Silva, e Cristina Ribeiro. 2017. "A Comparison of Research Data Management Platforms: Architecture, Flexible Metadata and Interoperability." *Universal Access in the Information Society* 16 (4):851–62. https://doi.org/10.1007/s10209-016-0475-y .

Bandrowski, Anita, Ryan Brinkman, Mathias Brochhausen, Matthew H Brush, Bill Bug, Marcus C Chibucos, Kevin Clancy, et al. 2016. "The Ontology for Biomedical Investigations," 1–19. https://doi.org/10.1371/journal.pone.0154556 .

Borgman, Christine. 2012. "Advances in Information Science: The Conundrum of Sharing Research Data." *Journal of the American Society for Information Science and Technology* 63 (6):1059–78. https://doi.org/10.1002/asi.

Brazma, Alvis, Pascal Hingamp, John Quackenbush, Gavin Sherlock, Paul Spellman, Chris Stoeckert, John Aach, et al. 2001. "Minimum Information about a Microarray Experiment (MIAME)— toward Standards for Microarray Data." *Nature Genetics* 29 (December): 365–71.

Bruce, Thomas R., e Diane Hillmann. 2004. "The Continuum of Metadata Quality: Defining, Expressing, Exploiting," in *Metadata in Practice*.

Castro, João Aguiar, Cristina Ribeiro, e João Rocha da Silva. 2014. "Creating Lightweight Ontologies for Dataset Description: Pratical Applications in a Cross-Domain Research Data Management Workflow." *IEEE/ACM Joint Conference on Digital Libraries*, no. September 2015: 313–16. https://doi.org/10.1109/JCDL.2014.6970185 .

Castro, João Aguiar, Ricardo Carvalho Amorim, Rúbia Gattelli, Yulia Karimova, João Rocha da Silva e Cristina Ribeiro. 2017. "Involving Data Creators in an Ontology-Based Design Process for Metadata Models." *Developing Metadata Application Profiles*, 181–214. https://doi.org/10.4018/978-1-5225-2221-8.ch008 .

Consultative Committee for Space Data Systems. 2012. "Reference Model for an Open Archival Information System (OAIS)." *Recommendation for Space Data System Practices*, no. Recommended Practice, Issue 2: 1–148. https://doi.org/10.1081/E-ELIS3-120044377 .

Costa, Michelli, e Tiago Braga. 2016. "Repositórios de Dados de Pesquisa No Mundo." *Cadernos BAD* 0 (2):80–95. https://www.bad.pt/publicacoes/index.php/cadernos/article/view/1585 .

Directorate-General for Research and Innovation (European Comission). 2018. "Turning FAIR into reality". Relatório técnico. pp 1-78. doi:10.2777/1524

Ferreira, Ana Luís. 2019. "Aplicação da ferramenta LabTablet em contexto laboratorial: Caso de estudo I3S". Dissertação de mestrado, Universidade do Porto.

González-Beltrán, Alejandra, Eamonn Maguire, Susanna-Assunta Sansone, e Philippe Rocca serra. 2014. "LinkedISA : Semantic Representation of ISA-Tab Experimental Metadata" 15 (Suppl 14): 1–15.

Gonçalves, Rafael S, e Mark A. Musen. 2019. "The Variable Quality of Metadata about Biological Samples Used in Biomedical Experiments." *Nature Publishing Group*, no. August 2018: 1–15. https://doi.org/10.1038/sdata.2019.21.

Hoehndorf, Robert, Paul N Schofield, e Georgios V. Gkoutos. 2015. "The Role of Ontologies in Biological and Biomedical Research: A Functional Perspective." *Briefings in Bioinformatics,* 16 (November 2014): 1069–80. https://doi.org/10.1093/bib/bbv011.

Kanza, Samantha, Cerys Willoughby, Nicholas Gibbins, Richard Whitby, Jeremy Graham Frey, Jana Erjavec, Klemen Zupančič, Matjaž Hren, e Katarina Kovač. 2017. "Electronic Lab Notebooks: Can They Replace Paper?" *Journal of Cheminformatics* 9 (1). Springer International Publishing:1–15. https://doi.org/10.1186/s13321-017-0221-3 .

