U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

MESTRADO EM CIÊNCIA DA INFORMAÇÃO

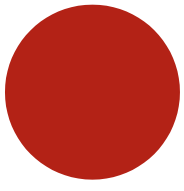
Avaliação da migração de registos de arquivo para dados ligados no projeto EPISA

Margarida Gouveia Augusto

M
2022

UNIDADES ORGÂNICAS ENVOLVIDAS

FACULDADE DE ENGENHARIA
FACULDADE DE LETRAS



Margarida Gouveia Augusto

Avaliação da migração de registos de arquivo para dados ligados no projeto EPISA

Dissertação realizada no âmbito do Mestrado em Ciência da Informação, orientada pelo Professor Gabriel David e pela Professora Cristina Ribeiro

Faculdade de Engenharia e Faculdade de Letras

Universidade do Porto

julho de 2022

Avaliação da migração de registos de arquivo para dados ligados no projeto EPISA

Margarida Gouveia Augusto

Dissertação realizada no âmbito do Mestrado em Ciência da Informação, orientada pelo Professor Gabriel David e pela Professora Cristina Ribeiro

Membros do Júri

Professor António Lucas Soares

Faculdade de Engenharia - Universidade do Porto

Professora Maria Manuel Borges

Faculdade de Letras - Universidade de Coimbra

Professor Gabriel David

Faculdade de Engenharia - Universidade do Porto

Agradecimentos

Aos meus pais, Sílvia e Carlos, por me apoiarem em todas as decisões e por serem os meus pilares. Obrigada por me ensinarem a ser resiliente e a nunca desistir.

À minha irmã Carolina e ao meu cunhado André, por todo o apoio incondicional. Queria conseguir expressar por palavras o orgulho que tenho neles, por serem para mim um exemplo de esforço e trabalho.

Ao David… pela imensa paciência, pelo companheirismo ao longo dos últimos cinco anos. Por sempre me apoiar nos momentos mais difíceis e por ter tempo para me ouvir.

Ao meu avô António pelas cantorias depois de almoço.

À minha avó Rosalina por sempre ter cuidado de mim.

Por fim, mas não menos importante, à Professora Cristina Ribeiro e ao Professor Gabriel David por se disponibilizaram para me ajudar no que fosse necessário ao longo dos últimos meses.

Resumo

O Arquivo Nacional da Torre do Tombo possui documentos originais desde o século IX até à atualidade e, através da utilização de metadados, é possível descrever os objetos culturais que possui. Utiliza, também, as normas ISAD (G) e ISAAR (CPF) para descrição arquivística. O Arquivo tem o seu sistema de descrição online há vinte anos e, recentemente, está a considerar a evolução para um sistema que favoreça a interoperabilidade com outros repositórios.

Para acompanhar a evolução tecnológica verificada no domínio das Humanidades Digitais, é necessário encontrar um novo modelo de dados que seja eficaz na recuperação da informação e a conectar com a rede internacional de dados ligados. Consequentemente, é preciso também um novo conjunto de ferramentas de descrição direcionadas para bibliotecas e arquivos. Estas considerações levaram a Torre do Tombo a escolher o CIDOC CRM (norma direcionada para a comunidade dos museus) e a promover o projeto EPISA – *Entity and Property Inference for Semantic Archives*. Este projeto tem como principal objetivo o desenho de um protótipo para uma plataforma de grafo de conhecimento em *open source*, bem como o desenvolvimento de algoritmos para garantir que a migração das descrições armazenadas segue as orientações do Conselho Internacional de Arquivos.

O presente trabalho tem como objetivos testar a qualidade do modelo de metadados, através do estudo do modelo ArchOnto e do CIDOC CRM para familiarização com o projeto, seguindo-se a comparação das descrições arquivísticas, realizadas no quadro das normas ISAD, com as realizadas manualmente no novo modelo. O objetivo seguinte é a caracterização do processo de migração do modelo antigo para o novo, a partir da recolha dos metadados de um conjunto de registos de batismos, casamentos e óbitos dos Arquivos Distritais da Guarda e do Porto, bem como registos de passaportes deferidos, do Arquivo Distrital de Bragança. Por fim, pretende-se testar a qualidade do acesso, através da análise do comportamento de pesquisa de dois grupos de utilizadores (utilizadores comuns e arquivistas). Para isso, foi elaborado um guião com diferentes tarefas a realizar tanto no sistema antigo DigitArq como na nova interface. A execução das tarefas foi observada e os comportamentos comparados.

Do presente trabalho resultaram assim *datasets* anotados com os metadados dos registos paroquiais, bem como o *dataset* com os registos de passaportes. Resultou, ainda, o guião para a experiência de avaliação da nova interface, bem como um documento com o registo das observações das atividades propostas no guião e respetiva análise qualitativa.

Palavras-chave: CIDOC CRM, descrição arquivística, avaliação de ferramentas de descrição

Abstract

The Torre do Tombo National Archive holds original documents from the 9th century to the present day and, by using metadata, it is possible to describe the cultural objects it holds. It also uses the ISAD (G) and ISAAR (CPF) standards for archival description. The Archive has had its online description system for twenty years and is recently considering the evolution towards a system that favours interoperability with other repositories.

To keep up with the technological evolution in the Digital Humanities, it is necessary to find a new data model that is effective in retrieving information and connecting it to the international network of linked data. Consequently, a new set of description tools aimed at libraries and archives is also needed. These considerations led Torre do Tombo to choose CIDOC CRM (standard directed to the museum community) and to promote the EPISA project - Entity and Property Inference for Semantic Archives. This project has as main goal the design of a prototype for an open source knowledge graph platform, as well as the development of algorithms to ensure that the migration of all stored content follows the guidelines of the International Council on Archives.

The present work aims to test the quality of the metadata model, through the study of the ArchOnto model and the CIDOC CRM to familiarize with the project, followed by the comparison of the archival descriptions, made in the framework of the ISAD standards, with those made manually in the new model. The next objective is the characterization of the migration process from the old model to the new one, based on the metadata collection of a set of baptism, marriage and death records from the District Archives of Guarda and Oporto, as well as deferred passport records from the District Archive of Bragança. Finally, we intend to test the quality of access, through the analysis of the search behaviour of two groups of users (ordinary users and archivists). For this, a script was developed with different tasks to be performed both in the old DigitArq system and in the new interface. The execution of the tasks was observed and their behaviour was compared.

This work resulted in annotated datasets with the metadata of parish records, as well as the dataset with the passport records. Another result was the script for the experiment to evaluate the new interface, as well as a document recording the observations of the activities proposed in the script and the respective qualitative analysis.

Keywords: CIDOC CRM, archival description, assessment of description tools

Sumário

1. Introdução 13
   1.1. Enquadramento da dissertação 13
   1.2. Abordagem metodológica 16
   1.3. Estrutura da dissertação 17
2. Web Semântica e Dados Ligados 19
   2.1. Resource Description Framework (RDF) 20
   2.2. SPARQL Protocol and RDF Query Language 21
   2.3. Ontology Web Language (OWL) 21
   2.4. Dados Ligados 22
3. Normas para a descrição de arquivos 25
   3.1. Norma ISAD (G) 25
   3.2. Norma ISAAR (CPF) 27
   3.3. Norma ISDIAH 29
   3.4. Norma ISDF 30
4. Modelos de dados ligados para o Património Cultural 32
   4.1. CIDOC CRM 32
   4.2. Modelos baseados no CIDOC CRM 33
      4.2.1. CRMarcheo 33
      4.2.2. CRMsci 33
      4.2.3. CRMdig 33
      4.2.4. CRMba 33
      4.2.5. CRMgeo 34
      4.2.6. CRMtex 34
      4.2.7. CRMinf 34
   4.3. Projetos de interesse 34
      4.3.1. Projeto ARIADNEplus 34
      4.3.2. Polish Digital Libraries Federation 35
      4.3.3. DIESEL Project - Archive Linked Open Data (aLOD) 36
      4.3.4. ArCo 37

4.3.5. SARI Documentation 37
4.3.6. LinkedArt 38
5. A avaliação do modelo de dados 39
5.1. Modelo de dados 39
5.2. Caracterização do processo de migração 41
5.2.1. Arquivo Distrital da Guarda 41
5.2.2. Arquivo Distrital do Porto 52
5.2.3. Arquivo Distrital de Bragança 54
5.3. Análise de resultados da migração 58
6. Avaliação da qualidade de acesso 61
6.1. Perguntas colocadas no guião 64
6.1.1. Pergunta 1 64
6.1.2. Pergunta 2 65
6.1.3. Pergunta 3 66
6.1.4. Pergunta 4 66
6.1.5. Pergunta 5 67
6.1.6. Pergunta 6 68
6.2. Análise dos resultados 69
7. Conclusões 77
7.1. Trabalho futuro 79
8. Referências bibliográficas 81
Apêndices 85
Apêndice 1 – Guião completo 86
Apêndice 2 – Campos utilizados no registo das respostas 88
Apêndice 3 – Organização do registo das respostas 89
Apêndice 4 – Resultados das pesquisas efetuadas 90

Lista de figuras

Figura 1 - Exemplo de uma descrição de um batismo (Arquivo Distrital da Guarda). 47
Figura 2 - Exemplo de uma descrição de um livro de registos de batismo (Arquivo Distrital da Guarda) 48
Figura 3 - Exemplo de uma descrição de um livro de registos de casamento (Arquivo Distrital da Guarda) 50
Figura 4 - Exemplo de uma descrição de um livro de registos de óbito (Arquivo Distrital da Guarda) 51
Figura 5 - Exemplo de uma descrição de um registo de passaporte deferido (Arquivo Distrital de Bragança) 57
Figura 6 - Interface da Pesquisa Simples 61
Figura 7 - Interface da Pesquisa Avançada 62
Figura 8 - Interface DigitArq+ 63
Figura 9 - Idades dos utilizadores comuns 69
Figura 10 - Idades dos arquivistas 72
Figura 11 - Resultado da pesquisa avançada (Pergunta 1) 90
Figura 12 - Resultado da pesquisa simples (Pergunta 1) 90
Figura 13 - Resultado da pesquisa no DigitArq+ (Pergunta 1) 91
Figura 14 - Resultado da pesquisa simples ao utilizar “Jerónima da Cruz” (Pergunta 2) 91
Figura 15 - Resultado da pesquisa simples ao utilizar “processo de Jerónima da Cruz” (Pergunta 2) 92
Figura 16 - Resultado da pesquisa avançada ao utilizar “Jerónima da Cruz” (Pergunta 2) 92
Figura 17 - Resultado da pesquisa avançada ao utilizar “processo de Jerónima da Cruz” (Pergunta 2) 93
Figura 18 - Resultado da pesquisa no DigitArq+ ao utilizar “Jerónima da Cruz” (Pergunta 2) 93
Figura 19 - Entidade mostrada pelo DigitArq+ (Pergunta 2) 94
Figura 20 - Resultado da pesquisa no DigitArq+ ao utilizar “processo de Jerónima da Cruz” (Pergunta 2) 94
Figura 21 - Resultado da pesquisa simples ao utilizar “Mosteiro de São Bernardo de Portalegre” (Pergunta 3) 95
Figura 22 - Resultado da pesquisa avançada ao utilizar “Mosteiro de São Bernardo de Portalegre” (Pergunta 3) 95
Figura 23 - Resultado da pesquisa no DigitArq+ ao utilizar “Mosteiro de São Bernardo de Portalegre” (Pergunta 3) 96
Figura 24 - Entidade mostrada pelo DigitArq+ (Pergunta 3) 96
Figura 25 - Resultado da pesquisa simples ao utilizar “Juízo da Índia e Mina” (Pergunta 4) 97

Figura 26 - Resultado da pesquisa avançada ao utilizar “Juízo da Índia e Mina” (Pergunta 4) 97
Figura 27 - Resultado da pesquisa no DigitArq+ ao utilizar “Juízo da Índia e Mina” (Pergunta 4) 98
Figura 28 - Resultado da pesquisa simples ao utilizar “registos de baptismos” (Pergunta 5) 98
Figura 29 - Resultado da pesquisa simples ao utilizar “registos de baptismos”, com o intervalo temporal 1811-1822 (Pergunta 5) 99
Figura 30 - Resultado da pesquisa simples ao utilizar “registos de baptismos”, com a data completa (Pergunta 5) 99
Figura 31 - Resultado da pesquisa avançada ao utilizar “registos de baptismos” (Pergunta 5) 100
Figura 32 - Resultado da pesquisa avançada ao utilizar “registos de baptismos”, com o com o intervalo temporal 1811-1822 (Pergunta 5) 100
Figura 33 - Resultado da pesquisa avançada ao utilizar “registos de baptismos”, com a data completa (Pergunta 5) 101
Figura 34 - Resultado da pesquisa no DigitArq+ ao utilizar “registos de baptismos” (Pergunta 6) 101
Figura 35 - Resultado da pesquisa simples ao utilizar “Um aquário de 2,40x1,00 m” (Pergunta 6) 102
Figura 36 - Resultado da pesquisa avançada ao utilizar “Um aquário de 2,40x1,00 m” (Pergunta 6) 102
Figura 37 - Resultado da pesquisa avançada ao utilizar o código de referência “PT/TT/AS/A/001/000142” (Pergunta 6) 103
Figura 38 - Resultado da pesquisa no DigitArq+ ao utilizar o código de referência “PT/TT/AS/A/001/000142” (Pergunta 6) 103

Lista de tabelas

Tabela 1 - Organização dos registos de batismo do Arquivo Distrital da Guarda 43
Tabela 2 - Tabela com as etiquetas gerais (registos de batismo) 44
Tabela 3 - Tabela utilizada nos documentos compostos não detalhados (registos de batismo) 44
Tabela 4 - Tabela de etiquetas específicas (registos de batismo) 45
Tabela 5 - Exemplo de cabeçalho de identificação de um livro de registos de batismos 46
Tabela 6 - Exemplo de um registo de batismo na amostra 46
Tabela 7 - Exemplo de um registo na amostra de um livro não detalhado 47
Tabela 8 - Explicação das etiquetas utilizadas (registos de casamento) 48
Tabela 9 - Cabeçalho de identificação da série relativa aos casamentos 49
Tabela 10 - Exemplo de um registo na amostra de um livro de registos de casamento 49
Tabela 11 - Significados das etiquetas (registos de óbitos) 50
Tabela 12 - Cabeçalho da amostra da série relativa aos óbitos 51
Tabela 13 - Exemplo de um registo na amostra de um livro de registos de óbito 51
Tabela 14 - Distribuição dos registos de batismo do Arquivo Distrital do Porto 53
Tabela 15 - Significados das etiquetas na série de registos de batismos do Arquivo Distrital do Porto 53
Tabela 16 - Distribuição dos registos de passaportes deferidos do Arquivo Distrital de Bragança 55
Tabela 17 - Explicação das etiquetas utilizadas 56
Tabela 18 - Exemplo de um cabeçalho de identificação de um livro de registos de passaportes 56
Tabela 19 - Exemplo de registo na amostra relativo a um passaporte deferido 57
Tabela 20 - Exemplo de outras etiquetas utilizadas 58
Tabela 21 - Campos tratados na Unidade de Instalação 0004 59
Tabela 22 - Campos tratados na Unidade de Instalação 0018 59
Tabela 23 - Informação sobre o dataset obtido após a migração 60
Tabela 24 - Resultados com todas as expressões 60
Tabela 25 - Resultados com as expressões não nulas 60
Tabela 26 - Explicação de campos utilizados no registo das respostas ao guião 88
Tabela 27 - Organização da folha de Excel no registo das respostas ao guião 89

1. Introdução

O Arquivo Nacional da Torre do Tombo, gerido pela Direção-Geral do Livro, dos Arquivos e das Bibliotecas (DGLAB) é uma das instituições mais antigas do país que, no decorrer dos anos, tem vindo a recolher uma vasta coleção de documentos históricos. A sua coleção possui documentos originais desde o século IX até à atualidade e utiliza metadados para descrever os objetos culturais, enquadrados pelas normas ISAD (G) e ISAAR (CPF) relativas à descrição arquivística.

Os metadados utilizados na descrição arquivística ajudam os utilizadores a encontrar e a contextualizar os documentos que procuram. Desta forma, o Arquivo Nacional da Torre do Tombo definiu o próprio sistema de descrição *online*, seguindo as normas do Conselho Internacional de Arquivos [1]. Os metadados utilizados dizem respeito, na sua maioria, a descrições textuais do contexto e ao conteúdo de documentos. Porém, estes não exploram todas as potencialidades da pesquisa de documentos, dado que não estão registadas todas as relações relevantes do contexto e conteúdo documentos existentes, em forma processável automaticamente. Esta questão é fulcral, dado que os metadados utilizados são a forma principal de chegar até aos documentos e é através daqueles que é possível encontrar diversos tipos de documentos como, por exemplo, documentos digitais com descrição e documentos físicos com e sem descrição. Desta forma, revela-se ser preciso gerar novas ferramentas para efetuar a descrição dos documentos, devendo ser mais flexíveis e também processáveis automaticamente.

Perante esta situação, o Arquivo Nacional da Torre do Tombo pretende renovar a infraestrutura de dados que existe e que é utilizada na descrição arquivística.

1.1. Enquadramento da dissertação

O EPISA – *Entity and Property Inference for Semantic Archives* – é um projeto que possui como parceiros o INESC TEC (Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência), a Universidade de Évora e a Direção Geral do Livro, dos Arquivos e das Bibliotecas que, por sua vez, é detentora da coleção de património cultural mais relevante, mais amplamente digitalizado e acedido.

Este projeto de investigação tem como principal objetivo o desenvolvimento de um protótipo para uma plataforma de grafo de conhecimento em *open source*, adotando o modelo de dados

[1] www.ica.org

para descrição arquivística que a DGLAB tem em estudo, para futura utilização. Dedica-se, de igual forma, ao desenvolvimento de algoritmos para garantir que a migração de todo o conteúdo armazenado fica de acordo com os padrões do Conselho Internacional de Arquivos para um modelo baseado em ontologias.

O Arquivo Nacional da Torre do Tombo concebeu e utiliza um sistema próprio para fazer a descrição de documentos, tendo por base as normas do Conselho Internacional de Arquivos. Desta forma, os metadados são compostos, sobretudo, pelas descrições textuais do contexto e conteúdo dos respetivos documentos. Porém, os documentos presentes no arquivo progrediram com o intuito de abranger quantidades crescentes de informação digital e, consequentemente, os requisitos de interoperabilidade cresceram. Perante esta ocorrência, revelou-se necessário gerar novas ferramentas para fazer a descrição dos documentos que fossem mais flexíveis e mais passíveis de ação mecânica. Estas características levaram a Torre do Tombo a escolher o CIDOC CRM – uma norma desenvolvida na comunidade museológica que pode ser traduzido sob a forma de um grafo, onde os "nós" representam as entidades e as "arestas" representam as relações existentes.

Contudo, este passo ainda requer um estudo aprofundado das técnicas de representação de informação, de migração de dados do sistema anterior e das interfaces de pesquisa mais adequadas. É justamente essas questões que o EPISA se dedica a resolver. A **primeira questão** é relativa à migração entre as normas do Conselho Internacional de Arquivos e a norma CIDOC CRM, dado que esta utiliza os modelos de tradução já existentes e, também, utiliza a inferência das novas relações com métodos semiautomatizados. A **segunda questão** diz respeito ao apoio que é fornecido à descrição, através de ferramentas que automatizam uma parte da geração dos registos de metadados mais complexos do CIDOC CRM. Por fim, a **terceira questão** é sobre a melhoria do acesso dos utilizadores aos arquivos, o que leva à promoção da interoperabilidade dos arquivos com a rede semântica. Esta questão só é resolvida com a melhoria da interface para os utilizadores como para máquinas (INESC TEC 2021).

Perante o exposto, é expectável que do EPISA surja uma alteração nos metadados do Património Cultural, no sentido de fornecer aos utilizadores um conhecimento do repositório, melhorando as ferramentas de pesquisa; uma maior presença nos agregadores, sobretudo na Europeana que utiliza uma abordagem semelhante; uma forma de lidar com metadados em diferentes plataformas, contribuindo, desta forma, "para a integração de recursos administrativos e de investigação no Arquivo Digital da Administração Pública" (INESC TEC and DGLAB 2019).

O EPISA é composto por oito tarefas. A **tarefa 1 – *Project management and dissemination of results*** – inclui todos os recursos necessários para a gestão do projeto e tem como atividades a organização de reuniões, a conceção de planos de trabalho, bem como a criação e publicação do website do projeto e a divulgação dos resultados. A **tarefa 2 – *Linked data model for cultural content*** – diz respeito à produção de um modelo de dados para o conteúdo cultural existente para que seja possível "interoperar com conjuntos de dados curados por múltiplas comunidades de prática" (INESC TEC and DGLAB 2019). Ou seja, é feito um mapeamento de todos os esquemas de metadados utilizados na informação cultural para os seus homólogos do CIDOC CRM. Desta forma, a tarefa dá origem a um modelo de dados que tem como base o CIDOC CRM para a descrição de conceitos e das suas relações em objetos do património cultural. Já a **tarefa 3 – *Semantic migration with relations inference*** – tem como objetivo a migração de descrições das normas utilizadas, nomeadamente ISAD (G), ISAAR (CPF), ISDIAH e ISDF para o modelo de dados CIDOC CRM, revelando-se fulcral trabalhar com as restrições adjacentes do domínio, bem como as suas características linguísticas e a definição de uma métrica para efetuar a avaliação da qualidade da informação extraída. A **tarefa 4 – *Knowledge propagation*** – tem como principais objetivos a propagação de informação sobre o grafo de inferência de relações relevantes para o grafo de conhecimento do EPISA, bem como a contribuição de dados dos arquivos para a Wikidata[2]. Assim, pretende-se que exista uma integração de arquivos na rede semântica global. Seguidamente, a **tarefa 5 – *Document mining for automatic metadata records*** – foca-se no desenvolvimento de ferramentas que visam facilitar o trabalho de cada profissional, enquanto efetuam a descrição dos objetos, através do uso do novo modelo de dados. Deste modo, aguardam-se resultados como, por exemplo, a avaliação da adequação dos algoritmos existentes para o modelo de dados interligados adotado e o desenvolvimento de uma interface gráfica para permitir aos profissionais uma revisão dos registos automáticos de metadados. A **tarefa 6 – *Knowledge graph management platform*** – tem o intuito de fornecer uma base para executar o algoritmo de extração de metadados, ao mesmo tempo que permite aos arquivistas gerir, de forma manual, o conteúdo do grafo. Assim, é adotada uma engenharia de software que inclui testes para uma integração contínua. A **tarefa 7 – *Exploration and querying interface*** – é voltada para o desenvolvimento de ferramentas tanto de exploração como de consulta, com o intuito de ajudar os arquivistas aquando da migração de registos ISAD para CIDOC CRM ou aquando da produção de novos registos de metadados. Pretende, também, dar ferramentas de visualização para que seja possível reutilizar o conhecimento no novo

[2] www.wikidata.org/wiki/Wikidata:Main_Page

arquivo de documentos representados por grafos.

A presente dissertação incide sobre a **tarefa 8 – *User evaluation*** – que reúne todas as atividades de avaliação a partir dos diferentes pontos de vista do utilizador. Esta tarefa é composta por cinco experiências, onde cada uma possui um objetivo concreto. A primeira tem o objetivo de testar a qualidade do modelo de metadados, escolhendo alguns registos e executando manualmente as descrições de acordo com o novo modelo. A segunda experiência envolve processar algoritmos de migração e descrever o conjunto de dados resultante, aplicando o algoritmo de migração aos registos de descrição completos no repositório da Torre do Tombo, bem como caracterizar o processo de migração. A terceira avalia a qualidade dos resultados das migrações, através da extração de uma amostra do repositório da Torre do Tombo, comparando sistematicamente as descrições originais com a representação migrada. A quarta experiência testa a qualidade da geração de metadados nos modos automático e semiautomático, através de uma amostra de documentos com transcrições disponíveis e da aplicação do algoritmo de geração de metadados aos conteúdos de cada documento. Finalmente, a quinta experiência testa a qualidade do acesso, partindo de uma análise dos registos do sistema da Torre do Tombo para determinar as tarefas de investigação mais representativas para diferentes tipos de utilizadores.

A presente dissertação visa contribuir para as primeira, terceira e quinta experiências, numa fase ainda não final do projeto EPISA, pelo que as conclusões obtidas não serão ainda as do fecho do projeto.

1.2. Abordagem metodológica

No momento de avaliação da qualidade do modelo de metadados foi utilizada a análise de documentos e de ficheiros, com o intuito de estudar os modelos e de comparar as descrições arquivísticas.

Já na caracterização do processo de migração foi utilizada a análise de metadados para que fosse possível saber que metadados seriam importantes para efetuar a recolha.

Por fim, na avaliação da qualidade de acesso foi utilizada a observação e o inquérito para analisar o comportamento de pesquisa dos diferentes grupos de utilizadores.

1.3. Estrutura da dissertação

A presente dissertação encontra-se dividida em sete capítulos, conforme será apresentado seguidamente.

O **primeiro capítulo** – *Introdução* – tem o objetivo de esclarecer o contexto da presente dissertação, explicando o seu enquadramento, bem como a abordagem metodológica utilizada.

Os **segundo, terceiro e quarto capítulos**, denominados *Web Semântica e Dados Ligados*, *Normas para a descrição de arquivos* e *Modelos de dados ligados para o Património Cultural*, respetivamente, são relativos à revisão da literatura. O segundo explicita o que é a Web Semântica, seguindo-se uma abordagem a tecnologias como *RDF*, *OWL* e *SPARQL*. O terceiro versa sobre as normas ISAD (G), ISAAR (CPF), ISDIAH e ISDF, mostrando no que é que cada uma consiste e como é que estão estruturadas. O quarto aborda o CIDOC CRM, explicando no que consiste e apresentando alguns estudos de aplicação baseados neste modelo. Seguidamente, são apresentados alguns projetos que se enquadram no domínio dos dados ligados no Património Cultural, sendo que alguns possuem o CIDOC CRM como base das suas ontologias. A pesquisa para estes capítulos ocorreu em bases de dados como, por exemplo, a SCOPUS[3], bem como no *website* da W3C[4], do Conselho Internacional de Arquivos e do modelo CIDOC CRM[5].

Por sua vez, o **quinto capítulo** – *A avaliação do modelo de dados* – é sobre a organização da ontologia *ArchOnto*, seguindo-se uma comparação entre as descrições arquivísticas efetuadas pela equipa do INESC TEC e as descrições arquivísticas efetuadas pela equipa da DGLAB relativamente a vinte e cinco registos previamente selecionados. Em paralelo com esta comparação, efetua-se uma avaliação da migração de dados, mostrando como se efetuou a caracterização do processo de migração, esclarecendo que amostra de registos foi analisada, a que Arquivos Distritais pertenciam e de que forma a recolha dos metadados dos registos foi organizada.

O **sexto capítulo** diz respeito à qualidade do acesso, comparando a pesquisa do DigitArq com a da nova interface (DigitArq+), através de um guião com algumas questões, tendo sempre em conta o grau de satisfação dos utilizadores.

A seguir, no **sétimo capítulo** são apresentadas as *Conclusões* do trabalho desenvolvido, sendo, também, apresentadas algumas sugestões de tarefas a serem realizadas futuramente.

[3] www.scopus.com
[4] www.w3.org
[5] www.cidoc-crm.org

Finalmente, após estes capítulos, encontram-se as referências bibliográficas utilizadas na elaboração da presente dissertação e os apêndices.

2. Web Semântica e Dados Ligados

A Web Semântica foi proposta há vinte anos (Berners-Lee, Hendler, and Lassila 2001) e originou o conjunto de bases de conhecimento que agora compõem a nuvem de Dados Ligados. As tecnologias da Web Semântica criaram as condições necessárias para motivar iniciativas como a *DBpedia*, referida mais abaixo. Por outro lado, estas tecnologias promoveram a interoperabilidade e, assim, permitiram que as coleções especializadas de informação se constituíssem como uma parte da rede global de conhecimento.

A Web Semântica é, também, definida como uma "coleção de conhecimentos em evolução, construída para permitir a qualquer pessoa na Internet acrescentar o que sabe e encontrar respostas às suas perguntas" (W3C 2002). A informação está presente de uma forma estruturada – em vez de estar em linguagem natural –, o que facilita o trabalho de cada pessoa e, também, dos computadores. Desta forma, as tecnologias da Web Semântica ocasionam a criação de armazéns de dados na *Web*, bem como estimulam a construção de vocabulários e a definição de regras para o tratamento dos dados.

Relativamente à sua estruturação, esta é feita através do registo de conhecimentos elementares interligados. Ou seja, o conhecimento é expresso através de declarações descritivas que mostram a existência de relações entre diversos recursos de informação. Esta estruturação foi padronizada como *Resource Description Framework*.

No que concerne aos nomes utilizados, estes são mais complicados, dado que podem não ser a melhor forma de identificar algo mais específico. A identificação é feita tanto a coisas como a relações e os nomes utilizados são endereços *web*, também denominados URL, tendo como objetivo ser útil na localização de informação.

É importante referir que na Web Semântica alguns factos (explícitos) levam a outros factos (inferidos), através do uso da lógica. Apesar de isto ser bastante intuitivo para uma pessoa, pode não ser assim tão simples para um computador. Porém, quando programados, estes podem ser muito úteis para reconhecer que determinados factos derivam, logicamente, de outros factos. Uma forma de ajudar os computadores a manipularem estes procedimentos lógicos é a criação de ontologias que podem ser comunicadas em diversos idiomas. É, neste preciso ponto, que o termo *"semântica"* se posiciona: na tentativa de tornar claros os significados para que um computador os possa processar. Apesar de nenhuma linguagem lógica ter sido previamente recomendada para a Web Semântica, os dados ligados são expressos por tecnologias tais como *RDF*, *OWL* e *SPARQL* que irão ser abordadas abaixo.

Dado que a Web Semântica é uma parte da *World Wide Web*, que faz com que os seres

humanos e os computadores trabalhem em cooperação, torna-se necessário falar um pouco sobre esta. Perante isto, é importante referir que a *World Wide Web* "tem-se desenvolvido como um meio de disponibilização de documentos para as pessoas" (Berners-Lee, Hendler, and Lassila 2001). Ou seja, através do aumento das páginas Web com dados direcionados aos computadores, é possível transformar a *Web* na Web Semântica. A principal característica da *World Wide Web* é ser universal. Desta forma, a informação varia ao longo de diversos eixos e um deles é a diferença entre informação produzida para seres humanos e informação produzida para computadores.

A *Web* desenvolveu-se de forma mais rápida como um meio de disponibilização de documentos para seres humanos do que de dados e informação que possam ser processados automaticamente. A Web Semântica tem como objetivo compensar esta situação (W3C 2012) e, desta forma, os computadores irão recuperar a semântica dos dados em questão, seguindo hiperligações para definir termos e regras fulcrais para raciocinar logicamente sobre estes e a infraestrutura resultante incentivará o desenvolvimento de serviços automatizados. Todavia, importa atentar que a maioria do conteúdo presente na *Web* é concebido para a leitura dos seres humanos e não para a manipulação dos programas de computador.

A Web Semântica dá uma organização ao conteúdo das páginas e dá origem a um ambiente onde os programas de computador referenciados nas páginas podem realizar tarefas para os utilizadores. Contudo, a "Web Semântica não é uma Web separada" (W3C 2012): é uma extensão da que existe atualmente, onde a informação possui significado e permite que pessoas e computadores trabalhem em cooperação. Então, para a Web Semântica funcionar, os computadores devem ter acesso à informação, bem como aos conjuntos de regras de inferência que devem levar a raciocínios automatizados.

Assim, é possível concluir que o desafio da Web Semântica é fornecer uma linguagem que seja capaz de exprimir dados e regras de raciocínio e que permita a exportação para a *Web* de qualquer sistema de representação de conhecimento existente.

Atualmente, acrescentar lógica à *Web* é a tarefa a que a comunidade da Web Semântica se dedica. Esta lógica, por sua vez, deve ser poderosa o suficiente para que possa descrever as propriedades mais complexas dos objetos.

2.1. Resource Description Framework (RDF)

O *Resource Description Framework* pode ser entendido como uma "linguagem para representar a informação sobre recursos na World Wide Web" (Manola and Miller 2004). Foi

criado com o intuito de representar metadados relativos a recursos da *Web* como, por exemplo, título, data de modificação, autor ou direitos de autor. O RDF pode, de igual forma, ser utilizado para representar a informação relativa a coisas que podem ser identificadas na *Web* – informação sobre artigos disponíveis ou a descrição das preferências de um utilizador.

Também tem o objetivo de facilitar em situações onde a informação é processada por aplicações específicas, através de um quadro comum que mostra a informação de modo que possa ser trocada entre aplicações sem perder o significado.

O RDF tem, então, como base a identificação de recursos através de identificadores *Web*, bem como da sua descrição em termos de propriedades simples e respetivos valores.

2.2. SPARQL Protocol and RDF Query Language

A linguagem SPARQL é utilizada para "expressar interrogações através de diversas fontes de dados, quer os dados sejam armazenados nativamente como RDF ou vistos como RDF" (W3C 2013).

A SPARQL pode ser utilizada para expressar interrogações através de fontes de dados, sendo um conjunto de especificações que baseiam protocolos para manipulação de conteúdo RDF. A formulação de interrogações vai desde a mais simples à mais complexa e pode incluir agregação de valores.

2.3. Ontology Web Language (OWL)

A OWL é "uma linguagem da Web Semântica concebida para representar um conhecimento rico e complexo sobre as coisas, grupos de coisas, e relações entre as coisas" (OWL Working Group 2012). Tem como base a lógica computacional e, desta forma, o conhecimento pode ser percorrido por programas de computador para tornar o conhecimento implícito em conhecimento explícito. Os documentos produzidos em OWL tanto podem ser publicados na *World Wide Web* como podem ser referidos a partir de outras ontologias.

Uma ontologia traduz-se num "conjunto de afirmações descritivas precisas sobre alguma parte do mundo" (Hitzler et al. 2012) que, normalmente, é mencionado como *"domínio"* ou *"assunto"* da ontologia. Para que o *"domínio"* seja descrito de forma precisa, é preciso que seja elaborado o vocabulário e que seja esclarecido o seu significado. Para além desta definição em linguagem natural, cada termo do vocabulário pode estar relacionado com os outros termos,

ajudando na sua caracterização.

O conhecimento é representado através de *Axiomas* (afirmações básicas que uma ontologia possui), *Entidades* (elementos utilizados para referir objetos) e *Expressões* (combinações de entidades para formar descrições).

Na linguagem OWL mostram-se objetos como *Indivíduos*, categorias como *Classes* e relações como *Propriedades*. Relativamente aos ***Indivíduos***, é de referir que são representados por objetos no domínio. Esta linguagem não pressupõe que nomes diferentes se referem a indivíduos diferentes. Ou seja, é importante clarificar se os indivíduos pertencentes ao domínio são os mesmos ou diferentes uns dos outros. No que concerne às ***Classes*** é importante mencionar que estas são "conjuntos que contêm os indivíduos" (Hitzler et al. 2012) e são descritas de modo formal, sendo organizadas em hierarquias como superclasse/subclasse onde as subclasses são especializações da respetiva superclasse. As ***Propriedades*** dizem respeito a relações binárias entre indivíduos e existem dois tipos principais de propriedades:

1. Propriedade de Objeto (*Object Property*): tem o intuito de unir um indivíduo a outro indivíduo;
2. Propriedade de Dados (*Data Property*): atribui valores de dados a objetos.

Existe, também, a Propriedade de Anotação (*Annotation Property*) que é utilizada para adicionar metadados a *Classes*, *Indivíduos* e outras propriedades.

2.4. Dados Ligados

As tecnologias da Web Semântica fornecem um ambiente onde é possível efetuar interrogações com dados, assim como extrair inferências usando vocabulários. Porém, para tornar a *"Web de Dados"* uma realidade, é importante ter todos os dados na *Web* disponíveis num "formato padrão, alcançável e gerível por ferramentas da Web Semântica" (W3C 2011). Além disso, não é só a Web Semântica que precisa de acesso a dados: também as relações entre dados devem ser disponibilizadas para criar uma Web de Dados. A recolha dos dados interrelacionados também pode ser designada como *Dados Ligados*. Para que seja possível criar estes *Dados Ligados*, é necessário que existam tecnologias que utilizem um formato comum para aceder a bases de dados em tempo real.

Desta forma, os *Dados Ligados* estão no centro daquilo que é a Web Semântica: integração em grande escala de dados na *Web* e um exemplo concreto de um grande Conjunto de Dados Ligados é a DBpedia, cujo intuito é tornar o "conteúdo da Wikipedia disponível em RDF" (W3C

2011).

A DBpedia é um projeto iniciado em 2007, pela Universidade Livre de Berlim e pela Universidade de Leipzig, na Alemanha e conta com a colaboração do *OpenLink Software*[6]. A DBpedia é criada a partir de informação estruturada, de geo-coordenadas e de *links* externos que são extraídos da *Wikipedia*, de forma automática.

A informação estruturada é semelhante a um grafo de conhecimento que está disponível na *Web*. Um grafo de conhecimento é um tipo de base de dados que "armazena o conhecimento numa forma legível e fornece um meio para a recolha, organização, partilha, pesquisa e utilização da informação" (DBpedia Association 2020). Esta informação está disponível nos termos da Licença *Creative Commons Attribution-ShareAlike 3.0* e a GNU *Free Documentation License*.

As suas bases de conhecimento têm uma importância cada vez mais fulcral na melhoria dos resultados da pesquisa na *Web*. Possuem diversas vantagens, relativamente às bases de conhecimento existentes, como a cobertura de muitos domínios; a representação de um acordo comunitário; a evolução constante à medida que a *Wikipedia* muda; e a diversidade de idiomas existente (DBpedia Association 2020).

Existem, ainda, entidades classificadoras que são englobadas dentro de quatro esquemas de classificação, a fim de cumprir diferentes requisitos de aplicação:

- **Categorias da Wikipedia**: A DBpedia contém uma representação SKOS do sistema de categorias da *Wikipedia*. A principal vantagem deste sistema é o facto de ser atualizado por diversos editores da Wikipédia. Porém, possui uma desvantagem que se traduz na inexistência de uma hierarquia adequada nas categorias.
- **YAGO**: este esquema de classificação foi criado com o intuito de mapear as categorias de páginas da *Wikipedia* e caracteriza-se pela profundidade e pela codificação da informação numa só classe. Embora a YAGO possua, de forma geral, um alcance com alta precisão, existem alguns erros e/ou omissões devido à geração automática.
- **UMBEL** (*Upper Mapping and Binding Exchange Layer*): é uma ontologia que foi criada para interligar conteúdos e dados da *Web* e deriva da *OpenCyc*[7]. Tem sido utilizada para ajudar na procura de conceitos e na sua definição, na classificação de interrogações e na integração e verificação de consistência de ontologias. Possui os seguintes meios para promover a interoperabilidade semântica da informação: uma

[6] Fornecedor de tecnologia segura e de alto desempenho.
[7] Projeto de inteligência artificial.

ontologia com diversos conceitos de referência, desenvolvida para fornecer pontos de mapeamento comuns e relacionar diferentes ontologias; e um vocabulário que auxilia o mapeamento da ontologia e que inclui expressões de relações de probabilidade distintas da identidade (DBpedia 2020).

- **Ontologia DBpedia**: esta ontologia foi criada a partir de especificações feitas de forma manual no *Wiki DBpedia Mappings*[8]. Cada versão desta ontologia corresponde a uma nova versão do conjunto de dados da DBpedia que contém dados extraídos das diferentes versões linguísticas da *Wikipedia* (DBpedia Association 2002).

Com o objetivo de cumprir todos os requisitos de diferentes aplicações, a DBpedia possui mecanismos de acesso, nomeadamente:

- **Dados Ligados**: conforme discutido em cima, é um método de publicação de dados RDF que se baseia em URI HTTP como sendo identificadores de recursos e no protocolo HTTP para recuperação de descrições de recursos.
- **SPARQL endpoint**: suporta várias extensões da linguagem de uma interrogação. Estas extensões revelaram-se úteis no desenvolvimento de outras aplicações, como, por exemplo, pesquisa de texto completo e funções agregadas.
- **Índice de pesquisa**: fornece uma pesquisa de cada etiqueta e culmina na semelhança entre *strings* e uma classificação de relevância de modo a encontrar as correspondências mais prováveis para um determinado termo.

[8] https://mappings.dbpedia.org/index.php/Main_Page

3. Normas para a descrição de arquivos

Neste capítulo são abordadas as principais normas que a Torre do Tombo tem utilizado para a representação dos objetos culturais.

3.1. Norma ISAD (G)

A Norma ISAD (G)[9] foi publicada no decorrer do Congresso do Conselho Internacional de Arquivos, em 2000. Contudo, em Portugal o trabalho foi iniciado há vinte anos no "âmbito da normalização da terminologia arquivística" (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2002) que permitiu a "consolidação de um léxico próprio e tendencialmente uniforme" (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2002).

O seu objetivo é "identificar e explicar o contexto e o conteúdo da documentação de arquivo" (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2002), com o intuito de promover a acessibilidade. Perante isto, é possível afirmar que este objetivo só é atingido com a elaboração de representações precisas e adequadas. A descrição arquivística possui um conjunto de regras que pertence a um processo com o intuito de:

a) "assegurar a produção de descrições consistentes, apropriadas e autoexplicativas;
b) facilitar a recuperação e a troca de informação sobre documentos de arquivo;
c) possibilitar a partilha de dados de autoridade;
d) tornar possível a integração de descrições provenientes de diferentes entidades detentoras num sistema unificado de informação" (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2002).

As regras estão organizadas em sete zonas:

- **Zona da identificação**: identificação da unidade de descrição;
- **Zona do contexto**: natureza da unidade de descrição;
- **Zona do conteúdo e estrutura**: tema e estrutura da unidade de descrição;
- **Zona das condições de acesso e de utilização**: acesso e disponibilização da

[9] ISAD(G): Norma geral internacional de descrição arquivística.

unidade de descrição;

- **Zona da documentação associada**: informação sobre os documentos relacionados;
- **Zona das notas**: informação específica que não se inclui nas restantes zonas apresentadas;
- **Zona do controlo da descrição**: elaboração da descrição arquivística.

Neste ponto, importa referir que a própria norma prevê que nem todos os elementos terão de ser, forçosamente, utilizados numa descrição, pelo que "é possível utilizar somente parte deles para elaborar uma determinada descrição" (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2002), sendo que estes elementos são o título, produtor, código de referência, bem como a data (ou datas), o nível de descrição e a dimensão da unidade de descrição.

Na *Descrição Multinível*, o fundo deve ser representado numa só descrição, caso esteja a ser descrito como um todo ou, caso contrário, as diversas partes podem ser descritas em separado. Porém, existem regras que a *Descrição Multinível* deve respeitar, nomeadamente a descrição do geral para o particular para representar a organização hierárquica do fundo, bem como das partes que o compõem; a informação importante para cada nível de descrição deve representar tanto o contexto como o conteúdo da unidade de descrição; a ligação entre descrições, que clarifica qual a posição na hierarquia da unidade de descrição; a não repetição de informação que serve para evitar redundâncias nas descrições arquivísticas relacionadas ao nível da hierarquia.

Devem, ainda, ser respeitados todos os elementos no momento da descrição do fundo arquivístico, dependendo da zona de informação descritiva (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2002):

- **Zona da Identificação**: engloba os campos do código (ou códigos) de referência, bem como título, data ou datas, nível de descrição e suporte e dimensão;
- **Zona do Contexto**: reúne os campos do nome do produtor (ou produtores), da história arquivística e, por fim, a forma de aquisição;
- **Zona do Conteúdo e da Estrutura**: aglomera o âmbito e conteúdo, a avaliação, seleção e eliminação, o ingresso adicional e o sistema de organização;
- **Zona das Condições de Acesso e de Utilização**: incorpora as condições de acesso,

as condições de reprodução, o idioma, as características físicas, os requisitos técnicos e os instrumentos de descrição utilizados;

- **Zona da Documentação Associada**: possui a existência e localização dos itens originais e das cópias, as unidades de descrição e a nota de publicação.
- **Zona de Notas**: tem um único campo existente que são as notas;
- **Zona do Controlo da Descrição**: junta notas do arquivista, as regras ou convenções e as datas de cada descrição.

3.2. Norma ISAAR (CPF)

A Norma ISAAR (CPF)[10] foi desenvolvida pela Comissão Ad Hoc do Conselho Internacional de Arquivos entre os anos de 1993 e de 1995, tendo sido originalmente publicada no ano de 1996. Contudo, a edição abordada é a segunda, publicada no decorrer de 2004.

Esta norma tem como finalidade partilhar as descrições dos produtores de cada item, bem como "promover a preparação de descrições consistentes, apropriadas e autoexplicativas das pessoas coletivas, das pessoas singulares e das famílias que os produziram" (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2004). Isto dá lugar a que sejam fornecidas regras gerais para normalizar a descrição arquivística dos produtores de itens e do contexto de produção, permitindo (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2004):

- Aceder aos arquivos tendo como base a descrição do contexto de produção;
- Compreender o contexto da produção e da utilização de cada arquivo;
- Identificar de forma precisa dos produtores dos itens, incluindo descrições das relações entre cada entidade;
- Transferir as descrições entre instituições e/ou sistemas.

A norma dá, igualmente, "orientações para a preparação de registos de autoridade arquivística que proporcionam descrições das entidades (pessoas coletivas, pessoas singulares e famílias) associadas à produção e gestão de arquivos" (Grupo de Trabalho para a Normalização da

[10] ISAAR (CPF): Norma Internacional de Registo de Autoridade Arquivística para Pessoas Coletivas, Pessoas Singulares e Famílias.

Descrição em Arquivo 2004). Por sua vez, os registos de autoridade devem utilizar-se na descrição de pessoas coletivas, singulares ou famílias como uma unidade que está inserida num sistema de descrição arquivística, bem como na criação e no uso de pontos de acesso e na documentação de relações entre diferentes produtores.

Os elementos da descrição de autoridade organizam-se em quatro zonas de informação (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2004):

- **Zona da Identificação**: identificação da entidade, definindo um ponto de acesso;
- **Zona da Descrição**: informação relativa à natureza, ao contexto e às atividades da entidade;
- **Zona das Relações**: registo e descrição das relações com pessoas coletivas, singulares e/ou famílias;
- **Zona do Controlo**: identificação do registo de autoridade, bem como a forma, o momento e o serviço onde este foi criado e mantido.

Diversos elementos da descrição são utilizados como pontos de acesso e, desta forma, devem ser dinamizadas regras para a sua normalização. Esta norma é para ser utilizada em estreita colaboração com a Norma ISAD (G) e, também, com outras normas a nível nacional.

Além do previamente exposto, existem elementos a serem tidos em conta no momento da descrição do fundo arquivístico, concretamente:

- **Zona da Identificação**: engloba os campos do tipo de entidade, da forma autorizada do nome, bem como outras formas deste e identificadores para pessoas coletivas;
- **Zona da Descrição**: reúne os campos da data de existência, da história, do lugar, do estatuto legal, das funções, de ocupações e atividades, entre outros;
- **Zona das Relações**: incorpora o nome ou o identificador da pessoa coletiva, da pessoa singular ou da família, bem como o tipo de relação e a descrição da relação;
- **Zona do Controlo**: possui o identificador do registo de autoridade, os identificadores da instituição, as regras, assim como as línguas e escritas e as fontes, entre outros.

Cada registo de autoridade é criado com o intuito de registar o contexto da produção dos itens do arquivo. Porém, para que a informação se revele útil, é preciso conectar cada registo de autoridade com a descrição dos itens. E quando a conexão ocorre é importante fazer a descrição

da natureza da relação existente entre a pessoa coletiva, a pessoa singular ou a família e a fonte relacionada. Isto só é possível através de identificadores e títulos do recurso relacionado; do tipo do recurso relacionado; da natureza da relação; e das datas do recurso relacionado (Grupo de Trabalho para a Normalização da Descrição em Arquivo 2004).

3.3. Norma ISDIAH

A Norma ISDIAH[11] foi elaborada pelo Comitê de Boas Práticas e Normas em Londres, Reino Unido, em março de 2008. Esta norma surgiu com o intuito de "criar um sistema de informação arquivística mais útil" (Conselho Internacional de Arquivos 2008), tendo "uma descrição separada e normalizada dos custodiadores" (Conselho Internacional de Arquivos 2008). Desta forma, a norma mostra que tipo de informação pode ser incluída nas descrições efetuadas pelas instituições de arquivo, dando, de igual forma, uma orientação sobre como as descrições em causa podem ser desenvolvidas.

Relativamente aos seus elementos de descrição, é de referir que estes estão organizados em seis áreas que devem ser respeitados no momento de descrição:

- **Área de Identificação**: mostra a identificação da instituição com acervo arquivístico e a definição dos pontos de acesso normalizados e possui elementos como o identificador, a(s) forma(s) autorizada(s) do nome, a(s) forma(s) paralela(s) do nome, bem como outra(s) forma(s) do nome e o tipo da instituição com acervo arquivístico;
- **Área de Contacto**: identifica os contactos da instituição e tem elementos como o(s) endereço(s), telefone, correio eletrónico e responsáveis para contato;
- **Área de Descrição**: tem informação sobre a história, a estrutura e as políticas de entrada de documentos da instituição, com elementos como a história da instituição com acervo arquivístico, o contexto geográfico e cultural, os mandatos/fontes de autoridade, a estrutura administrativa, as políticas de gestão e de entrada de documentos, bem como o(s) prédio(s), o acervo arquivístico e outros acervos e os instrumentos de pesquisa, guias e publicações;
- **Área de Acesso**: providencia informação relativa ao acesso à instituição, contendo elementos como o horário de funcionamento, as condições e requisitos para acesso e uso e a acessibilidade;
- **Área de Serviços**: dispõe de informação sobre os serviços técnicos oferecidos pela instituição com elementos como os serviços de pesquisa, os serviços de reprodução e as

[11] ISDIAH: Norma Internacional para Descrição de Instituições com Acervo Arquivístico.

áreas públicas;

- **Área de Controlo**: possui uma descrição da instituição, tendo um vasto conjunto de elementos como o identificador da descrição, o identificador da instituição, as regras e/ou convenções utilizadas, o status, o nível de detalhamento, as datas de criação, revisão ou obsolescência, o(s) idioma(s) e sistema(s) de escrita, as fontes e as notas de manutenção.

É, de igual forma, importante referir que existem três elementos que são de uso obrigatório: o identificador, a(s) forma(s) autorizada(s) do nome e o(s) endereço(s). Na descrição destas instituições, a utilização e o formato de cada elemento de caráter opcional devem ser previamente definidos pela própria instituição, atendendo ao uso pretendido de cada descrição. Muitos destes elementos podem ser utilizados como pontos de acesso, tendo em atenção regras ou convenções para a sua normalização. É, também, essencial que a norma seja utilizada em sintonia, nomeadamente, com as normas ISAD (G), ISAAR (CPF) e ISDF, bem como com as normas nacionais adotadas.

3.4. Norma ISDF

A Norma ISDF[12] foi elaborada pelo Comitê de Boas Práticas e Normas em Dresden, Alemanha, em maio de 2007, com o objetivo de fornecer "diretivas para a preparação de descrições de funções de entidades coletivas associadas à produção e manutenção de arquivos" (Conselho Internacional de Arquivos 2007). Assim, a norma esclarece que informação deve ser incluída nas descrições, orientando como estas descrições podem ser desenvolvidas num sistema de arquivo. Porém, o conteúdo dos elementos que está incluído em cada descrição deverá ser previamente determinado pelas regras que a instituição adote.

Estes elementos de descrição estão organizados em quatro áreas:

- **Área de Identificação**: procura definir um ponto de acesso normalizado, bem como mostrar, de forma específica, a função, onde constam elementos como o tipo, a(s) forma(s) autorizada(s) do nome, a(s) forma(s) paralela(s) do nome e, também, outra(s) forma(s) do nome e a classificação;
- **Área de Contextualização**: é o local onde está registada a informação sobre a natureza e contexto da função, tendo elementos como as datas, a descrição, a história e

[12] ISDF: Norma internacional para descrição de funções.

a legislação;

- **Área de Relacionamentos**: diz respeito ao registo e à descrição das relações com outras funções e possui elementos como a(s) forma(s) autorizada(s) do nome/Identificador da função relacionada, o tipo, a categoria do relacionamento, a descrição e as datas do relacionamento;
- **Área de Controlo**: possui uma descrição da função, bem como da instituição de arquivo, tendo um conjunto de elementos como o identificador da descrição da função, os identificadores da instituição, as regras e/ou convenções utilizadas, o status, o nível de detalhamento, as datas de criação, revisão ou obsolescência, o(s) idioma(s) e sistema(s) de escrita, as fontes e as notas de manutenção.

Perante o exposto, é referir que existem três elementos que são fulcrais: o tipo, a(s) forma(s) autorizada(s) do nome e o identificador da descrição da função. Contudo, diversos elementos podem ser utilizados como pontos de acesso, através do seguimento de regras ou convenções para a sua normalização. É, também, essencial que a norma seja utilizada em sintonia, nomeadamente, com as normas ISAD (G) e ISAAR (CPF), bem como com normas nacionais adotadas.

4. Modelos de dados ligados para o Património Cultural

No presente capítulo apresenta-se o modelo CIDOC CRM, bem como modelos baseados nesta ontologia utilizados em diversas áreas.

4.1. CIDOC CRM

O CIDOC CRM (*CIDOC Conceptual Reference Model*) é uma ontologia formal de referência, para melhorar a integração e a troca de informação no âmbito do Património Cultural. Isto é realizável através da definição de uma estrutura formal que é utilizada para descrever os conceitos e as relações em uso na documentação do Património Cultural. Assim, esta ontologia foi desenvolvida para promover a compreensão da informação sobre o Património Cultural e para que existisse uma linguagem comum para a formulação de requisitos em sistemas de informação (CIDOC CRM 2009a).

O CIDOC CRM é o resultado final de um trabalho realizado durante mais de vinte anos por parte do Comité Internacional para a Documentação do Conselho Internacional de Museus (ICOM) e foi desenvolvido tendo como base experiências de diversos projetos de integração de informação, com os objetivos de:

- Resolver algumas questões relacionadas com a documentação do património cultural para que a troca de informação em contexto global fosse possível;
- Permitir tanto a integração como a troca de informação sem perda de semântica;
- Fornecer uma infraestrutura extensível e com parâmetros bem definidos (Santos 2016).

Uma das inovações do CIDOC CRM é a estruturação de informação à volta de eventos temporais. Ou seja, os eventos são definidos "como entidades que agregam atores, factos e objetos (físicos e abstratos), localidades e duração de intervalo de tempo" (Santos 2016). Assim, permite "integrar o maior número de recursos de informação, proporcionando uma maior flexibilidade entre sistemas para se tornar mais compatível" (Santos 2016). A integração da informação é o foco do CIDOC CRM e esta deve ir para além dos limites de um documento com o intuito de relacionar a informação com um contexto mais amplo (Doerr 2005). Importa, de igual forma, realçar que o CIDOC CRM é uma ontologia que foi desenvolvida para museus. Contudo, já foi aplicada em outras áreas, conforme será apresentado.