Kennan, Mary Anne, and Lina Markauskaite. 2015. "Research Data Management Practices: A Snapshot in Time." *International Journal of Digital Curation* 10 (2):69–95. https://doi.org/10.2218/ijdc.v10i2.329.

Lide, David R. 1981. "Critical data for critical needs" Science, 19, 1343–1349.

Lyon, Liz. 2007. Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report. UKOLN, (June):1–65, 2011.

Malone, James, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. 2010. "Modeling Sample Variables with an Experimental Factor Ontology." *BioInformatics* 26 (8): 1112–18. https://doi.org/10.1093/bioinformatics/btq099 .

Mark D. Wilkinson, Michel Dumontier, Jsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Jun Zhao, and Barend Mons. 2016. "Comment: The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Nature* 3 (160018): 1–9. https://doi.org/10.1038/sdata.2016.18 .

McQuilton, Peter, Alejandra Gonzalez-Beltran, Philippe Rocca-Serra, Milo Thurston, Allyson Lister, Eamonn Maguire, e Susanna Assunta Sansone. 2016. "BioSharing: Curated and Crowd-Sourced Metadata Standards, Databases and Data Policies in the Life Sciences." *Database: The Journal of Biological Databases and Curation* 2016:1–8. https://doi.org/10.1093/database/baw075 .

Noy, Natalya F, e Deborah L McGuiness. 2000. "Ontology Development 101: A Guide to Creating Your First Ontology," 1–25.

NSF. 2005. «Long-lived Digital Data Collections: Enabling Research and Education in the 21st Century». http://www.nsf.gov/nsb/documents/2005/LLDDC_report.pdf.

Qin, Jian, Alex Ball, e Jane Greenberg. 2012. "Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data." *Proceedings International Conference on Dublin Core and Metadata Applications*, 62–71. https://doi.org/10.1007/s00799-013-0106-7 .

Read, Kevin. "Common Metadata Elements for Cataloguing Biomedical Datasets". figshare, July 29, 2015. https://doi.org/10.6084/m9.figshare.1496573.v1

Ribeiro, Cristina, Ricardo Carvalho Amorim, João Rocha da Silva, João Aguiar Castro e João Correia Lopes. 2016. "Projeto TAIL — Gestão de Dados de Investigação Da Produção Ao Depósito e à Partilha. *Cadernos BAD* 2: 256–64. https://www.bad.pt/publicacoes/index.php/cadernos/article/view/1603/pdf

Riley, Jenn. 2017. *Understanding Metadata: What Is Metadata, and What is it for? NISO Primer*. National Information Standards Organization (NISO). https://doi.org/10.1017/S0003055403000534 .

Sansone, Susanna Assunta; McQuilton, Peter; Roca-Serra, Philippe; Beltran-Gonzalez, Alejandra; Izzo, Izzo, Massimiliano, Lister, Allyson L. e Thurston, Milo. 2019. "FAIRsharing as a Community Approach to Standards, Repositories and Policies." *Nature Biotechnology* 37:350–69.

Silva, João Rocha da. 2016 "Usage-Driven Application Profile Generation Using Ontologies," Tese de doutoramento, Universidade do Porto. 211 https://repositorio-aberto.up.pt/handle/10216/83993

Taylor, Chris F, Dawn Field, Susanna-Assunta Sansone, Jan Aerts, Michael Ashburner, Catherine A Ball, Pierre-Alain Binz, et al. 2009. "Promoting Coherent Minimum Reporting Guidelines for Biological and Biomedical Investigations: The MIBBI Project." *Nature Biotechnology* 26 (8): 889–96. https://doi.org/doi:10.1038/nbt.1411 .