4.2. Modelos baseados no CIDOC CRM

4.2.1. CRMarcheo

Este modelo foi desenvolvido para apoiar o processo de escavação arqueológica, bem como todas as entidades e atividades relacionadas com esta. O modelo foi criado a partir de normas e modelos que já eram utilizados por instituições ligadas ao Património Cultural, evoluindo através de uma análise profunda dos metadados existentes e de documentação arqueológica real. Além disto, aproveita os conceitos fornecidos pela extensão *CRMsci*, herdando a maior parte dos princípios geológicos e estratigráficos. Neste sentido, o seu objetivo é facilitar a codificação semântica, o intercâmbio, a interoperabilidade e o acesso à documentação (CIDOC CRM 2009b).

4.2.2. CRMsci

Este modelo foi definido para ser utilizado como um esquema global para integrar metadados sobre observação científica, medições e dados processados em ciências descritivas e empíricas. O seu principal objetivo é facilitar a gestão, a integração, o intercâmbio, a mediação e o acesso a dados de investigação, através da descrição de relações semânticas (CIDOC CRM 2009g).

4.2.3. CRMdig

Este modelo serve para codificar metadados sobre as etapas e métodos de produção de produtos de digitalização e representações digitais sintéticas e distingue-se, comparativamente a outros modelos, pela inclusão dos processos iniciais de medição física e dos seus parâmetros, conseguindo ter descrições completas codificadas (CIDOC CRM 2009d).

4.2.4. CRMba

Este modelo serve para codificar metadados sobre a documentação de edifícios arqueológicos. Foi concebido para apoiar o processo de registo de edifícios arqueológicos e para registar as relações entre cada um dos componentes do edifício, bem como do edifício como um todo. É de referir, ainda, que este modelo incorpora partes do modelo *CRMgeo* e, também, dos modelos *CRMsci* e *CRMarchaeo* (CIDOC CRM 2009c).

4.2.5. CRMgeo

Este modelo deve ser utilizado para integrar propriedades espácio-temporais nas entidades temporais. Ou seja, o seu principal objetivo é dar um modelo relacionado com o CIDOC CRM para integrar a geoinformação, utilizando concetualizações, definições, padrões de codificação, bem como relações topológicas (CIDOC CRM 2009e).

4.2.6. CRMtex

Este modelo foi criado com o intuito de apoiar o estudo de documentos antigos, através da identificação de entidades textuais relevantes e da modelação do processo científico que está relacionado com a investigação destes documentos. O *CRMtex* destina-se, então, à identificação e definição das principais entidades envolvidas no estudo e na edição de textos manuscritos antigos para depois os descrever (CIDOC CRM 2009h).

4.2.7. CRMinf

Este modelo deve ser utilizado como um esquema global para integrar metadados sobre a argumentação e a elaboração de inferências em ciências descritivas e empíricas. O seu principal objetivo é "facilitar a gestão, a integração, a mediação, o intercâmbio e o acesso a dados sobre o raciocínio" (CIDOC CRM 2009f), através de uma descrição das relações semânticas entre as premissas, as conclusões e as atividades do raciocínio.

4.3. Projetos de interesse

Neste ponto são abordados projetos de interesse que foram selecionados porque se enquadram no domínio dos dados ligados no Património Cultural, sendo que alguns possuem, também, o CIDOC CRM como base das ontologias utilizadas.

4.3.1. Projeto ARIADNEplus

O Projeto ARIADNEplus é uma extensão do projeto ARIADNE (*Archaeological Research Infrastructure for Archaeological Data Networking in Europe*) e está ligado ao património arqueológico, sendo baseado na utilização do CIDOC CRM (ARIADNE 2019). Este projeto conseguiu integrar conjuntos de dados arqueológicos no seu registo, com mais de um milhão

de conjuntos de dados registados e geridos de acordo com os princípios FAIR (***F**indable*, ***A**ccessible*, ***I**nteroperable*, ***R**eusable*).

Uma das características mais importante que este projeto realça é a extensibilidade do CIDOC CRM dado que é possível acrescentar conceitos em qualquer subdomínio específico sempre que for necessário, sem que este perca a compatibilidade com outros subdomínios ou com o domínio (ARIADNE 2016). Isto é possível pois o CIDOC CRM possui uma equipa que, frequentemente, revê e atualiza a ontologia para responder a novos desafios que vão sendo colocados. Assim, criou diversas extensões para responder a necessidades de representação específicas em domínios gerais. Desta forma, cada extensão apresentada seguidamente é uma ontologia própria, onde as classes e as propriedades são especificações do CIDOC CRM (ARIADNE 2016).

4.3.2. Polish Digital Libraries Federation

Outro exemplo da aplicabilidade do CIDOC CRM fora do contexto de museus é o caso da *Polish Digital Libraries Federation* que defende que esta ontologia serve para "representar informação sobre o Património Cultural, especialmente coleções de museus" (Mazurek et al. 2011), que esta ontologia não é uma escolha natural para descrever coleções de bibliotecas digitais e que, contudo, foi escolhida como o formato de descrição, uma vez que foi criada para representar itens do Património Cultural, "que são definitivamente coleções de livros e impressões antigas, mas suficientemente universais para permitir a sua saída do mundo dos museus" (Mazurek et al. 2011).

A *Polish Digital Libraries Federation* possuía outras alternativas, além do CIDOC CRM, mas esta ontologia foi a escolhida, devido à sua:

- **Universalidade**: capaz de descrever tanto recursos bibliográficos, como obras de arte e artigos de museu;
- **Simplicidade**: organização numa hierarquia clara para todos os utilizadores;
- **Maturidade**: desenvolvida, mantida e utilizada frequentemente;
- **Popularidade**: o CIDOC CRM é conhecido e compreensível quando ligado a objetos do Património Cultural online.

Perante isto, a *Polish Digital Libraries Federation* considera que, apesar de desenvolvido para

a representação de conhecimento em contexto de museus, o CIDOC CRM é "suficientemente expressivo para descrever informação proveniente de bibliotecas digitais" (Mazurek et al. 2011).

4.3.3. DIESEL Project - Archive Linked Open Data (aLOD)

O projeto DIESEL é coordenado pelo Arquivo Federal da Suíça (*Swiss Federal Archive*), tendo parcerias com os Arquivos do Estado de Neuchâtel, os Arquivos do Estado do Valais, os Arquivos do Estado de Genebra, os Arquivos Cantonais de BaselStadt e os Arquivos da Cidade de Baden. Este projeto visa "dinamizar uma infraestrutura para efetuar pesquisas de palavras-chave e de respostas a algumas questões para dados relativos a organizações" (DIESEL 2014). Esta infraestrutura examina a forma como os dados estão distribuídos com o intuito de melhorar a interpretação e a execução das pesquisas de cada utilizador, respeitando as suas restrições de acesso. O projeto fornece, desta forma, uma versão em código aberto da infraestrutura com uma funcionalidade para pesquisa de palavras-chave.

Assim, desenvolveu uma nova abordagem, que permite efetuar uma pesquisa das políticas, através de fontes de informação distribuídas, a qual é mantida através da implementação e da adaptação de soluções que geram dados ligados de uma forma não-invasiva.

É, ainda, importante mencionar que, como o acesso e a integração da informação constituem os principais obstáculos para aproveitar a informação disponível, as organizações tendem a distribuí-la por diversos silos de dados em diferentes formatos. Diversas soluções de pesquisa semântica dependem de extensões de pesquisa de texto (por exemplo, *thesauri* com um domínio específico) e não fazem uso da semântica presente na informação no momento de interpretação de pesquisas. Ou seja, a semântica empregada pelo utilizador é utilizada para gerar pesquisas a partir de palavras-chave.

O resultado final do projeto é a geração de um novo paradigma de pesquisa, através de *Linked Enterprise Data*, com o objetivo de "aumentar a reutilização do conhecimento interno da organização, bem como a produtividade dos colaboradores e uma melhor utilização dos recursos internos" (DIESEL 2014).

O projeto DIESEL possui uma plataforma designada de **aLOD** – *Archive Linked Open Data*. Esta plataforma é gerida pelo Arquivo Federal da Suíça, bem como pelos Arquivos dos Estados de Neuchâtel, Valais, Genebra e pelos Arquivos Cantonais de Basel-Stadt. Através desta plataforma são feitas experiências relativas ao fornecimento de dados de arquivo como dados ligados. O objetivo é entender os "requisitos e oportunidades do fornecimento de informação

arquivística como dados ligados, compreender os requisitos técnicos para a criação e o fornecimento de dados ligados, partilhar o know-how com outros Arquivos, bem como tentar melhorar a facilidade de acesso para os utilizadores" (DIESEL 2017).

4.3.4. ArCo

O projeto ArCo é um "grafo do conhecimento do património cultural italiano" (ArCo 2019a) que se estabelece em diversos vocabulários para a descrição no domínio do Património Cultural do Catálogo Geral do Ministério Italiano do Património e Atividades Culturais.

A ontologia deste projeto está organizada através de uma "rede que modela os diferentes tipos de bens culturais existentes e os respetivos registos no catálogo" (ArCo 2019b). Atualmente tanto os bens culturais como os registos são geridos pelo sistema SIGEC[13] do Instituto Central Italiano de Catalogação e Documentação e o seu objetivo é recolher e validar os registos do catálogo dos bens culturais, excluindo bibliotecas e arquivos.

Como este projeto aborda assuntos como a catalogação, dados abertos ligados e ontologias, está destinado à realização de atividades como:

- Definição do processo de aquisição e produção de grafos de conhecimento, efetuando a sua ligação ao Catálogo Geral Oficial, através da extração automática de texto;
- Desenvolvimento de um projeto piloto destinado a melhorar o modelo ontológico, para que a comunidade científica também faça uso do grafo de conhecimento;
- Apoiar a criação de ferramentas para a aquisição e conversão de dados sobre o Património Cultural em dados interligados;
- Elaboração de um catálogo de serviços para a biblioteca digital.

4.3.5. SARI Documentation

O SARI – *Swiss Art Research Infrastructure* – é um projeto que tem o objetivo de dar acesso a dados de investigação específicos, a dados de recolha e a recursos visuais digitalizados. Estes dados são relativos aos campos da História, Arquitetura, Arqueologia, Estudos Religiosos e Humanidades Digitais. Através deste acesso, é criado um "ambiente de investigação de ponta

[13] *Sistema Informativo Generale del Catalogo*: plataforma baseada na web que gere todo o fluxo de catalogação.

para o ensino e investigação" (Universität Zürich 2018) que melhora a visibilidade e a acessibilidade aos dados de investigação.

O projeto utiliza uma infraestrutura da Web Semântica para "registar e expor recursos de referência úteis para projetos de História da Arte Digital e de Humanidades Digitais" (Universität Zürich 2018).

Além disto, o projeto deu origem a diversos recursos de documentação que são recolhidos em categorias como *Padrões de Modelação Ontológica*, *Modelos de Dados de Referência* e *Extensões Ontológicas*. Os *Padrões de Modelação Ontológica* traduzem-se em fórmulas que podem ser empregadas na modelação de entidades. Estes padrões resultam, então, de um estudo de caso e são explicitados através de um ficheiro Turtle. Os *Modelos de Dados de Referência* são um "modelo de descritores comuns com base na análise de fontes selecionadas" (SARI 2020) que têm relevância para a entidade a ser modelada. É possível fornecer implementações de referência a serem utilizadas por instituições e projetos não familiarizados com o CIDOC CRM, bem como criar diretrizes para gerar interfaces de entrada de dados semânticos do CIDOC CRM. Finalmente, a *Extensão Ontológica* diz respeito às propriedades e classes criadas como uma extensão do CIDOC CRM.

4.3.6. LinkedArt

O Linked Art é uma "comunidade que trabalha para criar um modelo partilhado baseado em *Linked Open Data* para descrever a Arte" (LinkedArt 2020a). Este modelo é um perfil de aplicação de metadados baseado, sobretudo, no CIDOC CRM e sobre o qual foi desenvolvido *software* capaz de adicionar padrões para a implementação que não são apropriados para o nível mais concetual da ontologia, abrangendo RDF e *Linked Open Data*.

Porém, o contexto do Linked Art é mais restrito do que o do CIDOC CRM, dado que o seu foco está nos casos de utilização de museus e não em toda a gestão do conhecimento do Património Cultural (LinkedArt 2020b).

5. A avaliação do modelo de dados

Neste capítulo estão presentes duas secções. A primeira diz respeito ao modelo de dados desenvolvido no projeto – *ArchOnto* –, explicando o que é, como está organizado, finalizando com uma comparação entre as descrições arquivísticas efetuadas pela equipa do INESC TEC e as descrições arquivísticas efetuadas pela equipa da DGLAB relativamente a vinte e cinco registos previamente escolhidos. A segunda é relativa aos metadados recolhidos de um conjunto de registos, mostrando de que forma foram organizados e tratados e, por fim, como é que estes demonstram a capacidade do modelo enquadrar metadados já existentes.

5.1. Modelo de dados

Esta primeira experiência teve como objetivo a familiarização com o projeto, através do estudo do modelo *ArchOnto* e do CIDOC CRM.

Desta forma, foi possível compreender o que é a *ArchOnto*: uma nova ontologia com diversos módulos criada no domínio dos arquivos e estruturada tendo como base o CIDOC CRM e tendo em consideração a informação presente na Norma ISAD (G). Contudo, a maior parte da informação dos elementos da ISAD (G) não seria facilmente colocada no CIDOC CRM, dado que este foi criado para museus (Koch, Ribeiro, and Teixeira Lopes 2020).

O processo de adaptação do CIDOC CRM passou por diversas fases. A primeira foi a criação da extensão *Data Property* para combater o número limitado de propriedades de dados no CIDOC CRM, onde a maioria das propriedades foram criadas para possuir informação dos campos da ISAD (G). Porém, esta abordagem levou à existência de um grande número de propriedades, conduzindo à criação da *ISAD Ontology*. Já as classes e as propriedades de objetos que existiam como extensões CIDOC CRM foram movidas para a ontologia *N-ary*. No seu conjunto, a ArchOnto possui cinco ontologias na sua base, além do CIDOC CRM. A ontologia *N-ary* foi criada para sistematizar a representação das instâncias onde se revela mais natural utilizar associações ternárias ou de cardinalidade superior. A *ISAD Ontology* foi desenvolvida para representar todos os elementos presentes na Norma ISAD (G) e, assim, é mantida toda a informação relativa às descrições arquivísticas existentes, sendo, também, possível manter a interoperabilidade de todos os dados representados na ISAD (G). A ontologia *DataObject* foi criada para fazer a validação dos valores utilizados nas propriedades do novo modelo de dados, possuindo classes e propriedades de dados utilizadas para validar dados na ontologia. Finalmente, para ligar a ontologia do CIDOC CRM com a ontologia *DataObject*, foi criada uma ontologia com a propriedade *hasValue* com o intuito de conectar as ontologias

(Koch, Ribeiro, and Teixeira Lopes 2020).

A seguir, foi efetuada uma comparação entre as descrições arquivísticas efetuadas pela equipa do INESC TEC e as descrições arquivísticas efetuadas pela equipa da DGLAB relativamente a vinte e cinco registos escolhidos de forma cuidadosa, sendo dez de diferentes níveis de descrição e quinze de classes representativas de registos, com diferentes séculos e diferentes suportes.

Observou-se que, enquanto a DGLAB representou de uma forma mais gráfica, o INESC TEC representou através de uma ontologia OWL.

O modelo utilizado pela equipa da DGLAB é muito estrito no uso do CIDOC CRM e na sua representação é possível observar que cada registo é uma agregação de documentos. O registo em si é referente a instrumentos de descrição, onde cada um possui um título, um tipo e um identificador. Já a agregação de documentos refere-se a um registo de descrição que tem um título, com um identificador.

É de mencionar o caso da propriedade *"P1 identifies"* que serve para identificação de, por exemplo, *"domain"*, *"range"*, *"notation"* e *"rds:label"* e que possui, entre outras, uma subpropriedade *"P48 is preferred identifier of"*. Neste caso concreto, a equipa da DGLAB utilizou a subpropriedade enquanto a equipa do INESC TEC utilizou a propriedade.

No que concerne à representação feita pelo INESC TEC, é possível observar a existência de diversas classes, nomeadamente *"Entidade Temporal"*, *"Local"*, *"Evento"*, *"Atividade"* e *"Dimensão"*. Nesta ontologia existem, também, as *Data Property* que servem para atribuir valores de dados a objetos e as *Object Property* que servem para ligar um indivíduo a outro indivíduo. Por sua vez, os *Indivíduos* são cerca de mil e quinhentos, entre os quais estão representados: datas ou intervalos temporais, autores, bem como listas de autoridades e outros documentos que definem uma determinada terminologia.

Nesta mesma representação existe, de igual forma, uma *Object Property* para especificar que existem diferentes níveis de descrição (*"ARP12 has level of description"*) que faz com que seja percetível entender qual o nível de descrição do documento em questão e outra (*"P106 is composed of")* para especificar que este documento é composto por outros documentos e que pertencem ao mesmo fundo de arquivo.

Por fim, o modelo da equipa da DGLAB não faz uso da norma ISAD (G), contrariamente ao modelo da equipa do INESC TEC, através de *Data Property* como, por exemplo, *"ISAD1 has title"*, *"ISAD7 has administrative history"*, *"ISAD24 has conditions governing reproduction"*, *"ISAD9 has scope"*.

5.2. Caracterização do processo de migração

A segunda atividade de avaliação desta dissertação visa contribuir para a terceira experiência do projeto EPISA, referida na secção 1.1. Tal como na secção anterior, o objetivo é avaliar qualitativamente a expressividade do modelo de dados desenhado sobre a ontologia ArchOnto, mas agora comparando a informação contida num conjunto de registos do sistema de descrição arquivística em uso, designado DigitArq, com a informação obtida após uma atividade de migração automática desses registos para o novo modelo. Para iniciar este processo de avaliação qualitativa do processo de migração foi necessário selecionar uma amostra de diversos registos presentes no **DigitArq**.

O DigitArq é um software de Gestão de Arquivos criado com vista à "simplificação e otimização do trabalho num Arquivo Definitivo" (DGLAB 2011) e é constituído por cinco módulos que procuram responder às necessidades de um profissional dos arquivos, sendo estes "a descrição arquivística, gestão de projetos de digitalização, publicação Web e a navegação e pesquisa" (DGLAB 2011). Este *software* assenta em três normas internacionais de descrição arquivística: ISAD, EAD (*Encoded Archival Description*) e OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*).

A amostra foi estruturada num conjunto de folhas de cálculo com o objetivo de funcionar como referência para avaliar o resultado do processo de migração. Foram incluídos 16372 unidades de descrição correspondentes a documentos simples e 129 correspondentes a documentos compostos, tipicamente unidades de instalação (livros) cujo conteúdo em documentos simples não se encontra descrito no DigitArq.

Apenas uma parte desta amostra foi efetivamente migrada, pelos que os resultados apresentados abaixo se referem apenas a esse subconjunto. A amostra completa será útil quando o resto da amostra for submetida a um ou vários processos de migração, para suportar a avaliação desses processos.

Descreve-se em seguida o conteúdo da amostra, a qual contém unidades de descrição relativas a batismos, casamentos e óbitos do **Arquivo Distrital da Guarda**; batismos, casamentos e óbitos do **Arquivo Distrital do Porto**; e, por fim, passaportes deferidos do **Arquivo Distrital de Bragança**.

5.2.1. Arquivo Distrital da Guarda

Os registos analisados pertencentes ao **Arquivo Distrital da Guarda** estão presentes num fundo denominado *"Paróquia da Sé – Guarda 1860/1911-03-31"* que contém sete séries.

Porém, foram apenas analisadas as séries relativas aos batismos[14], aos casamentos[15] e aos óbitos[16]. As restantes séries são relativas a registos de reconhecimentos e legitimações[17], livros de duplicados dos registos de batismos[18], livros de duplicados dos registos de casamentos[19] e livros de duplicados dos registos de óbitos[20].

Este fundo é relativo à paróquia da Sé, fundada na época medieval, sendo uma das paróquias urbanas da cidade da Guarda. Possui os documentos produzidos entre 1860 e 1911 e os documentos originais presentes derivam da Conservatória do Registo Civil da Guarda, tendo sido incorporados no Arquivo Distrital em 1992, em 1999 e, também, em 2012. Relativamente aos duplicados existentes – séries não analisadas –, é de mencionar que foram incorporados em 1987, em 1991 e em 1999. Contudo, estes são procedentes da Conservatória do Registo Civil de Viseu (Arquivo Distrital da Guarda 2011a).

Em cada série, todos os documentos foram descritos e organizados pela ordem de data de produção, não possuindo qualquer restrição de acesso, exceto os documentos que apresentam um mau estado de conservação, o qual condiciona, também, a sua reprodução. Por sua vez, os livros de registo de batismos entre os anos de 1622 e 1673, 1677 e 1859; os livros de registo de casamentos entre os anos de 1601 e 1640, 1645 e 1859 e os livros de registo de óbitos entre os anos 1601 e 1740, 1742 e 1859 são detidos pelo Arquivo Nacional da Torre do Tombo (Arquivo Distrital da Guarda 2011a).

No que concerne à série dos **batismos**, esta possui registos entre janeiro de 1860 e março de 1911, com uma extensão de cinquenta e dois livros. Estes livros estão organizados sob a forma de documentos compostos que, por sua vez, possuem entre 58 e 146 documentos simples, conforme é possível observar na Tabela 1. O Arquivo Distrital tem à disposição um microfilme que permite que os utilizadores consultem os registos entre 1622 e 1673 e entre 1677 e 1859. Contudo, os registos originais estão presentes no Arquivo Nacional da Torre do Tombo

[14] https://digitarq.adgrd.arquivos.pt/details?id=1199346

[15] https://digitarq.adgrd.arquivos.pt/details?id=1199379

[16] https://digitarq.adgrd.arquivos.pt/details?id=1199432

[17] https://digitarq.adgrd.arquivos.pt/details?id=1307261

[18] https://digitarq.adgrd.arquivos.pt/details?id=1256469

[19] https://digitarq.adgrd.arquivos.pt/details?id=1256496

[20] https://digitarq.adgrd.arquivos.pt/details?id=1256520

(Arquivo Distrital da Guarda 2011b). Os casos em que o número de registos é 1 correspondem a unidades de instalação cujo conteúdo não se encontra descrito de forma mais detalhada, ao nível de documento simples, de registo individual de um batismo.

Tabela 1 - Organização dos registos de batismo do Arquivo Distrital da Guarda

Documento composto	Número de registos
1	78
2	72
3	68
4	59
5	65
6	58
7	64
8	64
9	64
10	63
11	76
12	63
13	74
14	72
15	91
16	103
17	90
18	84
19	98
20	101
21	132
22	146
23	131
24	143
25	125
26	121
27	108
28	1
29	1
30	127
31	132
32	127
33	112
34	113
35	113
36	105
37	103
38	117
39	83
40	96
41	114
42	1
43	1
44	1
45	1
46	1
47	1
48	1
49	1
50	1
51	1
52	1

Seguidamente, apresenta-se um conjunto de tabelas que explicam o significado de cada etiqueta utilizada na amostra. A Tabela 2 mostra os elementos de informação utilizados em geral para a recolha dos metadados de todos os registos. Estes metadados foram recolhidos essencialmente dos campos ISAD Código de Referência, Datas de Produção, Âmbito e Conteúdo. A Tabela 3 mostra as etiquetas utilizadas nos casos dos documentos compostos 00028, 00029 e 00042 até 00052, que são livros de conteúdo não descrito. Por fim, a Tabela 4 mostra etiquetas mais específicas foram sendo criadas e, consequentemente, utilizadas à

medida que os metadados apareciam nos registos analisados e apenas nesses registos.

Tabela 2 - Tabela com as etiquetas gerais (registos de batismo)

FIELD	Explicação
RefCode	Código de referência
NomeBatismo	Nome da pessoa batizada
DateNasc	Data de Nascimento
NrRegisto	Número do registo de batismo
Pai	Nome do pai
NatPai	Naturalidade do pai
Mãe	Nome da mãe
NatMãe	Naturalidade da mãe

Tabela 3 - Tabela utilizada nos documentos compostos não detalhados (registos de batismo)

FIELD	Explicação
RefCode	Código de referência
Localidade	Localidade onde os registos pertencem
Âmbito	Âmbito e conteúdo do documento
Condições	Condições de acesso
Dimensão	Dimensão do documento

Tabela 4 - Tabela de etiquetas específicas (registos de batismo)

Notas	Notas de procuração
LocalNasc	Local de nascimento da pessoa batizada
Falecimento	Data de falecimento
LocalFal	Local de falecimento
LocalCas	Local de casamento
DateCas	Data de casamento
Esposa	Nome da esposa da pessoa batizada
LocalFalEsposa	Local de Falecimento da esposa
DateFalEsposa	Data de Falecimento da esposa
NatEsposa	Naturalidade da esposa
IdadeEsposa	Idade da esposa da pessoa batizada
PaiEsposa	Nome do pai da esposa
MãeEsposa	Nome da mãe da esposa
Esposo	Nome do esposo da pessoa batizada
NatEsposo	Naturalidade do esposo
LocalFalEsposo	Local de falecimento do esposo
FalecimentoEsposo	Data de falecimento do esposo
IdadeEsposo	Idade do esposo da pessoa batizada
PaiEsposo	Nome do pai do esposo
MãeEsposo	Nome da mãe do esposo
NatMãeEsposo	Naturalidade da mãe do esposo
Divorcio	Data do divórcio
EsposaSeg	Nome da segunda esposa
DateCasSeg	Data do segundo casamento
LocalCasEsposaSeg	Local do segundo casamento
LocalFalEsposaSeg	Local de falecimento da segunda esposa
FalecimentoEsposaSeg	Data de falecimento da segunda esposa
IdadeEsposaSeg	Idade da segunda esposa
NatEsposaSeg	Naturalidade da segunda esposa
EsposoSeg	Nome do segundo esposo
DateFalEsposoSeg	Data de falecimento do segundo esposo
PaiEsposaSeg	Nome do pai da segunda espoa
MãeEsposaSeg	Nome da mãe da segunda espoa
EsposaTer	Nome da terceira esposa
DateCasTer	Data do terceiro casamento
LocalCasEsposaTer	Local do terceiro casamento
NatEsposoSeg	Naturalidade do segundo esposo
LocalFalEsposoSeg	Local de falecimento do segundo esposo
FalecimentoEsposoSeg	Data de falecimento do segundo esposo
IdadeEsposoSeg	Idade do segundo esposo
MãeEsposoSeg	Nome da mãe do segundo esposo

A amostra foi organizada da seguinte forma. Para cada série utilizada, existe uma folha com um resumo do número de unidades de descrição por unidade de instalação, semelhante à Tabela 1. Existe a seguir uma folha com a lista das etiquetas gerais e específicas utilizadas, semelhante às Tabelas 2, 3 e 4. Nos casos das séries de registos de casamento e de óbitos, a amostra conclui-se com uma folha com os campos recolhidos relativamente a cada unidade de instalação. Nos casos das séries de registos de batismo e de passaportes, a estrutura é mais complexa, existindo uma folha para cada unidade de instalação, com tantos registos quantos os documentos simples que contém, isto é, cada batismo e cada passaporte, detalhando-se as

várias etiquetas consideradas relevantes para a avaliação da migração. Estas folhas, designadas pelo código de referência da unidade de instalação, incluem o URL referente à unidade de descrição do documento composto no DigitArq, para facilidade de consulta, com o respetivo código de referência e as datas de produção (Tabela 5).

Tabela 5 - Exemplo de cabeçalho de identificação de um livro de registos de batismos

https://digitarq.adgrd.arquivos.pt/details?id=1199347	RefCode	PT/ADGRD/PRQ/PGRD42/001/00001
Dataset Guarda 1860-01-02 a 1860-12-27		

Abaixo deste cabeçalho, foi registada a informação considerada relevante de cada um dos correspondentes documentos simples, incluindo o URL da sua descrição e as linhas com etiquetas que identificam o código de referência, o nome da pessoa batizada, a data de nascimento, o número do registo de batismo, o nome do pai e a sua naturalidade, o nome da mãe e a sua naturalidade (Tabela 6). Porém, existem, também, outras etiquetas que foram criadas à medida que a informação ia surgindo, não estando presentes em todas os registos, como, por exemplo, local de nascimento da pessoa batizada, data e local de falecimento, bem como data e local de casamento (Tabela 4).

Tabela 6 - Exemplo de um registo de batismo na amostra

https://digitarq.adgrd.arquivos.pt/details?id=1354415	**FIELD**	**ORIGINAL**
	RefCode	PT/ADGRD/PRQ/PGRD42/001/00001/000002
	NomeBatismo	Maria
	DateNasc	1859-12-12
	NrRegisto	2/1860
	Pai	João Henriques de Andrade
	NatPai	Gaia, Belmonte
	Mãe	Maria Justina de Oliveira
	NatMãe	Almeida

A título de exemplo deste caso concreto, é possível ver, através da Figura 1, que os metadados recolhidos foram retirados, respetivamente, dos campos do *Código de referência,* das *Datas de produção* e do *Âmbito e conteúdo*. O mesmo acontece com todos os registos de batismo analisados ao longo desta série, com exceção dos documentos compostos 00028, 00029 e 00042 até 00052.

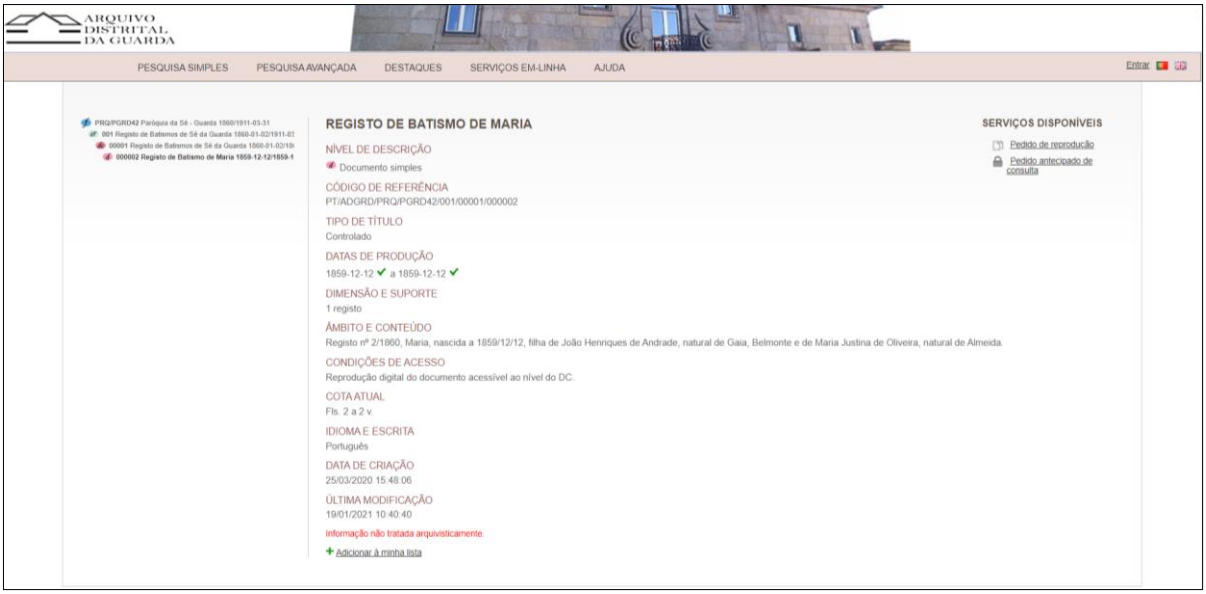


Figura 1 - Exemplo de uma descrição de um batismo (Arquivo Distrital da Guarda)

Devido ao facto de os documentos compostos 00028, 00029, 00042 até 00052 não possuírem qualquer documento simples, os metadados recolhidos centraram-se nos dados relativos ao código de referência, à localidade pertencente aos registos, ao âmbito e conteúdo do documento, às condições de acesso e, por fim, à dimensão do documento, conforme mostra a Tabela 7.

Tabela 7 - Exemplo de um registo na amostra de um livro não detalhado

https://digitarq.adgrd.arquivos.pt/details?id=1199367	**FIELD**	**ORIGINAL**
	RefCode	PT/ADGRD/PRQ/PGRD42/001/00028
	Localidade	/Sé/Guarda/Guarda/Portugal
	Âmbito	
	Condições	Reprodução digital do documento acessível a partir do repositório do ADGRD.
	Dimensão	1 liv.

Nestes registos de batismo, onde a tabela com os parâmetros recolhidos foi diferente, os metadados recolhidos foram retirados dos campos do *Código de referência*, da *Localidade*, do *Âmbito e conteúdo* (quando existente), das *Condições de acesso* e, por fim, da *Dimensão e suporte* (Figura 2).



Figura 2 - Exemplo de uma descrição de um livro de registos de batismo (Arquivo Distrital da Guarda)

Relativamente à série dos **casamentos**, esta tem registos entre 1860 e março de 1911, com uma extensão de cinquenta e dois livros. Semelhantemente a alguns registos de batismo, estes livros estão descritos apenas ao nível de documento composto, sem qualquer documento simples. O Arquivo Distrital tem na sua posse um microfilme que permite que os utilizadores consultem os registos dos casamentos realizados entre 1601 e 1640 e entre 1645 e 1859. Contudo, os registos originais estão presentes no Arquivo Nacional da Torre do Tombo (Arquivo Distrital da Guarda 2011c).

Neste caso, a lista de etiquetas e do seu significado mostra-se na Tabela 8.

Tabela 8 - Explicação das etiquetas utilizadas (registos de casamento)

FIELD	Significado
RefCode	Código de referência
Localidade	Localidade onde ocorreu o casamento
Âmbito	Âmbito e conteúdo do documento
Condições	Condições de acesso
Dimensão	Dimensão do documento

Para a recolha dos metadados destes registos na amostra, foi criada uma única folha para a série, com o respetivo nome. Nesta folha, semelhante ao exposto na série dos batismos, existe um cabeçalho com o URL da descrição da série, o código de referência, e a data de produção (Tabela 9).

Tabela 9 - Cabeçalho de identificação da série relativa aos casamentos

https://digitarq.adgrd.arquivos.pt/details?id=1199379	RefCode	PT/ADGRD/PRQ/PGRD42/002

Dataset Guarda 1860 a 1911-03-30

Os registos na amostra referentes a cada livro possuem, igualmente, o URL do documento composto, seguindo-se as linhas com etiquetas que identificam o código de referência, a localidade onde foi produzido o livro, o âmbito e conteúdo do documento, as condições de acesso e, finalmente, a dimensão do documento (Tabela 10).

Tabela 10 - Exemplo de um registo na amostra de um livro de registos de casamento

https://digitarq.adgrd.arquivos.pt/details?id=1199380	FIELD	ORIGINAL
	RefCode	PT/ADGRD/PRQ/PGRD42/002/00001
	Localidade	/Sé/Guarda/Guarda/Portugal
	Dimensão	1 liv.
	Âmbito	
	Condições	Consulta e reprodução proibida.

No caso do registo mostrado acima, os metadados são provenientes dos campos do *Código de referência*, da *Localidade*, da *Dimensão e suporte*, do *Âmbito e conteúdo* (quando existente) e das *Condições de acesso*, conforme mostra a Figura 3.



Figura 3 - Exemplo de uma descrição de um livro de registos de casamento (Arquivo Distrital da Guarda)

Por fim, na série dos **óbitos** existem registos entre janeiro de 1860 e março de 1911, com uma extensão de cinquenta e dois livros. A amostra desta série está organizada de forma semelhante à da dos casamentos, pois nenhum livro possui qualquer documento simples. O Arquivo Distrital tem, igualmente, um microfilme que permite aos utilizadores consultar os registos dos óbitos ocorridos entre 1601 e 1740 e entre 1742 e 1859, estando os registos originais presentes no Arquivo Nacional da Torre do Tombo (Arquivo Distrital da Guarda 2011d).

Foi, portanto, criada uma folha resumo, a que se segue uma folha a explicar o significado de cada etiqueta utilizada, conforme mostrado na Tabela 11. Seguidamente, de igual forma, criou-se outra folha com o nome da série, com o cabeçalho (Tabela 12) e uma folha com os registos de cada livro de óbitos, com o URL da descrição e as linhas com etiquetas para o código de referência, a localidade onde foi produzido o livro, a dimensão do documento e, por fim, as condições de acesso (Tabela 13).

Tabela 11 - Significados das etiquetas (registos de óbitos)

FIELD	ORIGINAL
RefCode	Código de referência
Localidade	Localidade onde ocorreu o óbito
Dimensão	Dimensão do documento
Condições	Condições de acesso

Tabela 12 - Cabeçalho da amostra da série relativa aos óbitos

https://digitarq.adgrd.arquivos.pt/details?id=1199432	RefCode	PT/ADGRD/PRQ/PGRD42/003

Dataset Guarda 1860-01-24 a 1911-03-27

Tabela 13 - Exemplo de um registo na amostra de um livro de registos de óbito

https://digitarq.adgrd.arquivos.pt/details?id=1199455	FIELD	ORIGINAL
	RefCode	PT/ADGRD/PRQ/PGRD42/003/00003
	Localidade	/Sé/Guarda/Guarda/Portugal
	Dimensão	1 liv.
	Condições	Reprodução digital do documento acessível a partir do repositório do ADGRD.

Neste registo concreto, os metadados são provenientes dos campos do *Código de referência*, da *Localidade*, da *Dimensão e suporte* e das *Condições de acesso*, conforme mostra a Figura 4.



Figura 4 - Exemplo de uma descrição de um livro de registos de óbito (Arquivo Distrital da Guarda)

5.2.2. Arquivo Distrital do Porto

Os registos pertencentes ao **Arquivo Distrital do Porto** estão presentes num fundo designado *"Paróquia de Aldoar 1640-05-15/1911-03-31"* que contém seis séries. Contudo, foram analisadas somente as séries relativas aos batismos[21], aos casamentos[22] e aos óbitos[23]. As restantes séries dizem respeito aos usos e costumes[24], aos legados e obrigações de missas[25] e, por fim, aos índices de registos[26].

Este fundo é relativo à paróquia de Aldoar e todos os registos datam entre 1640 e 1911. Esta freguesia pertenceu ao concelho de Bouças, tendo sido, posteriormente, anexada à cidade do Porto, em novembro de 1895. Por sua vez, os documentos aqui presentes foram reunidos pela 2ª Conservatória do Registo Civil do Porto no decorrer dos anos de 1998, 2004 e 2013. Já a sua ordenação é feita cronologicamente, enquanto o seu acesso e a sua reprodução estão condicionados, dependendo do estado de conservação de cada documento, bem como ao tipo de documentos e ao fim a que se destina (Arquivo Distrital do Porto 2012a).

Na série dos **batismos** existem registos entre fevereiro de 1644 e março de 1911 (Arquivo Distrital do Porto 2012b), organizados em 18 unidades de instalação e cada uma possui entre 28 e 304 documentos simples (Tabela 14).

[21] https://pesquisa.adporto.arquivos.pt/details?id=488458
[22] https://pesquisa.adporto.arquivos.pt/details?id=488477
[23] https://pesquisa.adporto.arquivos.pt/details?id=488490
[24] https://pesquisa.adporto.arquivos.pt/details?id=1241180
[25] https://pesquisa.adporto.arquivos.pt/details?id=1241183
[26] https://pesquisa.adporto.arquivos.pt/details?id=1241401

Tabela 14 - Distribuição dos registos de batismo do Arquivo Distrital do Porto

Unidade de instalção	Número de registos
UI 0001	1
UI 0002	161
UI 0003	287
UI 0004	109
UI 0005	127
UI 0006	304
UI 0007	145
UI 0008	184
UI 0009	149
UI 0010	156
UI 0011	175
UI 0012	178
UI 0013	201
UI 0014	210
UI 0015	189
UI 0016	44
UI 0017	105
UI 0018	28

Os dados da amostra foram organizados de forma semelhante aos do Arquivo Distrital da Guarda. As principais notas são as de que as unidades de descrição relativas aos livros UI0001, UI0002 e UI0003 não estão disponíveis e que foi utilizado um conjunto de etiquetas ligeiramente diferente (Tabela 15).

Tabela 15 - Significados das etiquetas na série de registos de batismos do Arquivo Distrital do Porto

FIELD	Explicação
RefCode	Código de referência
NomeBatismo	Nome da pessoa batizada
Pai	Nome do pai
Mãe	Nome da mãe
AMatHom	Nome do avô materno
AMatMul	Nome da avó materna
APatHom	Nome do avô paterno
APatMul	Nome da avó paterna
Padrinho	Nome do padrinho
Madrinha	Nome da madrinha
DateNasc	Data de Nascimento

A série relativa aos **casamentos** é composta por registos entre maio de 1640 e março de 1911 (Arquivo Distrital do Porto 2012c), com uma extensão de treze unidades de instalação onde

não existe qualquer documento simples.

Finalmente, na série dos **óbitos** estão presentes registos entre outubro de 1643 e março de 1911 (Arquivo Distrital do Porto 2012d), existentes ao longo de doze unidades de instalação também sem documentos simples descritos.

5.2.3. Arquivo Distrital de Bragança

Os registos analisados do **Arquivo Distrital de Bragança** estão presentes num fundo com o nome "*Governo Civil de Bragança 1834/1986*" que possui 843 livros e 123 maços.

No âmbito histórico, este fundo relata como, em 1835, surgiu a designação de Governador Civil, bem como as suas funções e competências. Ou seja, esta pessoa seria alguém com competências para funções administrativas de representação política e económica do Governo. Contudo, este cargo não surgiu nesse ano e, desta forma, especula-se que a sua origem remonta a 1832, aquando da menção ao Prefeito ou, ainda, ao Administrador Geral da Constituição de 1822 ou ao Corregedor de 1808. A origem do Prefeito surge, então, no Movimento Liberal e é através de Mouzinho da Silveira[27] que este cargo é incorporado no sistema administrativo em Portugal. Porém, com o passar do tempo, algumas competências desapareceram ou foram ajustadas com outras. Algumas destas competências passavam por fomentar a educação e a instrução pública; fomentar a agricultura, a pecuária, as minas e o comércio; criar e manter ativas escolas industriais, artes e ofícios, bem como instaurar cursos, exposições, feiras e mercados (Arquivo Distrital de Bragança 2009).

Este fundo é composto por vinte secções, nomeadamente uma secção designada "Passaportes 1844/1969" onde se encontra a série com os registos de passaportes deferidos analisados. Esta série - denominada "Registo de passaportes deferidos 1844-09-10/1969-03-24" - possui 43 unidades de instalação. Por sua vez, estas unidades de instalação possuem entre 76 e 3392 documentos. Por fim, é de referir que não existem livros com os registos destes passaportes entre dezembro de 1903 e outubro de 1904, assim como entre abril de 1905 e março de 1906 (Arquivo Distrital de Bragança 2006).

Uma dessas secções – *Passaportes 1844/1969* – é o local onde está, por sua vez, a série[28] onde se encontram as quarenta e três unidades de instalação com os registos analisados. Estas, por

[27] José Xavier Mouzinho da Silveira (1780-1849) foi jurisconsulto e político português, bem como uma das principais personalidades da revolução liberal.
[28] https://digitarq.adbgc.arquivos.pt/details?id=1228230

sua vez, possuem entre setenta e seis e três mil trezentos e noventa e dois documentos simples (Tabela 16).

Tabela 16 - Distribuição dos registos de passaportes deferidos do Arquivo Distrital de Bragança

Unidade de instalação	Número de registos
1140	76
1141	228
1142	288
1143	286
1144	383
1145	377
1146	498
1147	493
1148	391
1149	589
1150	1989
1151	2013
1152	1992
1153	1392
1154	322
1155	1987
1156	878
1157	1
1158	1
1159	1
1160	1
1161	1
1162	1
1163	1
1164	1
1165	1
1166	1
1145A	395
1145B	494
1146A	496
1146B	493
1147A	496
1148A	499
1149A	504
1149B	591
1149C	1434
1149D	1371
1149E	1988
1149F	1673
1149G	1674
1149H	3392
1149I	2408
1149J	2423

Após fazer a recolha do número de registos presente em cada unidade de instalação, procedeu-se à criação de uma folha onde consta o significado de cada etiqueta utilizada (Tabela 17).

Tabela 17 - Explicação das etiquetas utilizadas

FIELD	ORIGINAL
RefCode	Código de referência
Titular	Nome do titular do passaporte
NrRegisto	Número do registo do passaporte
Mãe	Nome da mãe
Pai	Nome do pai
Naturalidade	Naturalidade do titular do passaporte
Idade	Idade do titular do passaporte
Literacia	Literacia do titular do passaporte
Estado civil	Estado civil do titular do passaporte
Profissão	Profissão do titular do passaporte
Destino	Destino do titular do passaporte
Acomp	Acompanhantes do titular do passaporte

Tal como nas séries relativas a batismos, os dados sobre a amostra da série de passaportes estão organizados numa folha para cada unidade de instalação, com o correspondente cabeçalho (Tabela 18).

Tabela 18 - Exemplo de um cabeçalho de identificação de um livro de registos de passaportes

https://digitarq.adbgc.arquivos.pt/details?id=1228231	RefCode	PT/ADBGC/AC/GCBGC/PAS/077/01140
Dataset Bragança 1844-09-10 a 1858-10-25		

Em baixo (Tabela 19) está um exemplo de um registo da amostra que apresenta a informação considerada relevante de cada documento simples. Esta é composta pelo URL da descrição do documento em questão, seguindo-se as linhas com as etiquetas que dizem respeito ao código de referência, ao nome do titular do passaporte, bem como a sua naturalidade, a sua idade, a sua literacia, o seu estado civil, a sua profissão, o seu destino. Além destas etiquetas estão, também, presentes outras como o número do registo do passaporte, o nome dos pais e eventuais acompanhantes.

Tabela 19 - Exemplo de registo na amostra relativo a um passaporte deferido

https://digitarq.adbgc.arquivos.pt/details?id=1274310	FIELD	ORIGINAL
	RefCode	PT/ADBGC/AC/GCBGC/PAS/077/01140/000001
	Titular	Domingos Manuel Bandeira
	NrRegisto	1 - f.1
	Mãe	não consta
	Pai	não consta
	Naturalidade	Babe-Babe-Bragança
	Idade	32 anos
	Literacia	não consta
	Estado civil	Solteiro(a)
	Profissão	Barbeiro
	Destino	Rio de Janeiro - Brasil
	Acomp	

Para estes registos relativos foram recolhidos metadados dos campos do *Código de referência* e do Âmbito e conteúdo (Figura 5).



Figura 5 - Exemplo de uma descrição de um registo de passaporte deferido (Arquivo Distrital de Bragança)

Existem, ainda, outras etiquetas que foram criadas à medida que a informação foi surgindo, sendo as mais frequentes o nome da esposa, o nome dos filhos, de sobrinhos, irmãos, estando todos acompanhados, quando existentes, das respetivas idades (Tabela 20).

Tabela 20 - Exemplo de outras etiquetas utilizadas

Esposa	Nome da esposa
IdadeEsposa	Idade da esposa
NatEsposa	Naturalidade da esposa
ProfEsposa	Profissão da esposa
Esposo	Nome do esposo
Filha	Nome da flha
IdadeFilha	Idade da filha
EstadoFilha	Estado civil da filha
ProfFilha	Profissão da filha
Filho	Nome do filho
IdadeFilho	Idade do filho
ProfFilho	Profissão do filho
NatFilho	Naturalidade do filho
EstadoFilho	Estado civil do filho
Irmã	Nome da irmã
IdadeIrmã	Idade da irmã
NatIrmã	Naturalidade da irmã
Irmão	Nome do irmão
IdadeIrmão	Idade do irmão
ProfIrmão	Profissão do irmão
NatIrmão	Naturalidade do irmão

5.3. Análise de resultados da migração

Após uma reunião com um membro da equipa da Universidade de Évora, responsável pela migração de registos do DigitArq para a ArchOnto, foi possível comparar que metadados foram inicialmente considerados importantes e que metadados foram posteriormente adicionados e analisados por este membro da equipa.

O *dataset* analisado diz respeito a 35 registos de batismo do Arquivo Distrital do Porto, onde 20 registos pertencem à Unidade de Instalação 0004 e 15 registos pertencem à Unidade de Instalação 0018.

Desta forma, através da Tabela 21, relativa à unidade de instalação 0004, é possível observar que os campos tratados inicialmente foram os mesmos campos tratados por este membro da equipa. O mesmo acontece com a unidade de instalação 0018, apesar de, inicialmente, não serem recolhidos metadados relativos ao nome dos avós e dos padrinhos. Isto deveu-se ao facto de não constarem no registo, daí aparecerem nas tabelas como "NULL" (Tabela 22).

Tabela 21 - Campos tratados na Unidade de Instalação 0004

http://pesquisa.adporto.arquivos.pt/details?id=1374649	RefCode	PT/ADPRT/PRQ/PPRT01/001/0004/00003		RefCode	PT/ADPRT/PRQ/PPRT01/001/0004/00003
	Pai	Manuel Martins Ramos		Pai	Manuel Martins Ramos
	Mãe	Ana Joaquina Martins		Mãe	Ana Joaquina Martins
	AMatHom	João Martins de Oliveira		AMatHom	João Martins de Oliveira
	AMatMul	Joana Francisca da Silva		AMatMul	Joana Francisca da Silva
	APatHom	Manuel Martins Ramos		APatHom	Manuel Martins Ramos
	APatMul	Custódia Maria da Costa		APatMul	Custódia Maria da Costa
	Padrinho	António Martins de Oliveira		Padrinho	António Martins de Oliveira
	Madrinha	Josefina Maria de Jesus		Madrinha	Josefina Maria de Jesus
original = 9 de Novembro de 1811	DateNasc	1811-11-09		DateNasc	1811-11-09

Tabela 22 - Campos tratados na Unidade de Instalação 0018

http://pesquisa.adporto.arquivos.pt/details?id=1849176	RefCode	PT/ADPRT/PRQ/PPRT01/001/0018/000001		RefCode	PT/ADPRT/PRQ/PPRT01/001/0018/000001
	Pai	Domingos Rodrigues Torres		Pai	Domingos Rodrigues Torres
	Mãe	Conceição Maria d'Almeida		Mãe	de Conceição Maria
	AMatHom	NULL		AMatHom	NULL
	AMatMul	NULL		AMatMul	NULL
	APatHom	NULL		APatHom	NULL
	APatMul	NULL		APatMul	NULL
	Padrinho	NULL		Padrinho	NULL
	Madrinha	NULL		Madrinha	NULL
	DateNasc	1910-12-15		DateNasc	1910-12-15

Para chegar aos resultados que serão seguidamente apresentados, este membro da equipa começou por efetuar a conversão dos metadados em representação numérica para que fosse possível o processamento por um modelo de classificação. A seguir, para extrair as entidades e os eventos, utilizou o GATE: um conjunto de ferramentas de software em código aberto usado para resolver problemas relacionados com o processamento de texto (GATE 2018). Por sua vez, este processo está dividido em vários módulos que são geridos pela ANNIE. A ANNIE é uma coleção de algoritmos que faz extração de informação em texto não estruturado (University of Sheffield 2013).

Agrupou, também, as expressões em três categorias: **RefCode** (código de referência); **Date** (data de nascimento); **Person** (pais, avós e padrinhos). Estas são, ainda, divididas em dois grupos: **Total** (engloba todas as expressões) e **TotalNotNull** (possui expressões sobre os pais, mas caso o campo se encontre vazio, é considerado como um campo sem informação), conforme mostrado na Tabela 23.

Para avaliar os resultados da migração é conveniente ter presentes as seguintes métricas:

- **Precisão** (*Precision*): é a fração dos documentos recuperados que é relevante;
- **Recuperação** (*Recall*): é a fração dos documentos relevantes que foi recuperada;
- **F1-Score**: traduz-se na média harmónica entre a precisão e a recuperação.

Tabela 23 - Informação sobre o dataset obtido após a migração

Dataset Info				
Res	All	RefCode	Date	Person
Total	350	35	35	280
TotalNotNull	258	35	35	188

Perante isto, é possível observar (Tabela 24) que tanto os valores da precisão como da recuperação são bastante altos. O valor da precisão, bastante aproximado de 1, revela que praticamente toda a informação recuperada é, efetivamente, relevante, ou seja, corresponde ao anotado na amostra relativa aos registos do DigitArq. O valor da recuperação mostra que a maior parte das expressões presentes nos campos de texto foi devidamente extraída, dado que este valor é acima de 0.82. Consequentemente, o valor do F1-Score é, também, bastante alto.

Tabela 24 - Resultados com todas as expressões

Measures Full Correctness				
Measures	All	RefCode	Date	Person
Precision	0,977169	1	1	0,966443
Recall	0,829457	1	1	0,765957
F1-Score	0,897275	1	1	0,854599

Já na Tabela 25, é possível ver que, considerando apenas as expressões não nulas, os valores da precisão e da recuperação sobem. Neste caso, o valor da precisão é 1, significando que toda a informação recuperada é relevante. O valor da recuperação, sendo acima dos 0.84, mostra que grande parte das expressões presentes nos campos de texto foi extraída, conforme se pretendia. O valor do F1-Score mostra que a média harmónica entre aqueles dois valores é ainda mais elevada do que no caso anterior.

Tabela 25 - Resultados com as expressões não nulas

Measures Parcial Correctness				
Measures	All	RefCode	Date	Person
Precision	1	1	1	1
Recall	0,848837	1	1	0,792553
F1-Score	0,918239	1	1	0,884273

6. Avaliação da qualidade de acesso

Esta capítulo apresenta uma avaliação da qualidade de acesso, como primeiro contributo para a experiência 5 referida na secção 1.1. Esta avaliação está naturalmente condicionada pelo estado ainda intermédio da funcionalidade de pesquisa do protótipo que usa a ArchOnto.

O DigitArq é a plataforma de pesquisa do Arquivo Nacional da Torre do Tombo que tem o objetivo de permitir aos utilizadores o acesso remoto a determinados serviços que, anteriormente, só conseguiriam de forma presencial, como, por exemplo, visualizar documentos digitalizados ou reservar documentos para leitura presencial.

Desta forma, quando se acede a esta plataforma[29], a primeira entrada com que o utilizador se depara é com o modo mais básico da pesquisa (Figura 6), onde lhe é permitido colocar o termo ou a expressão que deseja pesquisar, bem como o intervalo temporal, caso este seja pretendido.