Taylor, Chris F, Norman W Paton, Kathryn S Lilley, Pierre-Alain Binz, Randall K Julian Jr, R Andrew, Weimin Zhu, et al. 2007. "The Minimum Information about a Proteomics Experiment (MIAPE)." *Nature Biotechnology* 25 (8): 887–93. https://doi.org/10.1038/1329.

Van den Eynden, Veerle, Louise Corti, Matthew Woollard, Libby Bishop, e Laurence Horton. 2011. *Managing and Sharing Data: A Guide to Good Practice. UK Data Archive.* https://doi.org/10.1016/0370-2693(94)91481-8.

Walport, Mark, e Paul Brest. 2011. Sharing research data to improve public health. The Lancet, v. 377, n. 9765, p. 537–539. https://doi.org/10.1016/S0140-6736(10)62234-9

Whyte, Angus, e Jonathan White. 2011. 'Making the Case for Research Data Management'. DCC Briefing Papers. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/briefing-papers/making-case-rdm

Willis, Craig, Jane. Greenberg, e Hollie White. 2012. “Analysis and Synthesis of Metadata Goals for Scientific Data.” *Journal of the American Society for Information Science and Technology* 63 (8):1505–20. https://doi.org/10.1002/asi .

ANEXOS

ANEXO I. Guião de Entrevista aos investigadores do I3S [32]

Guião de Entrevista

Investigador:
Data:

Esta entrevista enquadra-se na minha dissertação de mestrado e, por isso, o seu conteúdo servirá como material de apoio para tal. O objetivo principal é recolher informação acerca das práticas de produção de dados, gestão de dados e partilha de dados adotadas pelos investigadores em contexto laboratorial, durante o processo de investigação.

- Posso contar com a sua colaboração?

- A sua identidade apenas será divulgada (em trabalhos ou publicações resultantes deste projeto) se for do seu interesse e assim o permitir. Gostaria de fazer a sua participação anónima?

- Para efeitos de transcrição, será útil proceder à gravação áudio desta entrevista. A gravação será apagada assim que for transcrita. Autoriza a gravação?

- Tem alguma questão inicial?

Demografia

1. Qual o seu título profissional?

2. Qual o domínio de investigação no qual trabalha?

3. Apresente, brevemente, o projeto de investigação em que está envolvida/o (nome, domínio, objetivo, tipo de dados e fase em que se encontra).

4. Com que frequência contacta com dados de investigação?

Organização de dados

5. Onde são armazenados os dados que produz e de que forma os organiza?

6. Considera o método de organização que utiliza eficiente? Porquê?

7. Teve algum tipo de treino para a gestão de dados?

[32] Guião de entrevista elaborado por Ana Luís Ferreira do mestrado em Bioengenharia 2018/2019

8. Há alguém, dentro ou fora do grupo de investigação, a quem recorra para trocar ideias sobre organização de dados?

9. Quando pretende aceder a um conjunto de dados que já tenha armazenado, utiliza alguma estratégia para os localizar facilmente?

10. Identifica algum problema relacionado com a organização e pesquisa dos dados?

11. Tem algum plano para o armazenamento dos dados a longo prazo?

Descrição de dados

12. Tem o hábito de fazer anotações acerca dos dados?

-Se sim, de que forma o faz?

13. Utiliza alguma ferramenta de apoio à anotação dos dados que produz? Qual?

14. Está familiarizada/o com o conceito de metadados?

- Se sim: o que entende por metadados?

- Se não: explico -> Metadados é o termo utilizado para nos referirmos aos "dados sobre os dados", são uma forma de descrever os dados. Por exemplo, quando uma amostra deve repousar à temperatura ambiente, o valor da temperatura que anotamos é um metadado OU o Mendeley pede-nos para inserirmos título, autor e data sobre os artigos - isso são metadados.

15. Os dados que produz são, normalmente, acompanhados de metadados?

- Se sim, de que forma faz esta associação? Segue uma metodologia específica?

16. Tem por hábito utilizar dispositivos eletrónicos/digitais? Considera útil a utilização de uma ferramenta digital para a recolha de metadados/anotações acerca dos dados? De que forma poderia beneficiar o seu trabalho e quais os requisitos para essa ferramenta?