Figura 6 - Interface da Pesquisa Simples

Na barra do topo é possível observar que também existe o modo de pesquisa avançada (Figura 7). Esta pesquisa, por sua vez, permite ao utilizador efetuar uma pesquisa mais refinada, através da pesquisa pelo código de referência, pelo título ou pela pesquisa com intervalos temporais como a data de criação do registo, a data da última modificação do registo ou outro intervalo que o utilizador pretenda. Também é possível efetuar pesquisas através da cota atual,

[29] https://digitarq.arquivos.pt/

dos níveis de descrição (fundo, secção, unidade de instalação, entre outros), bem como pesquisar noutros campos (âmbito e conteúdo, autor intelectual ou unidades de descrição relacionadas) e efetuar uma pesquisa por assuntos (cidades, cargos). Nos campos relativos ao título e à cota atual, o utilizador pode selecionar se o termo ou a expressão inserida contém todas as palavras, se contém qualquer uma das palavras e se contém ou não a frase exata.

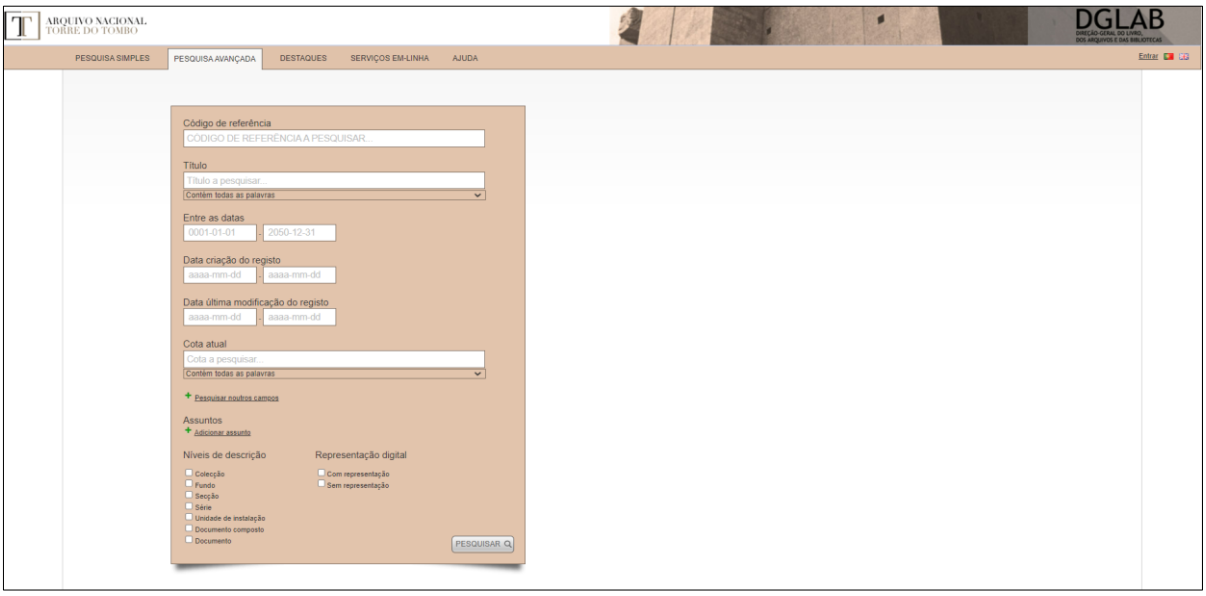


Figura 7 - Interface da Pesquisa Avançada

Por outro lado, o protótipo do projeto EPISA DigitArq+ (Figura 8) até ao momento só possui uma barra de pesquisa onde o utilizador pode inserir o termo ou expressão a pesquisar. Após efetuar a pesquisa, a interface mostra os resultados que foram devolvidos, bem como as entidades que estão associadas a estes registos de arquivo.

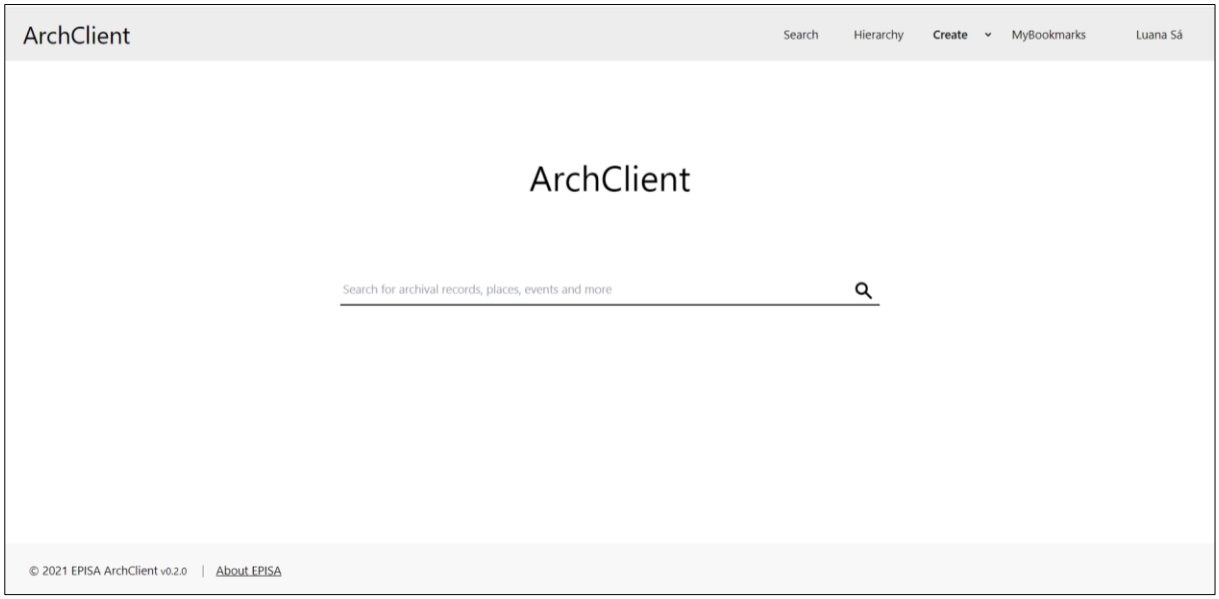


Figura 8 - Interface DigitArq+

Com o intuito de avaliar a qualidade do acesso à nova interface – DigitArq+ –, optou-se por elaborar um guião de atividades de pesquisa (Apêndice 1 – Guião completo) a executar por um conjunto de utilizadores no DigitArq e no DigitArq+, procedendo-se à observação dos correspondentes comportamentos e à sua análise comparativa.

Este guião traduz-se num conjunto de seis perguntas que foram construídas do ponto de vista de quatro categorias. A **primeira categoria** incidia sobre encontrar documentos e, para isso, desenvolveram-se as perguntas 1 e 2. A **segunda categoria** era relativa à quantidade de documentos que resultam de determinadas pesquisas (perguntas 3 e 4). A **terceira categoria** tinha o intuito de testar o efeito das datas em pesquisas, refletindo-se na pergunta 5. Finalmente, através da pergunta 6, a **quarta categoria** tinha o intuito de explorar a estrutura hierárquica de cada coleção existente.

Por sua vez, as perguntas foram elaboradas tendo como base os conteúdos correspondentes à migração previamente efetuada para que fosse possível obter resultados em ambas as plataformas.

Para o registo destas respostas, foi preparado um documento Excel que possuía três folhas: a primeira com a explicação das etiquetas utilizadas nas folhas seguintes (Apêndice 2 – Campos utilizados no registo das respostas); a segunda com o registo das respostas dos utilizadores comuns e a terceira com o registo das respostas dos arquivistas.

Desta forma, tanto a folha dos utilizadores comuns como a folha dos arquivistas possuíam uma

célula com o número do participante (participante 1, participante 2, etc.), seguindo-se um conjunto de quadros do lado esquerdo, que correspondiam às respostas obtidas no DigitArq, e do lado direito os quadros que correspondiam às respostas obtidas no DigitArq+. Por sua vez, estes quadros estavam divididos por perguntas, existindo células para registar a forma da resposta e a resposta final que o utilizador dava, seguindo-se eventuais erros cometidos e as notas. Estas linhas eram iguais para as pesquisas no DigitArq e no DigitArq+ (Apêndice 3 – Organização do registo das respostas).

Por baixo, existe outro quadro para registar a informação relativa ao tipo de contacto com o DigitArq (se se tratava de um utilizador que já teve algum contacto com a plataforma, se nunca usou ou se usou poucas vezes); um possível comentário ou sugestão que o utilizador pudesse mencionar relativamente à nova interface; se o utilizador consideraria útil ter a opção de selecionar determinados filtros que, posteriormente, estivessem sempre ativos durante várias pesquisas e, por fim, idade e formação académica (Apêndice 3 – Organização do registo das respostas).

6.1. Perguntas colocadas no guião

Nesta secção apresentam-se os resultados esperados em cada tarefa do guião. Para evitar dispersão, as figuras com os ecrãs de pesquisa mencionados encontram-se no Apêndice 4 – Resultados das pesquisas efetuadas.

6.1.1. Pergunta 1

Para a primeira pergunta do guião – Que documento é devolvido quando se pesquisa a Cota Atual "Empresa Pública do Jornal O Século, Joshua Benoliel, Cota Desconhecida 011"? – era necessário efetuar uma pesquisa que devolvesse o documento "Venda de louça preta: as Sras. D. Maria Luísa de Brito Rocha e Melo, D. Helena de Brito Rocha e Melo, D. Ana da Cunha e Meneses Pinto Cardoso (Lumiares), D. Maria Luísa Barroso da Câmara, o Sr. Francisco Queirós de Andrade Pinto, e as Sras. D. Francisca Seabra de Lacerda, D. Maria de Brito Rocha e Melo e D. Maria Emília Cabral da Silva".

Esta questão tinha o objetivo de encontrar o documento através da informação presente na sua Cota atual. Para isso, era necessário efetuar uma pesquisa avançada (Figura 11) dado que neste tipo de pesquisa existe um campo específico para pesquisar pela Cota atual.

Também seria possível encontrar o documento em questão através da pesquisa simples (Figura

12). Contudo, ao efetuar a procura pela pesquisa avançada, o DigitArq devolve diretamente o documento pedido, enquanto pela pesquisa simples o DigitArq iria devolver dois documentos, sendo que só um é que corresponde à Cota atual pedida.

No que concerne ao resultado esperado de resposta a esta questão, é de referir que, na pesquisa simples, deveria aparecer somente um resultado, dado que a Cota Atual é única para cada documento. Por outro lado, na pesquisa avançada, o DigitArq devolve o documento pretendido, abrindo-o diretamente após a sua pesquisa.

Já no DigitArq+, não foi possível efetuar esta pesquisa pela Cota Atual. Desta forma, foi pedido aos utilizadores que efetuassem uma pesquisa que devolvesse o mesmo documento, mas não podendo utilizar o título completo do mesmo. Ou seja, só seria possível aceder ao documento através de parte do seu título ou do código de referência. Esta interface devolveu somente um resultado, conforme é possível observar na Figura 13.

6.1.2. Pergunta 2

Esta questão – *Qual o código de referência do documento composto sobre o processo de Jerónima da Cruz?* – tinha o objetivo de testar a capacidade de encontrar documentos, através de informação presente nos respetivos campos de descrição.

Esta pesquisa devolvia o documento com o título "Processo de Jerónima da Cruz" e tanto se podia efetuar uma pesquisa simples como uma pesquisa avançada. No caso da pesquisa simples, poder-se-ia pesquisar por "Jerónima da Cruz" ou por "processo de Jerónima da Cruz". A primeira pesquisa devolvia vinte e quatro resultados (Figura 14), sendo que o documento pretendido é dos últimos a surgir, enquanto na segunda pesquisa devolve somente cinco resultados, onde o documento pretendido é o terceiro a aparecer (Figura 15). No caso da pesquisa avançada, ao pesquisar por "Jerónima da Cruz" o DigitArq devolve três documentos compostos, sendo um deles o documento pretendido (Figura 16). Quando se pesquisa por "processo de Jerónima da Cruz" devolve apenas dois documentos compostos, sendo um deles o pretendido (Figura 17).

Apesar destes resultados, esperava-se que, no DigitArq, quando se efetuasse a pesquisa "processo de Jerónima da Cruz", devolvesse somente resultados de processos com nomes semelhantes ao pesquisado como, por exemplo, os documentos compostos "Processo de Jerónima da Cruz ou Soares" e "Processo de Joana da Cruz". Este resultado deve-se ao facto de o DigitArq efetuar a pesquisa simples em todos os campos de descrição dos documentos como, por exemplo, o código de referência, o âmbito e conteúdo ou a cota atual. Assim, devolve todos

os documentos que contenham a expressão de pesquisa ou uma parte da mesma nos campos de descrição existentes.

No que concerne ao **DigitArq+**, quando se efetua uma pesquisa por "Jerónima da Cruz", esta interface devolve somente um resultado que é o documento composto (Figura 18) e, também, mostra uma entidade que está associada a este documento que, neste caso, é o nome da pessoa em causa (Figura 19).

Porém, também é possível chegar a este documento composto através da expressão "Processo de Jerónima da Cruz". Neste caso, a interface só devolve o documento composto e não associa qualquer entidade (Figura 20).

6.1.3. Pergunta 3

A terceira pergunta – *Qual é o número de resultados devolvidos quando se efetua uma pesquisa sobre o Mosteiro de São Bernardo de Portalegre?* – tinha o intuito de averiguar a quantidade de documentos que podem resultar de determinadas pesquisas.

Desta forma, ao efetuar a pesquisa simples sobre o Mosteiro de São Bernardo de Portalegre no Arquivo Distrital de Portalegre, este devolveu cento e seis resultados (Figura 21). Defende-se a posição de que nem todos eram relevantes, dado que alguns dos resultados devolvidos não pertencem ao fundo pesquisado. É de referir que o ideal seria devolver o fundo relativo ao Mosteiro de São Bernardo de Portalegre como primeiro resultado, seguindo-se todos os documentos que lhe pertencessem e, desta forma, estes poderiam conduzir o utilizador até ao fundo. Quando se faz uma pesquisa avançada, através do campo do Título, o Arquivo Distrital devolve somente catorze resultados, sendo que o primeiro é o próprio fundo e os restantes são documentos simples que fazem parte do mesmo (Figura 22).

Relativamente ao DigitArq+, esta interface só devolve um resultado, não possuindo qualquer entidade associada (Figura 23). Porém, se se efetuar a pesquisa por "São Bernardo", além de devolver como resultado este mesmo fundo, o DigitArq+ mostra uma entidade - Convento de São Bernardo, da Ordem de Cister (Figura 24).

6.1.4. Pergunta 4

A quarta pergunta – Efetuar uma pesquisa que devolva as secções existentes no fundo "Juízo da Índia e Mina". Registar o número de secções encontradas. – tinha o intuito de verificar a quantidade de documentos que podem resultar de determinadas pesquisas, à semelhança da

pergunta 3, mas incluindo explicitamente o nível de descrição pretendido.

Para isso, era necessário que o utilizador efetuasse uma pesquisa que mostrasse as secções que o fundo "Juízo da Índia e Mina" possui. Num primeiro momento, ao efetuar uma pesquisa simples com a expressão "Juízo da Índia e Mina" e, seguidamente, selecionar a opção "Secção" do lado direito (Figura 25), o DigitArq irá devolver cinco resultados, mas apenas três pertencem a este fundo. Se se efetuar o mesmo percurso na pesquisa avançada (Figura 26), também com a expressão "Juízo da Índia e Mina", o DigitArq não irá devolver qualquer documento, dado que é feita uma pesquisa pelo título e não existe nenhuma secção com esse título.

O esperado era que a resposta obtida fossem as vinte e cinco secções presentes no fundo. Contudo, os resultados obtidos justificam-se pelo facto de a própria secção não ter diretamente a expressão "Juízo da Índia e Mina" e, por isso, não aparece nos resultados devolvidos. Ou seja, o DigitArq não está a usar a hierarquia presente em todos os documentos para efetuar a pesquisa. Caso a pesquisa através da hierarquia fosse utilizada, o DigitArq deveria devolver todas as secções. Quando o utilizador quer fazer uma pesquisa hierárquica, o DigitArq deveria considerar tudo o que está sob um determinado resultado. Assim, seria possível procurar o fundo e a informação relativa a este deveria estar presente nos diversos níveis de descrição existentes. Deste modo, o utilizador já poderia ser capaz de aplicar qualquer filtro que iria obter sempre a resposta mais completa possível.

Por fim, o DigitArq+ devolve apenas um documento que é o próprio fundo, não tendo qualquer entidade associada (Figura 27).

6.1.5. Pergunta 5

A pergunta 5 – Ao efetuar uma pesquisa sobre "registos de baptismos", registe o número de resultados devolvidos. Seguidamente, fazer a mesma pesquisa considerando o intervalo temporal entre 1811 e 1822, registando, igualmente, o número de resultados. – tinha o objetivo de testar o efeito das datas nas pesquisas.

Ao efetuar a pesquisa simples no Arquivo Distrital do Porto, são devolvidos 13077 resultados, entre fundos, séries, unidades de instalação e documentos simples (Figura 27). Porém, quando a pesquisa é restringida ao intervalo temporal pedido, o número de resultados devolvidos desce abruptamente para 1214 resultados (Figura 29). O mesmo acontece na pesquisa avançada onde, numa primeira fase, são devolvidos 12599 documentos (Figura 31) e, após a introdução do intervalo temporal, o número de resultados decresce para 817 (Figura 32).

Os resultados previamente mencionados são os obtidos quando nos campos específicos para as datas se colocam somente os anos (neste caso, 1811 e 1822). Mas, ao colocar as datas completas (neste caso, *1811-01-01* e *1822-12-31*), existe a diferença de um resultado. Ou seja, quando se coloca apenas o ano, o Arquivo Distrital devolve 1214 resultados, enquanto quando se colocam as datas completas, o Arquivo Distrital devolve 1213 resultados (Figura 30). O mesmo acontece na pesquisa avançada: com a utilização dos anos o Arquivo Distrital devolve 817 resultados e com a utilização das datas completas, devolve 816 resultados (Figura 33). Esta diferença dos resultados ocorre devido a uma unidade de instalação[30] que na sua data de produção tem apenas "1811", o que faz com que não seja abrangida quando se coloca um intervalo temporal específico.

Por fim, relativamente ao resultado no Digitarq+, é de referir que esta interface devolve vinte e oito resultados e não possui nenhuma entidade, não sendo possível fazer restrições com intervalos temporais (Figura 34).

6.1.6. Pergunta 6

A pergunta 6 – Efetue uma pesquisa para encontrar o nome do fundo, bem como o nome dos diferentes níveis de descrição, onde está presente o documento simples "Um aquário de 2,40x1,00 m" com o código de referência PT/TT/AS/A/001/000142. – tinha o objetivo de evidenciar a existência de uma estrutura hierárquica no DigitArq. Esta, por sua vez, mostra de que forma os documentos são organizados.

Quando se procura este documento através da pesquisa simples do DigitArq, são devolvidos três resultados com o mesmo nome, por isso é que a pergunta se faz acompanhar do código de referência para que seja possível identificar o documento em causa (Figura 35). Por outro lado, quando se procura este documento através da pesquisa avançada do DigitArq, são devolvidos dois resultados (Figura 36) caso a pesquisa seja feita somente através do título. Quando a pesquisa é feita através do código de referência, o DigitArq abre diretamente o documento (Figura 37).

Já o DigitArq+ devolve apenas um resultado, não possuindo entidades associadas. Neste caso, o documento é devolvido quer a pesquisa seja feita através do título quer seja através do código de referência (Figura 38).

[30] https://pesquisa.adporto.arquivos.pt/details?id=827166

6.2. Análise dos resultados

Nesta secção é apresentada a análise dos resultados obtidos nas sessões de resposta ao guião. Os utilizadores foram segmentados em dois grupos, o dos utilizadores comuns e o dos arquivistas.

Começando com o grupo dos **utilizadores comuns**, conforme é possível observar na Figura 9, este grupo é composto por doze pessoas com a idade compreendida entre os 20 e os 40 anos. Relativamente à formação académica, é de referir que predominam os cursos relacionados com Ciências Humanas e com Ciências Exatas como, por exemplo, Ciência da Informação e de Engenharia Informática, respetivamente. Deste conjunto de doze participantes, metade referiu que foi a primeira vez que esteve em contacto com o DigitArq, seguindo-se o contacto através do projeto EPISA. Apenas dois participantes referiram que já tinham utilizado o DigitArq, embora poucas vezes.

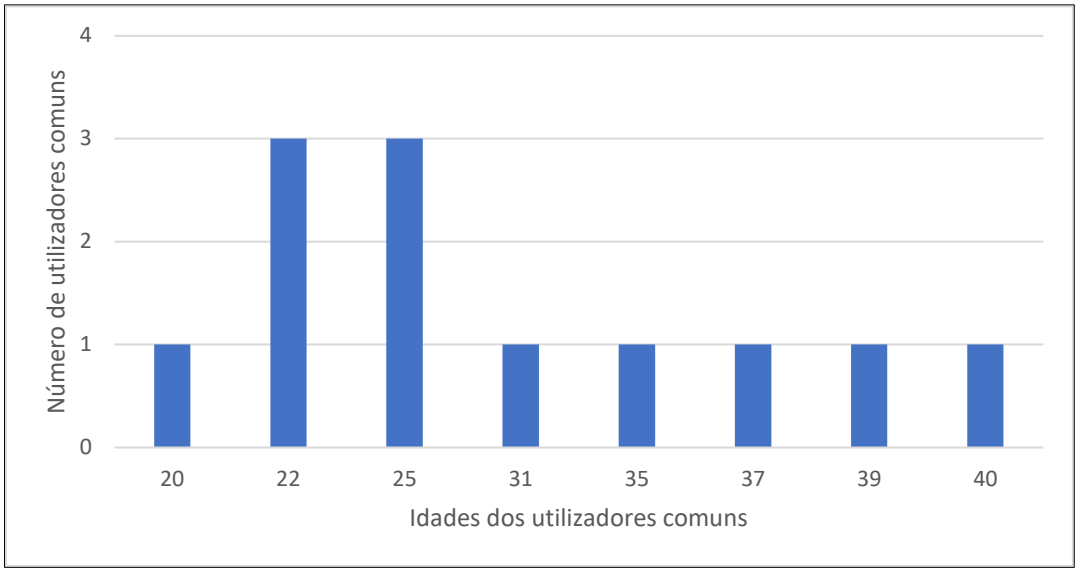


Figura 9 - Idades dos utilizadores comuns

Num primeiro momento, foi pedido a cada participante que efetuasse as pesquisas no DigitArq e, seguidamente, no DigitArq+.

Na **Pergunta 1**, sete utilizadores comuns usaram a pesquisa simples e cinco usaram a pesquisa avançada. Relativamente aos utilizadores que usaram a pesquisa simples é de referir que na maioria dos casos foi utilizada a cota conforme apresentada no enunciado e, mesmo com a devolução de dois resultados, os utilizadores não tiveram dificuldade em identificar o documento pretendido. É de salientar um participante que começou por pesquisar através da

expressão “Empresa Pública do Jornal O Século” e, percebendo que não chegaria ao documento, utilizou a cota conforme apresentada no enunciado, continuando na pesquisa simples.

Por outro lado, os participantes que utilizaram a pesquisa avançada, chegaram ao documento pretendido através do campo específico para a Cota atual, utilizando a cota conforme o enunciado e facilmente chegaram ao documento pretendido. Neste ponto, é de mencionar que um participante começou por pesquisar com a expressão “Empresa Pública do Jornal O Século” na pesquisa simples, tendo acabado por chegar ao resultado através da pesquisa avançada, utilizando a cota conforme o enunciado.

No que concerne às pesquisas no DigitArq+, não sendo possível efetuar esta pesquisa pela Cota Atual, foi pedido que os utilizadores efetuassem uma pesquisa tendo como base a informação presente no DigitArq e, caso utilizassem o título, este não poderia ser utilizado na sua totalidade. Desta forma, seis participantes chegaram ao documento através do título, apesar da existência de algumas variações na escrita (por exemplo: “venda de louça preta”, “VENDA DE LOUÇA PRETA”, “Venda de Louça Preta”) e os restantes utilizaram o código de referência.

Na **Pergunta 2**, seis participantes utilizaram a pesquisa simples e seis usaram a pesquisa avançada. Na pesquisa simples os participantes pesquisaram por “Jerónima da Cruz” ou por “processo de Jerónima da Cruz”. No primeiro caso, o DigitArq devolveu três páginas com resultados e somente uma pessoa selecionou a opção “Documento composto” presente nos filtros apresentados ao lado. Os restantes inquiridos procuraram em todas as páginas até encontrar o documento composto pretendido.

Já na pesquisa avançada, a expressão de pesquisa “processo de Jerónima da Cruz” só foi utilizada por três pessoas, sendo que as outras pesquisas foram feitas com a expressão “Jerónima da Cruz”. Um destes participantes iniciou a procura pela pesquisa simples, mas, face ao número de resultados apresentados, optou pela pesquisa avançada.

A seguir, quando pedido para efetuar a mesma pesquisa no DigitArq+, os participantes chegaram ao documento pretendido através de expressões como “processo de Jerónima da Cruz”, “Jerónima da Cruz”, “PROCESSO DE JERÓNIMA DA CRUZ” ou apenas “Jerónima”, não havendo qualquer dificuldade na procura do mesmo.

Na **Pergunta 3**, apenas dois inquiridos utilizaram a pesquisa avançada. Nesta pergunta, todos os participantes utilizaram a mesma expressão de pesquisa: “Mosteiro de São Bernardo de Portalegre”, não havendo sinais de dificuldade para dar resposta à mesma. Neste caso, o que variou foi o número de resultados, dado que a pesquisa simples devolveu 106 resultados e a pesquisa avançada devolveu 14 resultados.

No momento da pesquisa no DigitArq+, a expressão de pesquisa utilizada foi a mesma, tendo sido devolvido somente um resultado.

Na **Pergunta 4**, obtiveram-se oito respostas pela pesquisa simples e quatro respostas pela pesquisa avançada. Na pesquisa simples, seis participantes iniciaram a sua pesquisa por colocar na barra de pesquisa a expressão "Juízo da Índia e Mina" e, após o DigitArq devolver uma resposta, selecionaram o filtro "Secção" presente no lado direito da página. Os restantes dois participantes utilizaram a mesma expressão, mas optaram por procurar o fundo no conjunto de resultados, acedendo ao mesmo e contando as secções apresentadas.

Relativamente à pesquisa avançada, é de referir que três pessoas colocaram a expressão "Juízo da Índia e Mina" no Título, selecionando "fundo" na parte dos filtros. Dado que esta pesquisa abre diretamente o fundo, os participantes contaram as secções. Neste ponto, importa realçar uma pesquisa efetuada por um participante. Este começou por pesquisar "Juízo da Índia e Mina" no Título, selecionou "fundo" como filtro. A seguir, fez a mesma pesquisa sem o filtro do nível de descrição. Por fim, efetuou a pesquisa através do código de referência do fundo e, após o DigitArq devolver diversos resultados, selecionou "Secção" como filtro. Assim chegou à resposta de vinte e três secções.

No DigitArq+, a maioria dos participantes efetuou a pesquisa através da expressão "Juízo da Índia e Mina", tendo um resultado como resposta. Contudo, perante esta situação, um participante efetuou a pesquisa pelo código de referência, obtendo dez resultados, onde apenas um estava identificado como secção.

Na **Pergunta 5** todos os participantes utilizaram a pesquisa simples para responderem à pergunta colocada. Numa primeira fase, todas as pesquisas devolveram, logicamente, o mesmo número de resultados. A seguir, quando questionados sobre o número de resultados devolvidos pelo DigitArq quando se aplicava o intervalo temporal entre 1811 e 1822, é que existiu uma variação na forma de pesquisa e, consequentemente, no número de resultados devolvidos.

Quando questionados sobre o porquê de pesquisarem utilizando somente o ano, as respostas passavam por ser aquilo que abrangeria mais documentos, por ser o que estava presente no enunciado do guião ou por assumirem que a plataforma devolvia utilizando só o ano. Por outro lado, quando questionados sobre a utilização das datas completas, os participantes alegaram que era a informação apresentada por defeito na caixa de seleção.

No DigitArq+, quase todas as pesquisas foram efetuadas através da expressão "registos de baptismos". Porém, um participante efetuou a pesquisa por "registos de baptismo", que se traduziu na diferença de um resultado com o título "Índice de registos de baptismo".

Para responder à **Pergunta 6**, cinco participantes utilizaram a pesquisa simples, enquanto sete optaram pela pesquisa avançada. Na pesquisa simples, dois dos participantes procuraram o documento pelo título, um pesquisou pelo código de referência, enquanto outros dois participantes efetuaram a pesquisa pelo código de referência colocando, também, o título do documento. Na pesquisa avançada, seis participantes utilizaram o código de referência e um fez uso tanto do código de referência como do título do documento.

Relativamente ao entendimento da estrutura hierárquica presente na interface, a maioria dos participantes conseguiu identificar a presença da hierarquia, conseguindo, também, identificar os diferentes níveis de descrição. Contudo, outros participantes procuraram esta informação nos campos de descrição do documento em questão.

No DigitArq+, quatro participantes pesquisaram pelo título, enquanto os restantes oito pesquisaram pelo código de referência colocando. Nesta interface, dois participantes conseguiram identificar a hierarquia através do código de referência, tendo como base a informação do DigitArq, enquanto outros dois participantes conseguiram identificar que se tratava de um documento simples ao abrirem o documento.

No que concerne ao grupo dos **arquivistas**, é de referir que foram questionadas quatro pessoas, com idades compreendidas entre os 36 e os 55 anos (Figura 10).

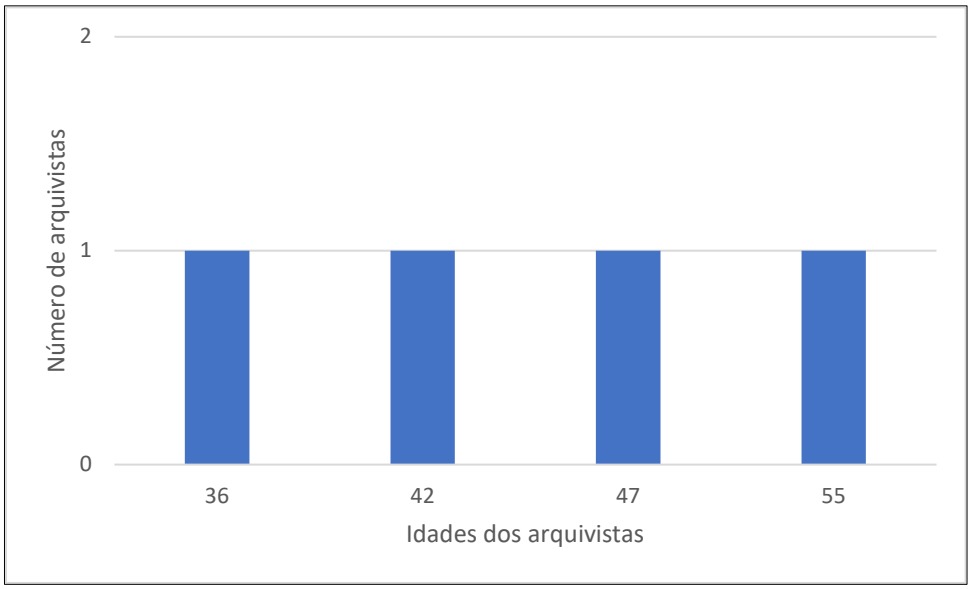


Figura 10 - Idades dos arquivistas

Relativamente à formação académica, é de referir que predomina o curso de Ciência da Informação, existindo uma pessoa com o curso de Ciências e Tecnologias da Documentação e Informação. É, ainda, de destacar que somente um arquivista só teve contacto uma vez com a plataforma, sendo que os restantes tiveram contacto mais que uma vez.

Num primeiro momento, foi pedido a cada arquivista que efetuasse as pesquisas no DigitArq e, seguidamente, no DigitArq+.

Na **Pergunta 1**, três dos arquivistas utilizaram a pesquisa avançada e somente um utilizou a pesquisa simples. Relativamente à pesquisa avançada, é de mencionar que todos os arquivistas utilizaram o campo específico para efetuar uma pesquisa pela Cota atual de qualquer documento. Contudo, começaram por pesquisar os termos "Empresa Pública do Jornal O Século" ou os termos "Empresa Pública do Jornal O Século, Joshua Benoliel". Como estas pesquisas não devolviam o resultado esperado, optaram por pesquisar pela Cota conforme a apresentada no enunciado do guião.

Na pesquisa simples ocorreu o mesmo com a pesquisa através dos termos "Empresa Pública do Jornal O Século", seguindo-se a pesquisa com a Cota apresentada no enunciado.

Já no DigitArq+, não sendo possível efetuar esta pesquisa pela Cota Atual, foi pedido aos arquivistas que efetuassem uma pesquisa tendo como base a informação presente no DigitArq e, caso utilizassem o título, este não poderia ser utilizado na sua totalidade. Assim, começaram por tentar com partes da Cota como, por exemplo, "Joshua Benoliel", "jornal o século" ou "empresa pública". A seguir, chegaram ao documento pretendido através de pesquisas com expressões como "louça preta", "maria luisa de brito rocha" que é um dos nomes que consta no título e "VENDA DE LOUÇA PRETA". Um dos arquivistas efetuou a sua pesquisa através do código de referência.

Na **Pergunta 2**, semelhante ao ocorrido na pergunta 1, três arquivistas utilizaram a pesquisa avançada e apenas um utilizou a pesquisa simples. Na pesquisa simples, a expressão utilizada foi "jerónima da cruz" e o arquivista inquirido procurou nas três páginas que o DigitArq devolveu, chegando, assim, ao documento pretendido.

Relativamente à pesquisa avançada, é de referir que os arquivistas utilizaram as expressões "jeronima da cruz" ou "Jerónima da Cruz". Somente um arquivista aplicou o filtro "documento composto". Neste caso, o DigitArq devolveu três resultados, sendo mais fácil chegar ao documento pretendido.

No DigitArq+, as expressões utilizadas foram "jeronima da cruz", "Jerónima da Cruz" e "jerónima da cruz", não havendo qualquer dificuldade na procura do mesmo dado que esta

interface só devolveu um resultado.

Na **Pergunta 3**, três arquivistas utilizaram a pesquisa simples e somente um utilizou a pesquisa avançada. Nesta pergunta, todos os arquivistas utilizaram a mesma expressão de pesquisa - "Mosteiro de São Bernardo de Portalegre". Neste caso, o que variou foi o número de resultados, dado que a pesquisa simples devolveu 106 resultados e a pesquisa avançada devolveu 14 resultados, não havendo sinais de dificuldade para dar resposta à questão.

No momento da pesquisa na nova interface, a expressão de pesquisa utilizada foi a mesma, tendo sido devolvido apenas um resultado.

Na **Pergunta 4**, todos os arquivistas chegaram a uma resposta pela pesquisa avançada. Porém, dois dos arquivistas inquiridos começaram por efetuar uma pesquisa simples com a expressão "Juízo da Índia e Mina". A seguir, passaram para a pesquisa avançada, utilizando a mesma expressão de pesquisa no campo do Título, selecionaram o resultado que devolvia este fundo e contaram as secções existentes.

Os restantes arquivistas inquiridos efetuaram uma pesquisa avançada com a expressão "Juízo da Índia e Mina" no campo do *Título*, selecionaram o resultado que devolvia este fundo e contaram as secções existentes.

Ao efetuarem a mesma pesquisa na nova interface, todos utilizaram a mesma expressão de pesquisa, tendo sido devolvido apenas um resultado.

Na **Pergunta 5**, três arquivistas utilizaram a pesquisa avançada e as expressões utilizadas passaram por "Registo de Baptismo", "registos de baptismos" e "Registos baptismo". Esta diferença traduziu-se, naturalmente, no número de resultados devolvidos.

A seguir, quando questionados sobre o número de resultados devolvidos pelo DigitArq quando se aplicava o intervalo temporal entre 1811 e 1822, é que existiu uma variação na forma de pesquisa e, consequentemente, no número de resultados devolvidos. Conforme ocorrido nas sessões com os utilizadores comuns, quando os arquivistas foram questionados sobre o porquê de pesquisarem utilizando somente o ano, as respostas passavam por ser o que estava presente no enunciado do guião ou por concluírem que a plataforma devolvia utilizando só o ano. Por outro lado, quando questionados sobre a utilização das datas completas, os participantes alegaram que era a informação apresentada por omissão na caixa de seleção e para que o DigitArq devolvesse resultados que abrangessem o ano inteiro.

Por outro lado, um arquivista começou por utilizar a pesquisa simples para a primeira parte da pergunta e utilizou a pesquisa avançada para aplicar os intervalos temporais, sempre com a expressão "registos de baptismos".

No momento das pesquisas no DigitArq+, a maioria dos arquivistas utilizou a expressão de pesquisa "registos de baptismos". Porém, um arquivista efetuou a pesquisa por "registos baptismo" e por "registos de batismo", que se traduziu na inexistência de resultados. Supõe-se que isso aconteça por falta de configuração da pesquisa.

Para responder à **Pergunta 6**, apenas um arquivista utilizou a pesquisa simples, enquanto três arquivistas optaram pela pesquisa avançada. Na pesquisa simples, o arquivista efetuou a pesquisa pelo título do documento. Na pesquisa avançada, um arquivista utilizou o título do documento e os restantes utilizaram o código de referência.

Relativamente ao entendimento da estrutura hierárquica presente no DigitArq, todos os arquivistas a identificaram, bem como os diferentes níveis de descrição e o respetivo nome.

No DigitArq+, dois arquivistas efetuaram a pesquisa pelo título e os outros dois pelo código de referência. Um dos inquiridos conseguiu identificar a hierarquia através do código de referência, baseando-se na informação presente no DigitArq, enquanto os restantes arquivistas conseguiram identificar que se tratava de um documento simples ao abrirem o documento.

Relativamente ao tempo utilizado para responder ao guião, é de referir que os utilizadores comuns demoraram entre 14 e 27 minutos; enquanto os arquivistas demoraram entre 20 e 37 minutos.

No final da resposta ao guião, foi pedido a todos os inquiridos que fizessem um comentário ou dessem uma sugestão relativamente à nova interface, sendo também questionados sobre a utilidade de possuir filtros que estivessem sempre ativos durante um processo de pesquisa.

No que concerne ao **comentário e/ou sugestão** relativo à nova interface, os inquiridos mencionaram aspetos como a falta de filtros como intervalos temporais; a falta de uma pesquisa através dos níveis de descrição; a falta da estrutura hierárquica para se conseguir entender em que nível de descrição o documento se encontra; a falta de uma caixa para a pesquisa avançada, conforme está presente no DigitArq para que seja possível refinar as pesquisas seja pelas datas ou pelo nível de descrição. Esta pesquisa avançada deveria ter outras possibilidades de pesquisa como, por exemplo, a pesquisa hierárquica, onde ocorreria uma associação de objetos digitais, através das propriedades definidas nos respetivos níveis superiores; a pesquisa através do código de referência deveria abrir diretamente o documento, como ocorre no DigitArq.

Alguns arquivistas referiram que esperavam que a nova interface tivesse ajuda para que os utilizadores não tivessem dúvidas sobre como efetuar as pesquisas ou onde as efetuar. Realçaram, de igual forma, a distinção existente entre registos arquivísticos e entidades

quando o DigitArq+ devolvia os resultados. Por fim, todos os participantes referiram que a nova interface é mais simples e intuitiva e que é mais rápida a devolver resultados, comparativamente ao DigitArq.

Relativamente à existência de uma opção onde diversos **filtros** de pesquisa estivessem sempre ativos, todos os participantes afirmaram que gostariam de ter essa opção. Desta forma, foram exemplificados casos onde os filtros poderiam ser intervalos temporais, ter a opção de efetuar uma pesquisa hierárquica ou uma pesquisa por todos os campos ou ter uma opção para selecionar o nível de descrição pretendido também na caixa de pesquisa simples, além do que já existe na caixa de pesquisa avançada.

7. Conclusões

Em jeito de conclusão, é de referir que a Torre do Tombo tem utilizado as normas ISAD (G), ISAAR (CPF), ISDIAH e ISDF para representar os objetos culturais que custodia. A primeira norma tem o objetivo de explicar o conteúdo dos documentos de um arquivo para facilitar a sua acessibilidade. A segunda norma tem como finalidade partilhar as descrições dos produtores de cada item e promover descrições consistentes das pessoas coletivas, das pessoas singulares e das famílias que os produziram. A terceira norma surgiu com o intuito de mostrar que tipo de informação pode ser incluída nas descrições feitas acerca das próprias instituições de arquivo. A última norma foi elaborada para ter diretivas que dirigissem a preparação de descrições de funções de entidades coletivas, esclarecendo que informação deve ser incluída nas descrições.

No que diz respeito à revisão de literatura, é possível afirmar que a Web Semântica, sendo parte da *World Wide Web* e fazendo com que os seres humanos e computadores trabalhem em cooperação, necessita de desenvolver mais linguagens recomendadas. Os dados ligados são suportados por tecnologias como RDF, OWL e SPARQL. O RDF foi criado para representar metadados relativos a recursos da *Web* como o título ou o autor, com o objetivo de facilitar situações onde a informação seja processada por aplicações específicas. Já a OWL tem na sua base a lógica computacional e, assim, o conhecimento pode ser percorrido por programas de computador para tornar o conhecimento implícito em conhecimento explícito. Por fim, a SPARQL pode ser utilizada para expressar *queries* através de múltiplas fontes de dados, sendo um conjunto de especificações que usam protocolos para manipulação de conteúdo RDF. Por outro lado, não é possível falar em Web Semântica sem falar em Dados Ligados, os quais integram em grande escala os dados na *Web,* e para que seja possível criar estes Dados Ligados, é preciso que existam tecnologias que utilizem um formato comum para aceder a bases de dados em tempo real.

Relativamente aos modelos de dados ligados para o Património Cultural, o CIDOC CRM é uma ontologia de referência para melhorar a integração e a troca de informação no âmbito do Património Cultural que foi desenvolvida para promover a compreensão da informação sobre este. Uma das suas inovações é a estruturação da informação à volta de eventos temporais. Apesar de ter sido desenvolvida para museus, já foi aplicada em outras áreas como, por exemplo, escavação arqueológica (CRMarcheo); observação científica (CRMsci); documentação de edifícios arqueológicos (CRMba); propriedades espácio-temporais nas entidades temporais (CRMgeo).

Ainda neste contexto surgem alguns projetos de interesse que foram selecionados por se

enquadrarem no domínio dos dados ligados no Património Cultural ou por possuírem o CIDOC CRM como base das ontologias.

Após estes capítulos relativos à revisão de literatura, surge a secção sobre o processo de migração, caracterizado através da recolha dos metadados de diversos registos presentes no Arquivo Distrital da Guarda, no Arquivo Distrital do Porto e no Arquivo Distrital de Bragança. Para este efeito, todos os metadados foram estruturados em listas que continham as etiquetas respetivas.

Esta dissertação visou, de igual forma, avaliar a qualidade do acesso através de um guião com algumas tarefas para que cada grupo de utilizadores realizasse tanto no DigitArq como no DigitArq+. Desta forma, ao observar o comportamento de pesquisa destes grupos, foi possível avaliar a forma como as pesquisas são efetuadas. Neste ponto, é possível concluir que a pesquisa mais utilizada pelo grupo de utilizadores comuns foi a pesquisa simples, contrariamente ao ocorrido nas sessões de resposta com o grupo de arquivistas; o grupo de arquivistas possui, naturalmente, outra perceção da forma de pesquisa para responder a cada questão colocada e, desta forma, o grupo sabia em que local poderia refinar a pesquisa, tendo, igualmente, uma maior predisposição para interagir com o DigitArq+.

De uma forma geral, é possível concluir que os dois grupos notaram que a pesquisa avançada era mais rápida na devolução de resultados do que a pesquisa simples, dado que a pesquisa avançada é mais refinada e, consequentemente, mais direcionada. Porém, a maioria dos utilizadores comuns respondeu à maioria das questões através da pesquisa simples e só na Pergunta 2 é que o número de participantes igualou na utilização das pesquisas simples e avançada. É de destacar a Pergunta 6 que é uma exceção, na medida em que a maioria dos participantes optou pela pesquisa avançada para que fosse possível utilizar o código de referência do documento. Por outro lado, a maioria dos arquivistas teve tendência a responder pela pesquisa avançada. Neste caso, é de ressalvar a Pergunta 3 onde três dos quatro arquivistas inquiridos utilizaram a pesquisa simples para responder à questão em causa.

7.1. Trabalho futuro

Como trabalho futuro, devem considerar-se as sugestões dadas por todos os participantes das sessões de resposta ao guião realizadas na última fase desta dissertação.

Desta forma, os passos seguintes seriam adicionar uma caixa para que fosse possível efetuar pesquisas avançadas para, desta forma, os utilizadores poderem refinar as suas pesquisas; adicionar zonas específicas para que seja possível filtrar as pesquisas com intervalos temporais. Estes intervalos temporais devem estar disponíveis tanto no momento antes do DigitArq devolver os resultados como depois, sendo que estes filtros já estariam pré-definidos conforme já ocorre no DigitArq; acrescentar caixas que permitam ao utilizador selecionar em que níveis de descrição quer efetuar a pesquisa, bem como a estrutura hierárquica para se conseguir entender em que nível de descrição o documento se encontra e caixas onde seja possível selecionar vários filtros de pesquisa e tê-los sempre ativos ao longo de diversas pesquisas.

Por outro lado, é necessário acrescentar uma secção com ajuda direcionada a todos os utilizadores, facilitando-lhes, assim, caso surjam dúvidas sobre como efetuar as pesquisas ou onde as efetuar. Além disto, seria, também, importante que esta ajuda contivesse toda a informação relativa à organização e estruturação da documentação, bem como o significado dos diferentes símbolos que constituem a estrutura hierárquica.

Sendo que uma das novidades do DigitArq+ é a existência de entidades, seria, também, importante que este fosse capaz de efetuar a ligação entre indivíduos, através destas. Por exemplo, quando um utilizador efetuasse uma pesquisa sobre uma determinada pessoa, o DigitArq+ poderia ser capaz de providenciar uma resposta que contivesse pessoas relacionadas com esta, seja por serem família ou por estarem historicamente ligadas.

Com o guião também se concluiu que faria sentido o DigitArq+ possuir uma opção para efetuar pesquisas em modo hierárquico. Ou seja, através da Pergunta 4 percebeu-se que o DigitArq só devolve todas as secções do fundo Juízo da Índia e Mina caso a pesquisa seja feita através do código de referência. Quando se efetua uma pesquisa simples e, após devolução dos resultados, se seleciona "Secção" nas opções dos níveis de descrição, o DigitArq só devolve três das vinte e cinco secções que este fundo tem. Isto acontece dado que as secções não possuem a expressão "Juízo da Índia e Mina", conforme explicado na Pergunta 4 (6.1.4). Com uma pesquisa hierárquica, o DigitArq+ poderia devolver todas as secções presentes neste fundo, através da herança dos níveis superiores que refletiria nos níveis inferiores.

Desta forma, o DigitArq+ manteria os dois modos de pesquisa: a pesquisa simples e pesquisa avançada. A pesquisa simples seria, então uma pesquisa em todos os campos, conforme já acontece no DigitArq. Já a pesquisa avançada continuaria com o formulário já existente onde

é possível colocar condições como, por exemplo, pesquisa pelo código de referência, pela cota atual, por assuntos ou pelos níveis de descrição e acrescentar-se-ia a pesquisa no modo hierárquico onde se iria buscar as propriedades definidas nos níveis superiores.

8. Referências bibliográficas

ArCo. 2019a. “ArCo - Guide.” 2019. http://wit.istc.cnr.it/arco/?lang=en#page-top.

———. 2019b. “ArCo - Primer Guide v1.0.” 2019. http://wit.istc.cnr.it/arco/primer-guide-v1.0-en.html.

ARIADNE. 2016. “Advanced Research Infrastructure for Archaeological Dataset Networking in Europe.” 2016. http://www.ariadne-infrastructure.eu/.

———. 2019. “ARIADNEplus Deliverables – ARIADNE PLUS.” 2019. https://ariadne-infrastructure.eu/about-ariadne/.

Arquivo Distrital da Guarda. 2011a. “Paróquia Da Sé - Guarda.” 2011. https://digitarq.adgrd.arquivos.pt/details?id=1199345.

———. 2011b. “Registo de Batismos de Sé Da Guarda.” 2011. https://digitarq.adgrd.arquivos.pt/details?id=1199346.

———. 2011c. “Registo de Casamentos de Sé Da Guarda.” 2011. https://digitarq.adgrd.arquivos.pt/details?id=1199379.

———. 2011d. “Registo de Óbitos de Sé Da Guarda.” 2011. https://digitarq.adgrd.arquivos.pt/details?id=1199432.

Arquivo Distrital de Bragança. 2006. “Registo de Passaportes Deferidos.” 2006. https://digitarq.adbgc.arquivos.pt/details?id=1228230.

———. 2009. “Governo Civil de Bragança.” 2009. https://digitarq.adbgc.arquivos.pt/details?id=1226012.

Arquivo Distrital do Porto. 2012a. “Paróquia de Aldoar.” 2012. https://pesquisa.adporto.arquivos.pt/details?id=488455.

———. 2012b. “Registos de Baptismos.” 2012. https://pesquisa.adporto.arquivos.pt/details?id=488458.

———. 2012c. “Registos de Casamentos.” 2012. https://pesquisa.adporto.arquivos.pt/details?id=488477.

———. 2012d. “Registos de Óbitos.” 2012. https://pesquisa.adporto.arquivos.pt/details?id=488490.

Berners-Lee, Tim, James Hendler, and Ora Lassila. 2001. “The Semantic Web.” In *Scientific American ,* 284:34–43. https://doi.org/10.1007/978-1-4020-4749-5_3.

CIDOC CRM. 2009a. “Home | CIDOC CRM.” 2009. http://www.cidoc-crm.org/.

———. 2009b. “Home | CRMarchaeo.” 2009. https://cidoc-crm.org/crmarchaeo/.

———. 2009c. “Home | CRMba.” 2009. https://cidoc-crm.org/crmba/.

———. 2009d. “Home | CRMdig.” 2009. https://cidoc-crm.org/crmdig/.

———. 2009e. “Home | CRMgeo.” 2009. https://cidoc-crm.org/crmgeo/.

———. 2009f. “Home | CRMinf.” 2009. https://www.cidoc-crm.org/crminf/home-4.

———. 2009g. “Home | CRMsci.” 2009. https://cidoc-crm.org/crmsci/.

———. 2009h. “Home | CRMtex.” 2009. https://cidoc-crm.org/crmtex/.

Conselho Internacional de Arquivos. 2007. “ISDF: Norma Internacional Para Descrição de Funções.” www.ica.org.

———. 2008. “ISDIAH: Norma Internacional Para Descrição de Instituições Com Acervo Arquivístico.” *Aids*, 10–11. www.ica.org.

DBpedia. 2020. “About: UMBEL.” 2020. https://dbpedia.org/page/UMBEL.

DBpedia Association. 2002. “About: Http://Dbpedia.Org/Ontology/.” 2002. https://dbpedia.org/ontology/.

———. 2020. “About DBpedia.” 2020. https://www.dbpedia.org/about/.

DGLAB. 2011. “Nova Versão DIGITARQ.”

DIESEL. 2014. “The Pitch.” 2014. https://diesel-project.eu/.

———. 2017. “Archive Linked Open Data (ALOD).” 2017. https://diesel-project.eu/archive-linked-open-data-alod/.

Doerr, Martin. 2005. “The CIDOC CRM, an Ontological Approach to Schema Heterogeneity.” *Drops.Dagstuhl.De*. http://drops.dagstuhl.de/opus/volltexte/2005/35.

GATE. 2018. “GATE.Ac.Uk - Index.Html.” 2018. https://gate.ac.uk/.

Grupo de Trabalho para a Normalização da Descrição em Arquivo. 2002. “ISAD(G): Norma Geral Internacional de Descrição Arquivística.” *Ministério Da Cultura, Instituto Dos Arquivos Nacionais/Torre Do Tombo*, 97. http://arquivos.dglab.gov.pt/servicos/documentos-tecnicos-e-normativos/lista-de-documentos/%5Cnhttp://act.fct.pt/acervodocumental/documentos-tecnicos-e-normativos/.

———. 2004. “ISAAR (CPF): Norma Internacional de Registo de Autoridade Arquivística

Para Pessoas Colectivas, Pessoas Singulares e Famílias." www.ica.org.

Hitzler, Pascal, Bijan Parsia, Peter F Patel-schneider, Nuance Communications, and Sebastian Rudolph. 2012. "OWL 2 Web Ontology Language Primer (Second Edition)." World Wide Web Internet And Web Information Systems. 2012. https://www.w3.org/TR/owl2-primer/.

INESC TEC. 2021. "Inferência de Entidades e Propriedades Para Arquivos Semânticos - INESC TEC." 2021. https://www.inesctec.pt/pt/projetos/episa#intro.

INESC TEC, and DGLAB. 2019. "EPISA—Entity and Property Inference for Semantic Archives."

Koch, Inês, Cristina Ribeiro, and Carla Teixeira Lopes. 2020. "ArchOnto, a CIDOC-CRM-Based Linked Data Model for the Portuguese Archives." In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12246 LNCS:133–46. https://doi.org/10.1007/978-3-030-54956-5_10.

LinkedArt. 2020a. "Home." 2020. https://linked.art/index.html.

———. 2020b. "Linked Art - ICOM CIDOC -ICOM CIDOC." 2020. https://cidoc.mini.icom.museum/working-groups/linked-art/.

Manola, Frank, and Eric Miller. 2004. "RDF Primer." https://www.w3.org/TR/rdf-primer/.

Mazurek, Cezary, Krzysztof Sielski, Justyna Walkowska, and Marcin Werla. 2011. "Applicability of CIDOC CRM in Digital Libraries." In *Proceedings of „CIDOC 2011 – Knowledge Management and Museums" Conference, Sibiu, Romania, 4 – 9 September 2011*. http://purl.org/dc/terms/.

OWL Working Group. 2012. "OWL - Semantic Web Standards." W3C.Org. 2012. https://www.w3.org/OWL/.

Santos, Hercules Pimenta. 2016. "Modelo CIDOC CRM." *Brazilian Journal of Information Science: Research Trends* 10 (1). https://doi.org/10.36311/1981-1640.2016.v10n1.07.p56.

SARI. 2020. "SARI Documentation." 2020. https://docs.swissartresearch.net/.

Universität Zürich, Kunsthistorisches Institut. 2018. "Swiss Art Research Infrastructure (SARI)." https://swissartresearch.net/.

University of Sheffield. 2013. "Module 2 : Introduction to IE and ANNIE," 1995–2013.

W3C. 2002. “How the Semantic Web Works.” 2002. https://doi.org/10.1007/978-1-4020-4749-5_3.

———. 2011. “Linked Data.” The Serial Librarian. 2011. https://www.w3.org/standards/semanticweb/data.

———. 2012. “Semantic Web.” W3.Org. 2012. https://www.w3.org/standards/semanticweb/.

———. 2013. “SPARQL.” 2013. https://www.w3.org/2001/sw/wiki/SPARQL.