17. Considera a anotação dos dados uma prática relevante? Em que sentido?

18. Acha que apenas através dos dados e informação que guarda acerca dos mesmos, estes seriam facilmente interpretados e utilizados por alguém externo ao projeto? E no caso de ser alguém envolvido no projeto, daqui a 2 anos?

Partilha, reutilização, publicação e repositórios

20. Trabalha em equipa: como é coordenada a partilha de dados entre elementos do grupo de investigação (se esta for necessária)?

21. Já alguma vez partilhou dados de um projeto? Se sim, qual o motivo?

22. Os dados deste projeto são partilhados ou existe interesse em partilhá-los com pessoas externas ao grupo de investigação?

-Se sim, como e porquê?

-Se não, qual o motivo?

23. Considera que os dados com que trabalha têm potencial de reutilização para si (projetos em que está envolvida/o) ou para projetos de outros domínios? De que forma?

24. Já explorou dados depositados em algum repositório?

25. Utiliza ou já utilizou dados (*raw data*) criados por outra pessoa ou instituição, por exemplo, publicados num repositório?

-Se sim, foi uma experiência positiva ou negativa? Porquê?

26. Vê alguma vantagem na partilha de dados resultantes da sua investigação?

-Se sim, quais?

27. Estaria interessado/a em ter acesso aos dados de algum projeto de investigação no qual não tenha participado? Isso traria alguma vantagem para algum projeto em que esteja envolvida/o?

ANEXO II. Guião para as experiências de interação com a plataforma Dendro

Localização da máquina da experiência	Credenciais de acesso aos utilizadores
Utilização da Máquina Virtual Ubuntu Server 18.04.2 (64 bit) e aceder ao servidor http://192.168.56.101:3001/	Utilizador: up2014***** Password: *********

Demonstração de funcionamento do Dendro

Para esta fase inicial foi criado um projeto de demonstração dentro da conta do utilizador. Este projeto tem no seu interior duas pastas com um grupo de ficheiros no seu interior, assim como uma lista de metadados preenchidos associados às pastas e a cada um dos ficheiros. O principal objetivo passou por demonstrar o resultado das tarefas de descrição.

Registo dos utilizadores

Para acelerar o processo, todos os projetos foram criados dentro de uma conta de utilizador (neste caso a minha). Desta forma, não foi necessário criar diferentes contas para os diferentes utilizadores.

Criação de projetos

Cada projeto criado corresponde a cada investigador participante nas experiências. O nome do projeto será "Researcher (X)" e a descrição será "Small data description exercise with a (laboratory´s name) researcher". Dentro de cada projeto, o investigador procederá à simulação de uma descrição de dados a partir da utilização da ontologia "MIBBI". Não foi pedido a nenhum investigador que depositasse um "dataset".

Guião

1. Dar a noção do conceito de descritor/metadado, caso seja necessário;
2. Apresentar sumariamente os objetivos da avaliação destas tarefas;
3. Demonstrar o funcionamento da plataforma Dendro

a. Criação de projetos e sua administração (adicionar colaboradores, descrever projeto, etc.)
b. Layout da interface, criação de pastas e *upload* de ficheiros;
c. Navegação pelas diferentes listas de descritores;
d. Seleção e preenchimento de descritores;

4. Apresentar novamente os descritores da ontologia "MIBBIUP"
5. Pedir aos investigadores que façam uma simulação de uma experiência de descrição de dados, apenas utilizando a ontologia MIBBIUP;
6. Deixar os investigadores à vontade durante a interação com a plataforma;
7. Registar e responder a todas as dúvidas que possam surgir;
8. Cronometrar a experiência;

Perguntas finais

O que achou da interface da plataforma Dendro?
Quais as maiores dificuldades durante a utilização do Dendro?
Que sugestões tem para a plataforma Dendro?
Sugere mais algum descritor necessário para a interpretação dos seus dados de investigação?