Apêndices

Apêndice 1 – Guião completo

Guião para a experiência de avaliação da nova interface

O presente guião surge no âmbito de uma dissertação com o título "Avaliação da migração de registos de arquivo para dados ligados no projeto EPISA", para o Mestrado de Ciência de Informação. Este guião tem como principal objetivo avaliar uma nova interface que está a ser desenvolvida no âmbito do projeto EPISA, comparando-a com a interface anterior (DigitArq). Desta forma, solicito a sua colaboração para efetuar um conjunto de tarefas acompanhadas, enquanto se anota as respostas e os resultados encontrados, seguindo-se algumas questões genéricas de controlo e de sugestões.
Os dados recolhidos serão usados apenas para fins académicos da referida tese, e de melhoria da nova interface.

Muito obrigada,
Margarida Augusto

Pergunta 1: Que documento é devolvido quando se pesquisa a Cota Atual "*Empresa Pública do Jornal O Século, Joshua Benoliel, Cota Desconhecida 011*"?

- Torre do Tombo: https://digitarq.arquivos.pt/

Pergunta 1.1: Efetue uma pesquisa através da nova interface que devolva o mesmo documento, sem utilizar o título completo.

- Interface: https://episa-labs.inesctec.pt/archclient/

Pergunta 2: Qual o código de referência do documento composto sobre o processo de Jerónima da Cruz?

- Torre do Tombo: https://digitarq.arquivos.pt/

Pergunta 2.1: Efetue a mesma pesquisa através da nova interface.

- Interface: https://episa-labs.inesctec.pt/archclient/

Pergunta 3: Qual é o número de resultados devolvidos quando se efetua uma pesquisa sobre o Mosteiro de São Bernardo de Portalegre?

- Arquivo Distrital de Portalegre: https://digitarq.adptg.arquivos.pt/

Pergunta 3.1: Efetue a mesma pesquisa através da nova interface.

- Interface: https://episa-labs.inesctec.pt/archclient/

Pergunta 4: Efetuar uma pesquisa que devolva as secções existentes no fundo "Juízo da Índia e Mina". Registar o número de secções encontradas.

- Torre do Tombo: https://digitarq.arquivos.pt/

Pergunta 4.1: Efetue a mesma pesquisa através da nova interface.

- Interface: https://episa-labs.inesctec.pt/archclient/

Pergunta 5: Ao efetuar uma pesquisa sobre "registos de baptismos", registe o número de resultados devolvidos. Seguidamente, fazer a mesma pesquisa considerando o intervalo temporal entre 1811 e 1822, registando, igualmente, o número de resultados.

- Arquivo Distrital do Porto: https://pesquisa.adporto.arquivos.pt/

Pergunta 5.1: Efetue a mesma pesquisa através da nova interface.

- Interface: https://episa-labs.inesctec.pt/archclient/migration

Pergunta 6: Efetue uma pesquisa para encontrar o nome do fundo, bem como o nome dos diferentes níveis de descrição, onde está presente o documento simples "Um aquário de 2,40x1,00 m" com o código de referência PT/TT/AS/A/001/000142.

- Torre do Tombo: https://digitarq.arquivos.pt/

Pergunta 6.1: Efetue a mesma pesquisa através da nova interface.

- Interface: https://episa-labs.inesctec.pt/archclient/

Apêndice 2 – Campos utilizados no registo das respostas

Tabela 26 - Explicação de campos utilizados no registo das respostas ao guião

Forma de resposta	Como efetuou a pesquisa para chegar ao que era pedido
Resposta final	Qual a resposta que o participante deu
Eventuais erros cometidos	Erros que possam ter sido cometidos
Notas	Notas importantes a registar ao longo de toda a experiência
Contacto com o DigitArq	Tipo de contacto que tem com o DigitArq (se já utilizou o DigitArq várias vezes, se nunca utilizou ou se utilizou poucas vezes)
Tempo	Tempo que demorou a responder à pergunta
Comentário/Sugestão	Comentário e/ou sugestão relativamente à nova interface
Filtros	Considera que seria útil possuir alguns filtros que estivessem sempre ativos durante todo o processo de pesquisa?
Formação	Formação académica

Apêndice 3 – Organização do registo das respostas

Tabela 27 - Organização da folha de Excel no registo das respostas ao guião

Participante 1					
DigitArq			DigitArq+		
Pergunta 1	Forma de resposta		Pergunta 1	Forma de resposta	
	Resposta final			Resposta final	
	Eventuais erros cometidos			Eventuais erros cometidos	
	Notas			Notas	
Pergunta 2	Forma de resposta		Pergunta 2	Forma de resposta	
	Resposta final			Resposta final	
	Eventuais erros cometidos			Eventuais erros cometidos	
	Notas			Notas	
Pergunta 3	Forma de resposta		Pergunta 3	Forma de resposta	
	Resposta final			Resposta final	
	Eventuais erros cometidos			Eventuais erros cometidos	
	Notas			Notas	
Pergunta 4	Forma de resposta		Pergunta 4	Forma de resposta	
	Resposta final			Resposta final	
	Eventuais erros cometidos			Eventuais erros cometidos	
	Notas			Notas	
Pergunta 5	Forma de resposta		Pergunta 5	Forma de resposta	
	Resposta final			Resposta final	
	Eventuais erros cometidos			Eventuais erros cometidos	
	Notas			Notas	
Pergunta 5	Forma de resposta				
	Resposta final				
	Eventuais erros cometidos				
	Notas				
Pergunta 6	Forma de resposta		Pergunta 6	Forma de resposta	
	Resposta final			Resposta final	
	Eventuais erros cometidos			Eventuais erros cometidos	
	Notas			Notas	
Contacto com o DigitArq					
Comentário/Sugestão					
Filtros					
Tempo					
Idade					
Formação					

Apêndice 4 – Resultados das pesquisas efetuadas



Figura 11 - Resultado da pesquisa avançada (Pergunta 1)



Figura 12 - Resultado da pesquisa simples (Pergunta 1)

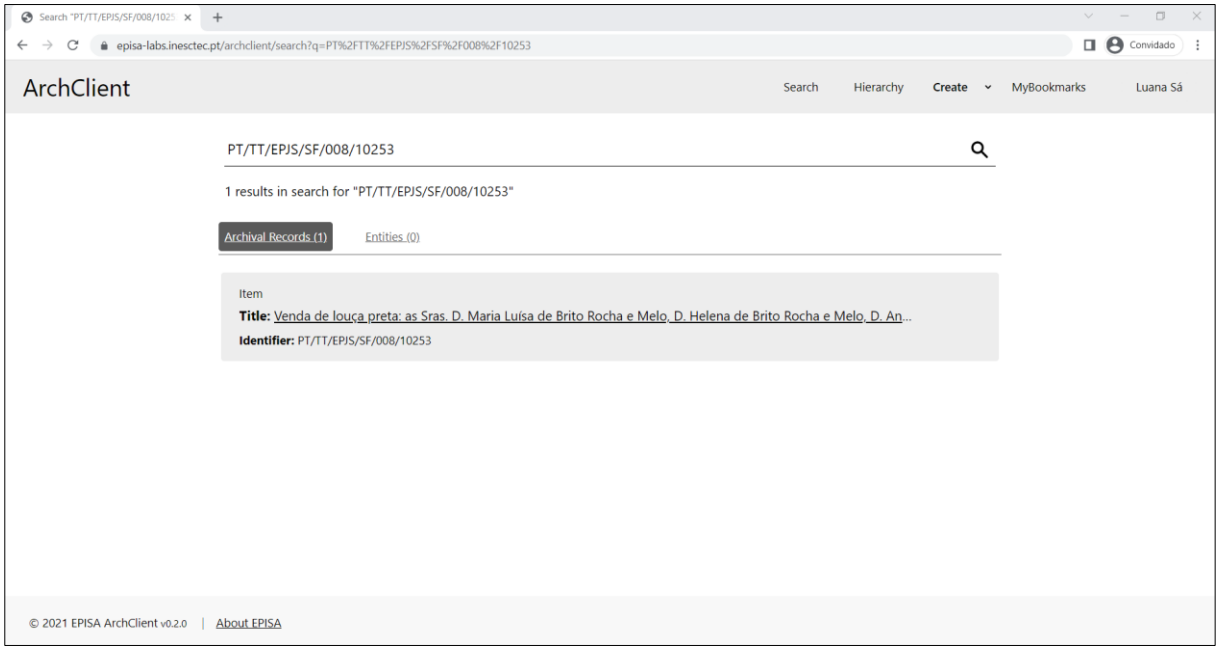


Figura 13 - Resultado da pesquisa no DigitArq+ (Pergunta 1)

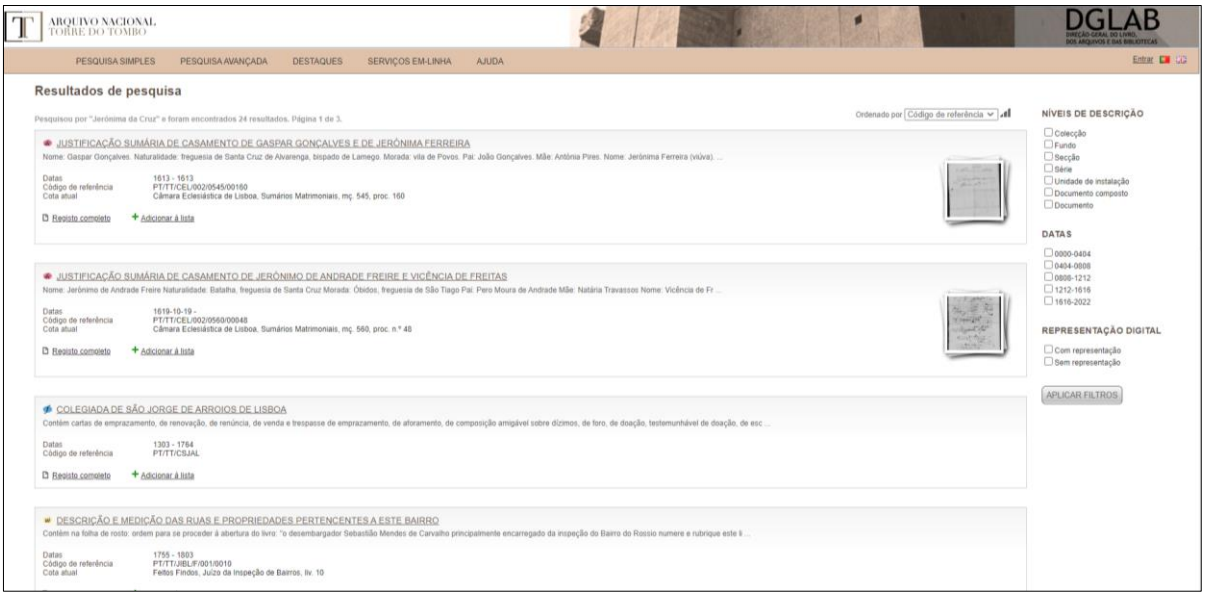


Figura 14 - Resultado da pesquisa simples ao utilizar "Jerónima da Cruz" (Pergunta 2)

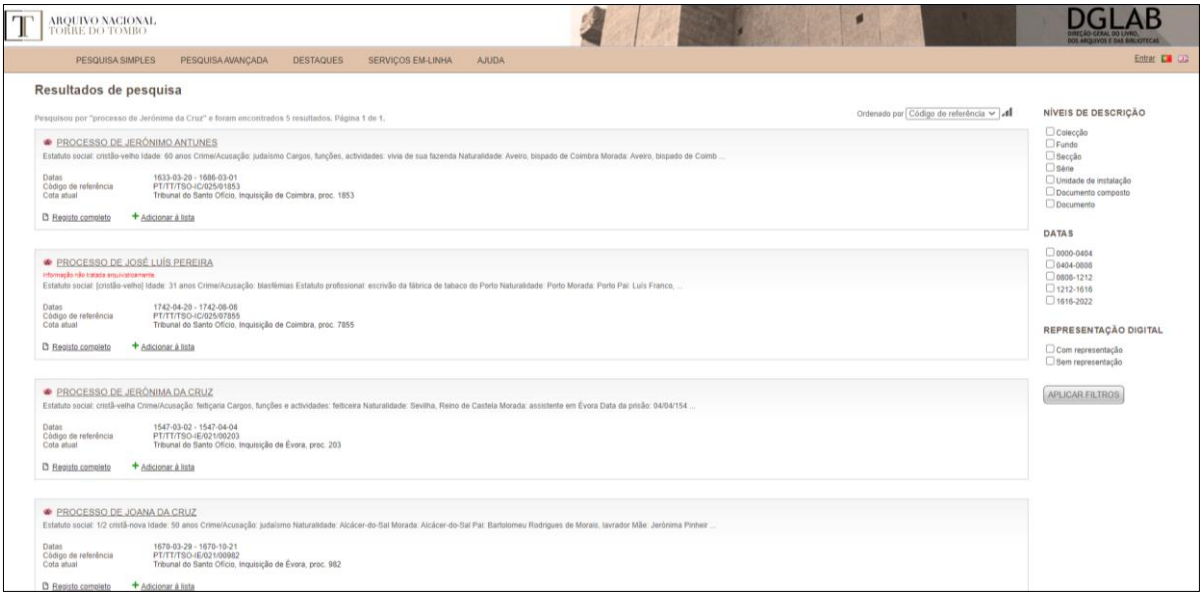


Figura 15 - Resultado da pesquisa simples ao utilizar "processo de Jerónima da Cruz" (Pergunta 2)

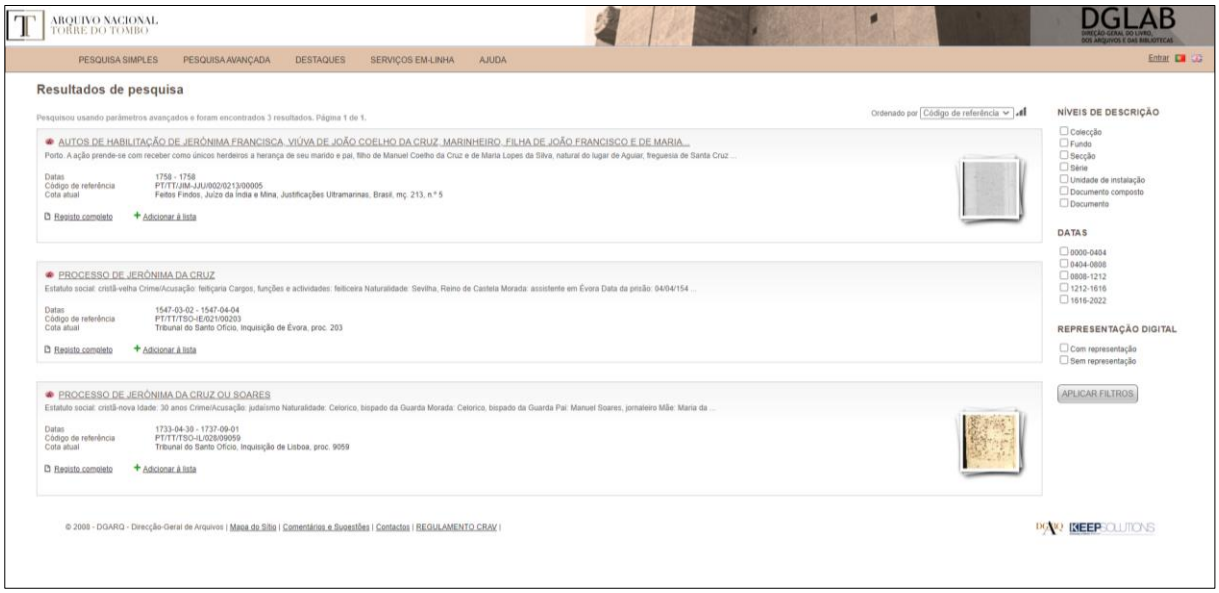


Figura 16 - Resultado da pesquisa avançada ao utilizar "Jerónima da Cruz" (Pergunta 2)

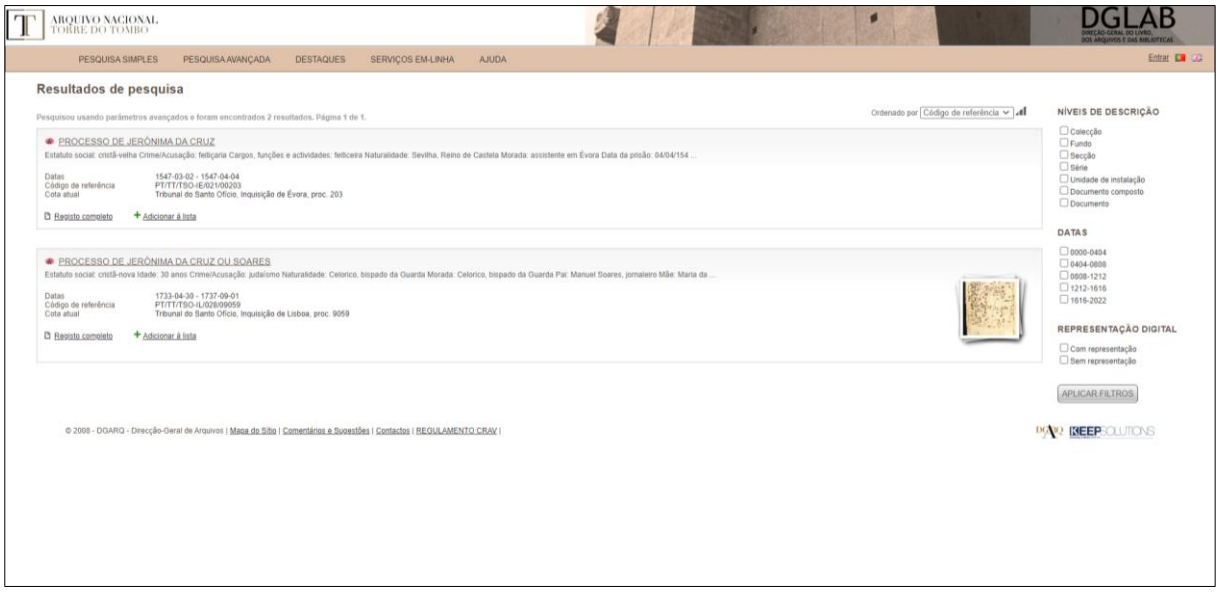


Figura 17 - Resultado da pesquisa avançada ao utilizar "processo de Jerónima da Cruz" (Pergunta 2)

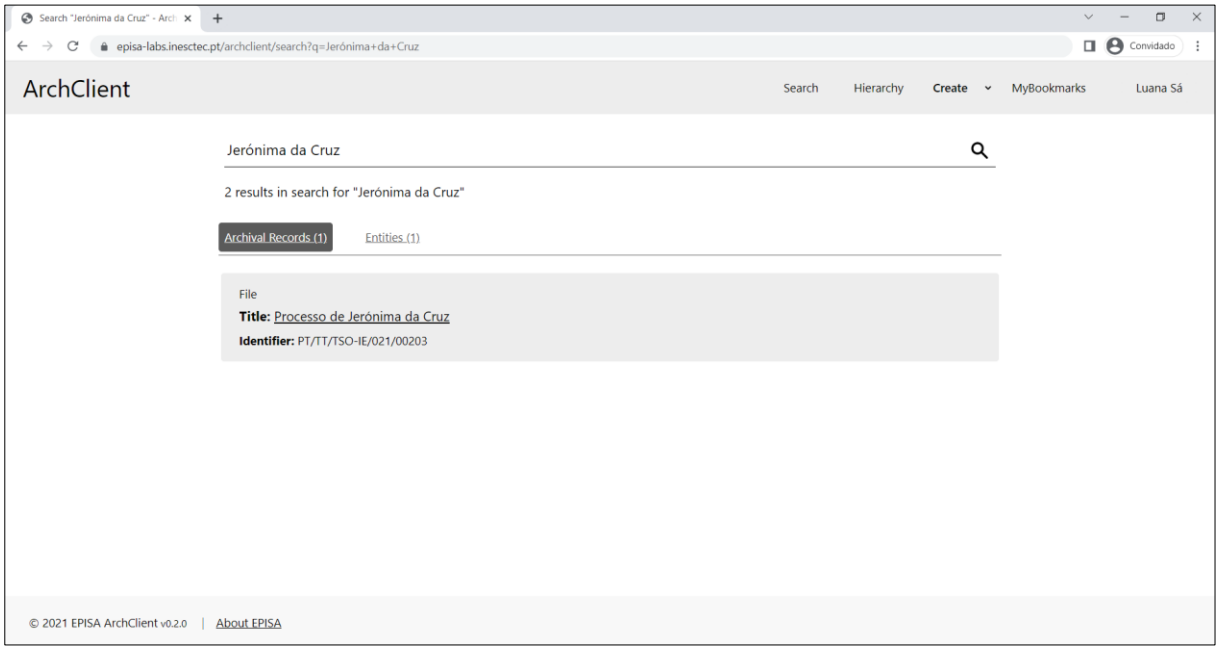


Figura 18 - Resultado da pesquisa no DigitArq+ ao utilizar "Jerónima da Cruz" (Pergunta 2)

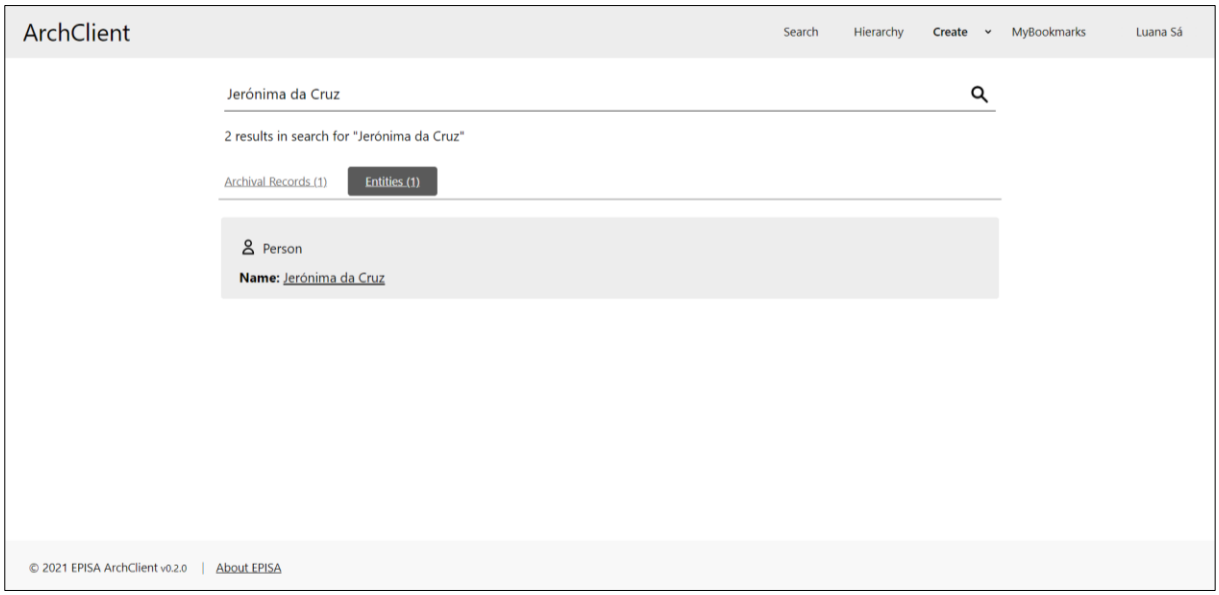


Figura 19 - Entidade mostrada pelo DigitArq+ (Pergunta 2)

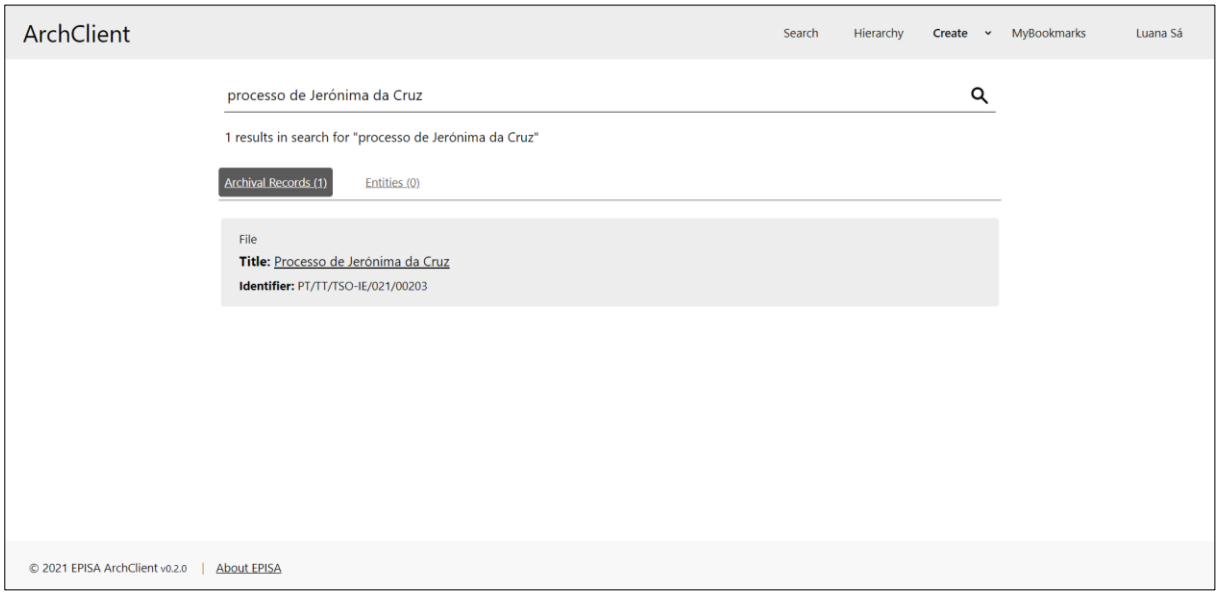


Figura 20 - Resultado da pesquisa no DigitArq+ ao utilizar "processo de Jerónima da Cruz" (Pergunta 2)

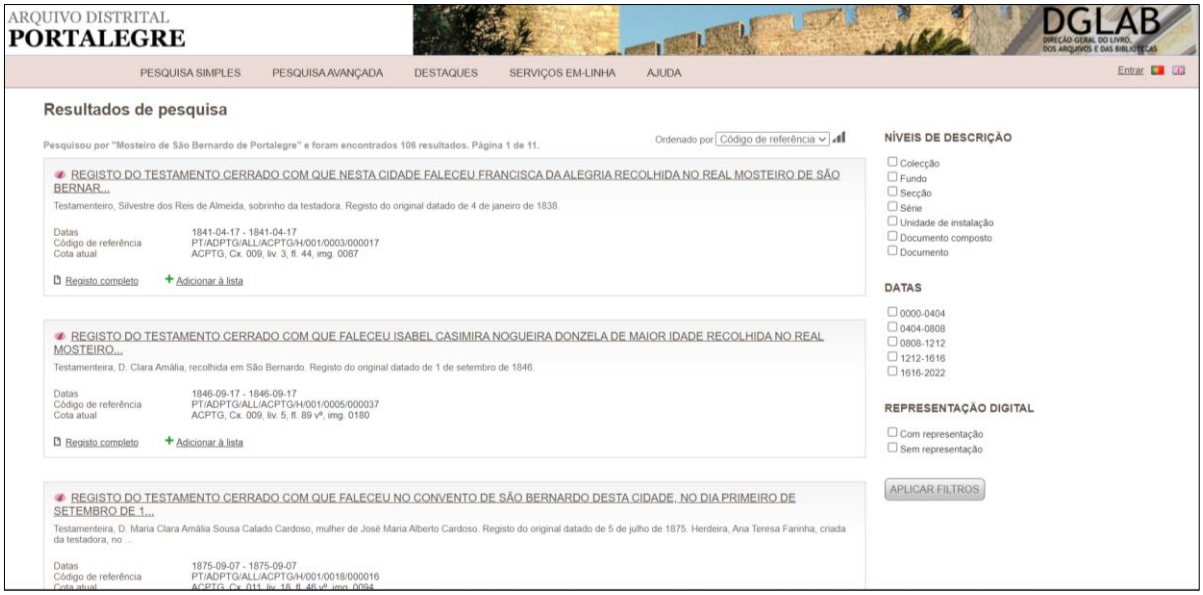


Figura 21 - Resultado da pesquisa simples ao utilizar "Mosteiro de São Bernardo de Portalegre" (Pergunta 3)

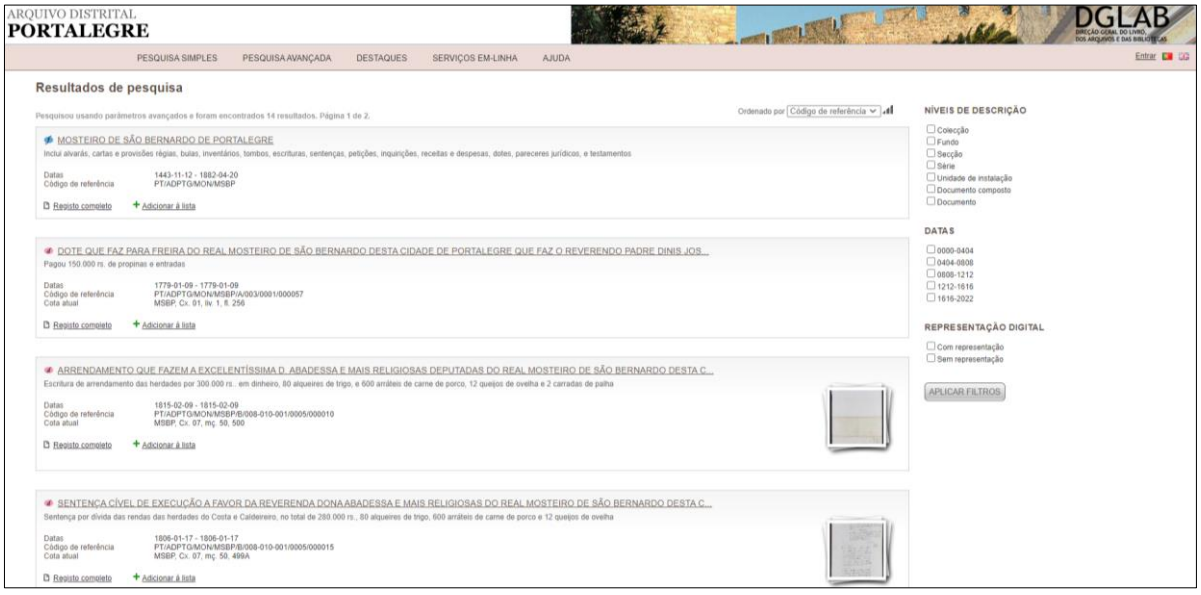


Figura 22 - Resultado da pesquisa avançada ao utilizar "Mosteiro de São Bernardo de Portalegre" (Pergunta 3)

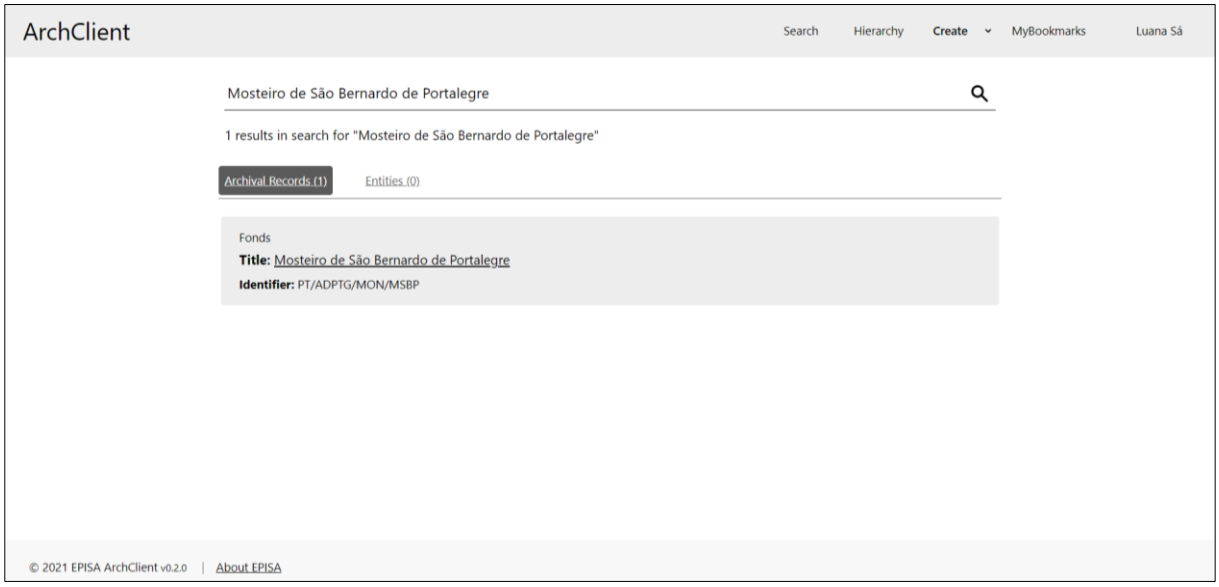


Figura 23 - Resultado da pesquisa no DigitArq+ ao utilizar "Mosteiro de São Bernardo de Portalegre" (Pergunta 3)

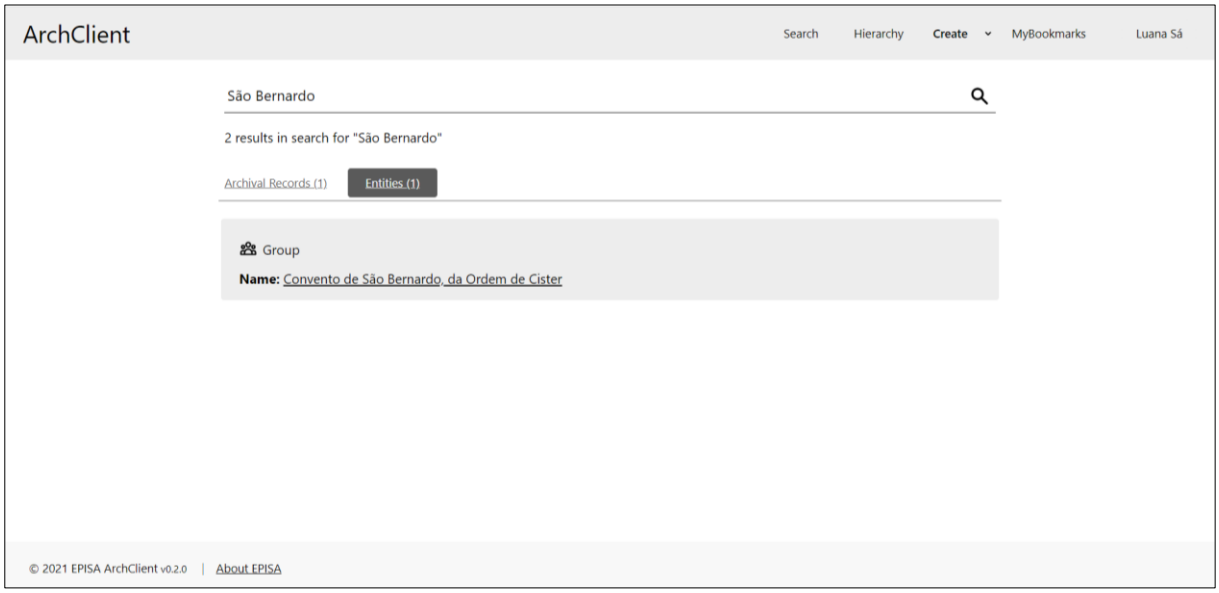


Figura 24 - Entidade mostrada pelo DigitArq+ (Pergunta 3)

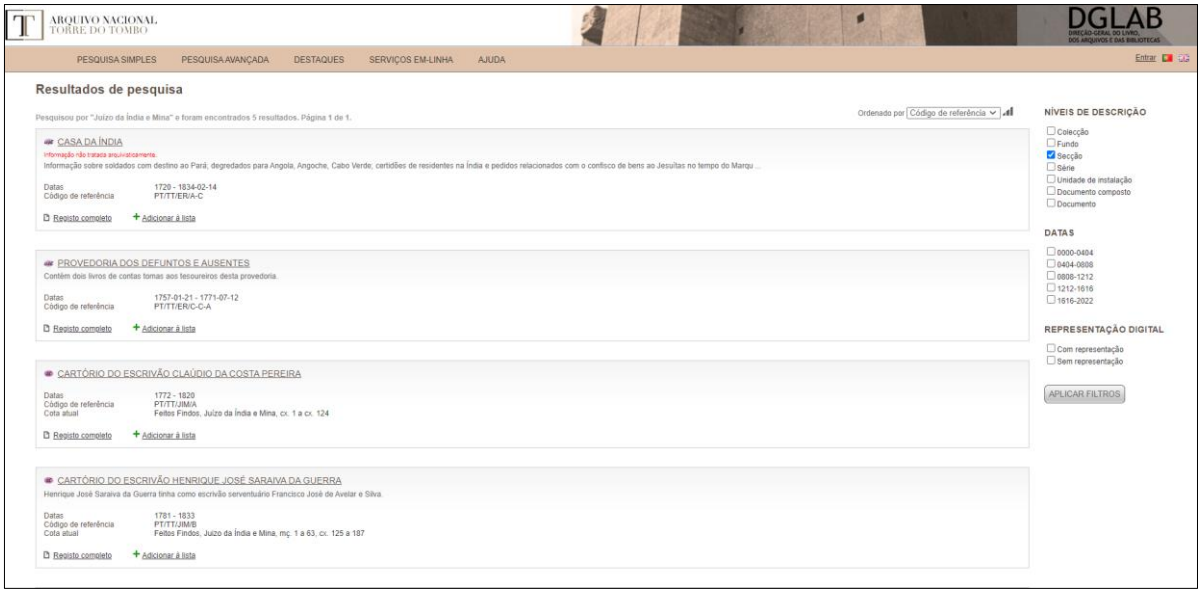


Figura 25 - Resultado da pesquisa simples ao utilizar "Juízo da Índia e Mina" (Pergunta 4)

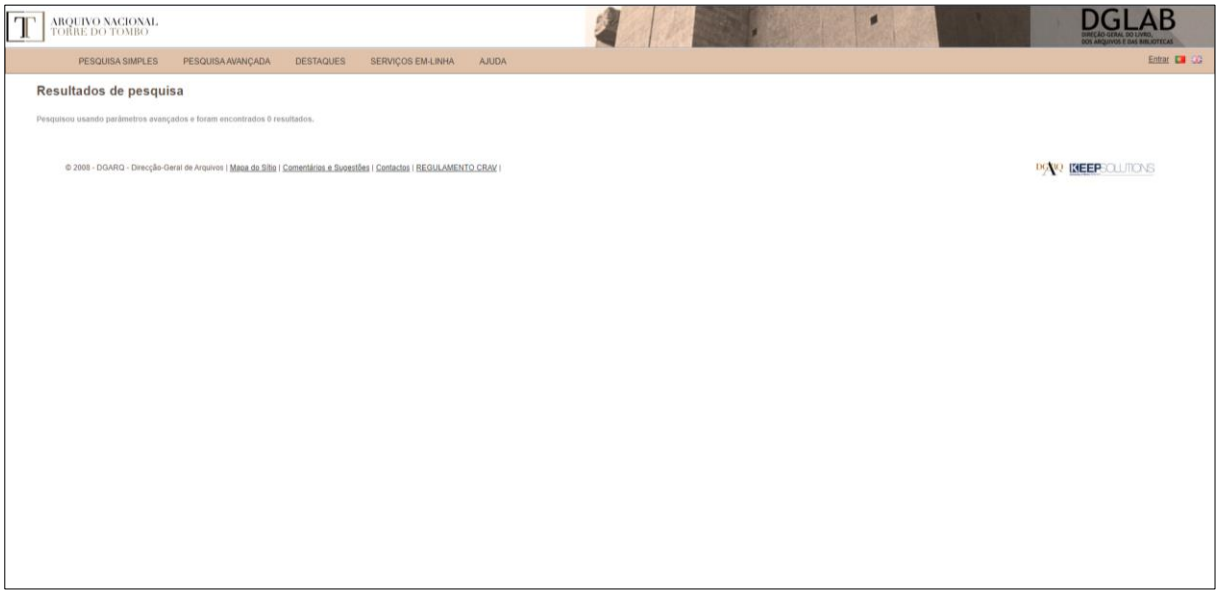


Figura 26 - Resultado da pesquisa avançada ao utilizar "Juízo da Índia e Mina" (Pergunta 4)

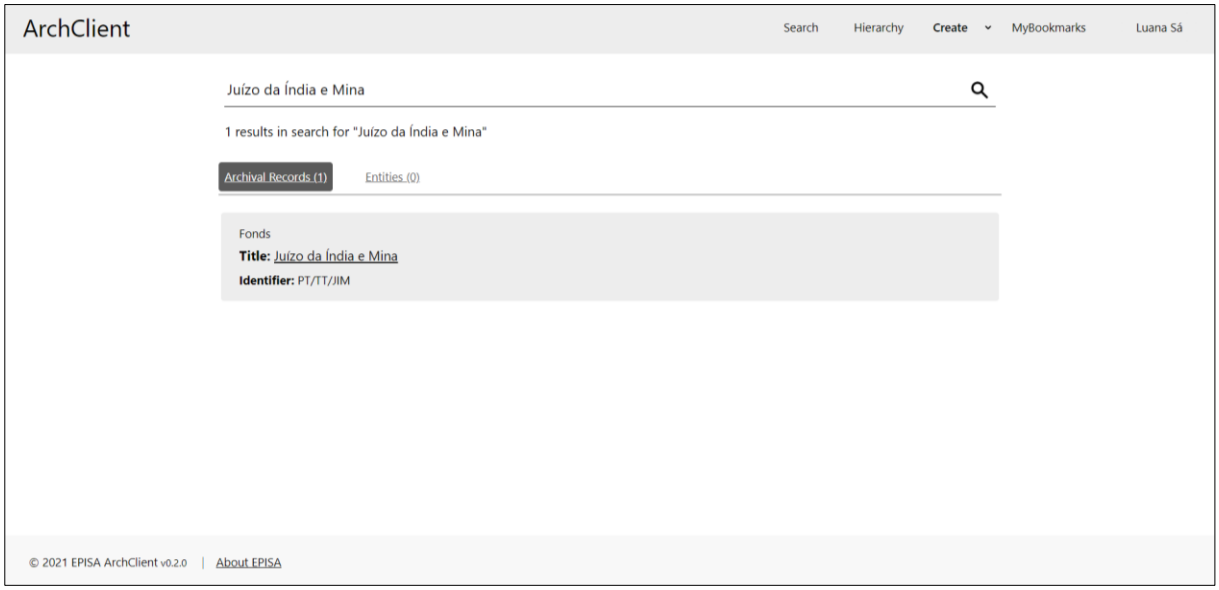


Figura 27 - Resultado da pesquisa no DigitArq+ ao utilizar "Juízo da Índia e Mina" (Pergunta 4)

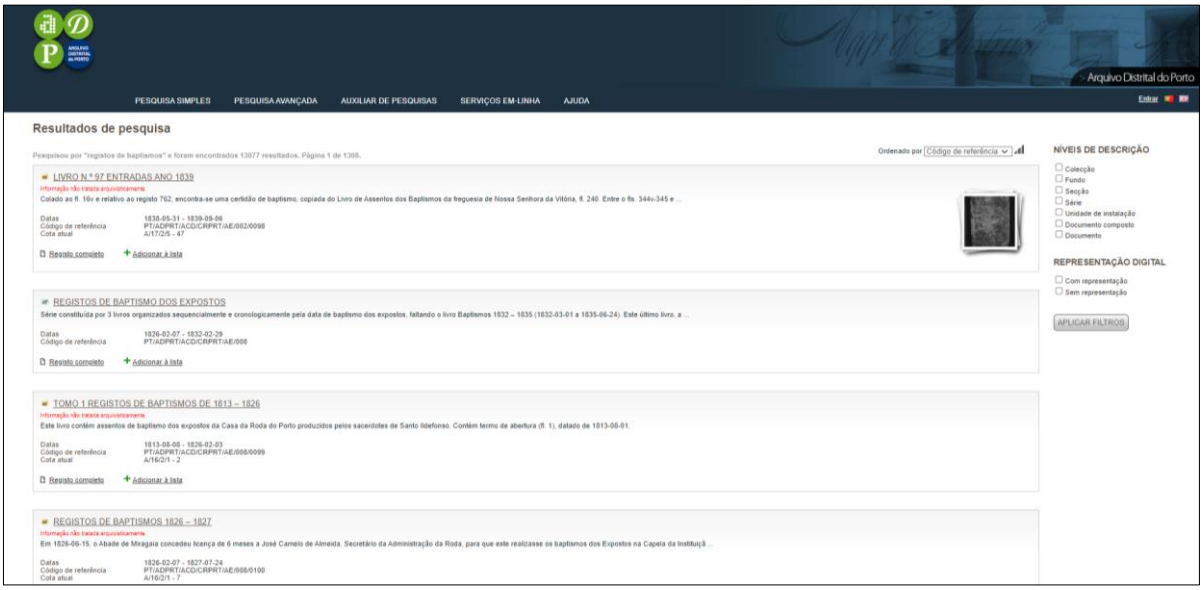


Figura 28 - Resultado da pesquisa simples ao utilizar "registos de baptismos" (Pergunta 5)

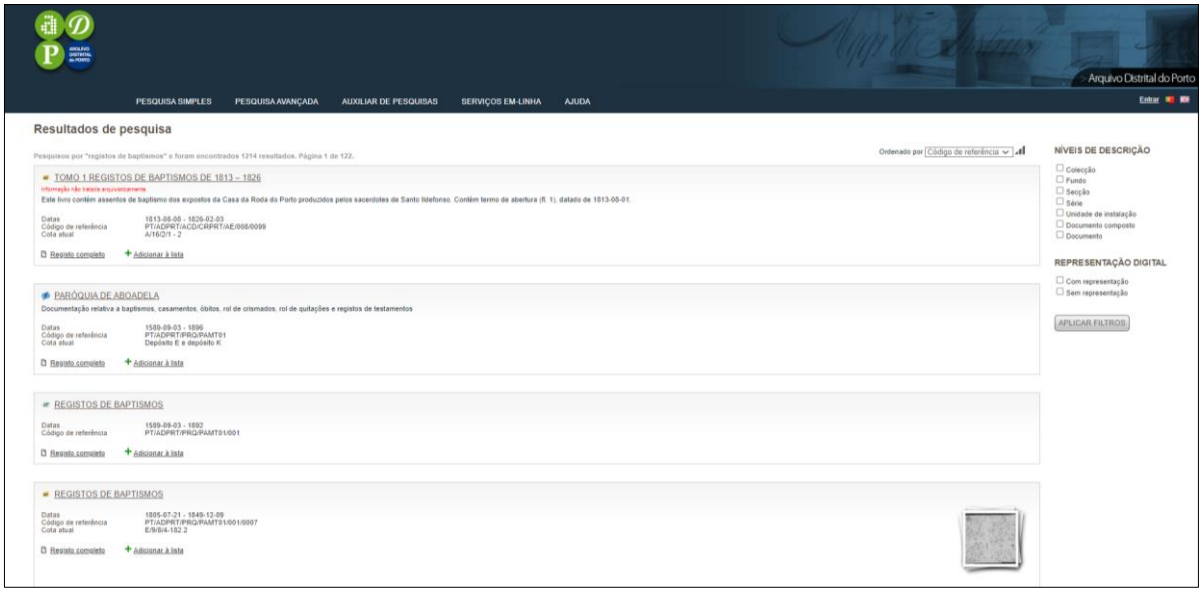


Figura 29 - Resultado da pesquisa simples ao utilizar "registos de baptismos", com o intervalo temporal 1811-1822 (Pergunta 5)

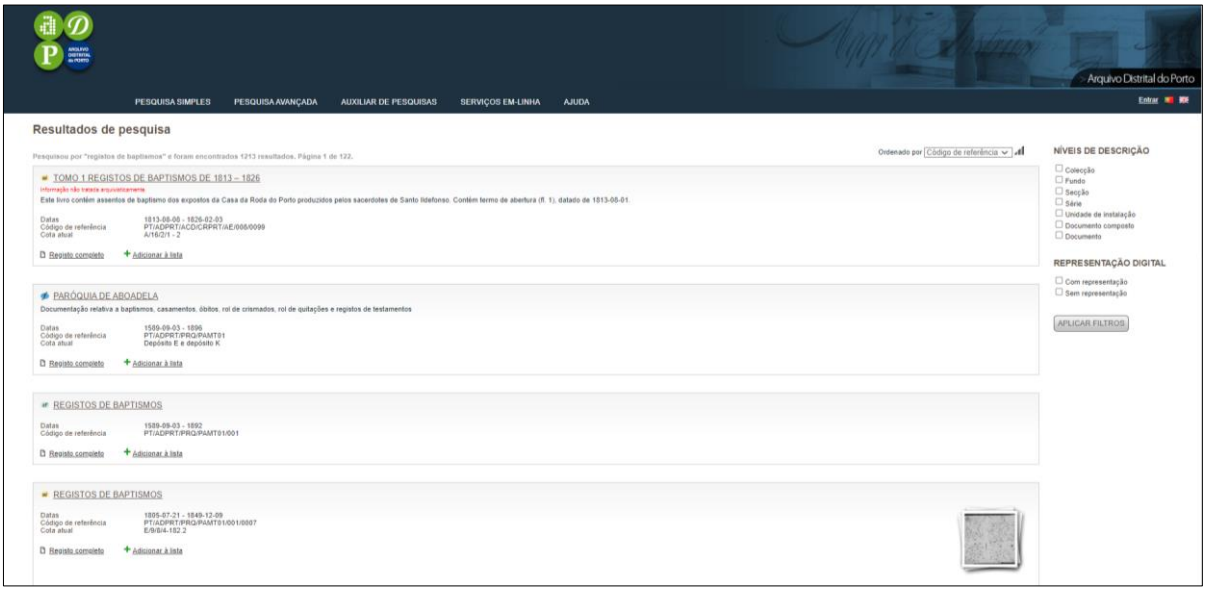


Figura 30 - Resultado da pesquisa simples ao utilizar "registos de baptismos", com a data completa (Pergunta 5)

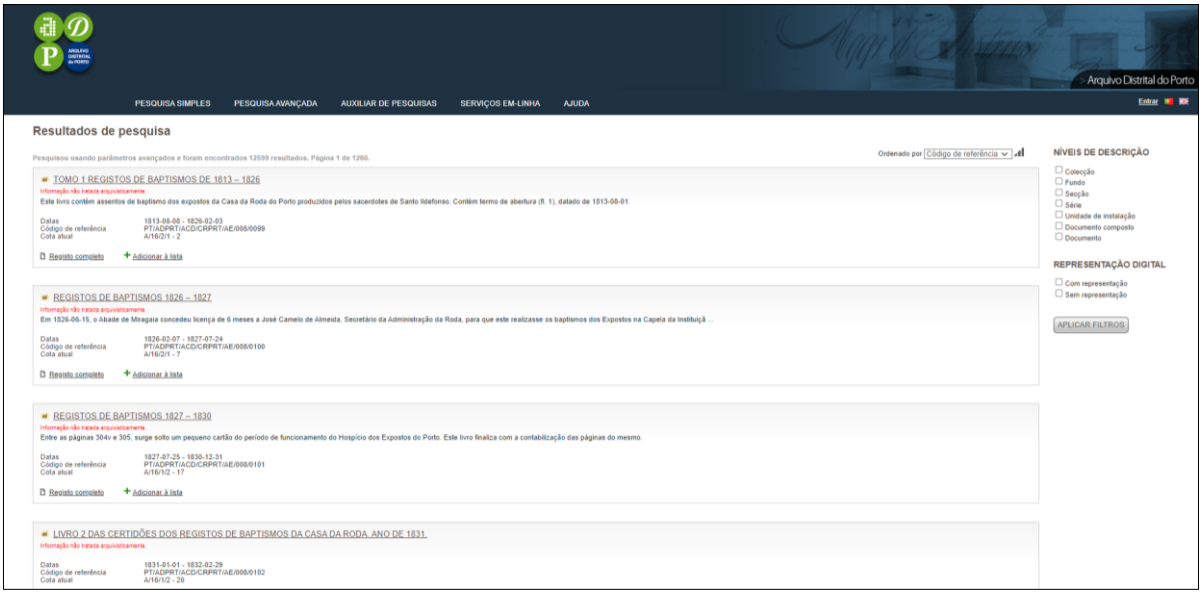


Figura 31 - Resultado da pesquisa avançada ao utilizar "registos de baptismos" (Pergunta 5)

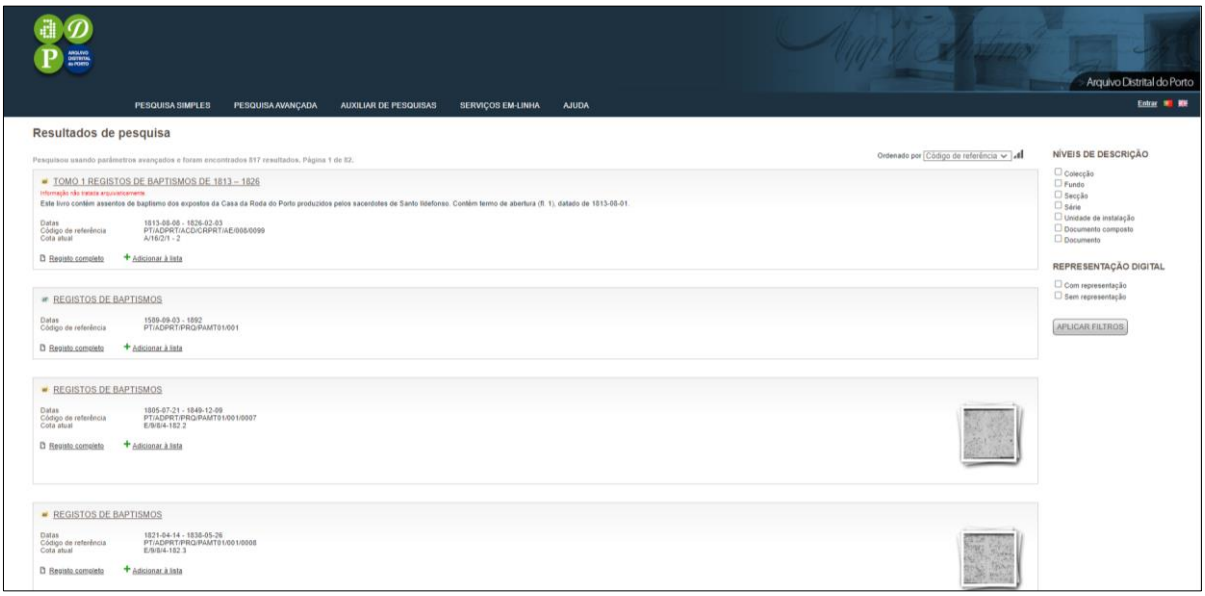


Figura 32 - Resultado da pesquisa avançada ao utilizar "registos de baptismos", com o com o intervalo temporal 1811-1822 (Pergunta 5)

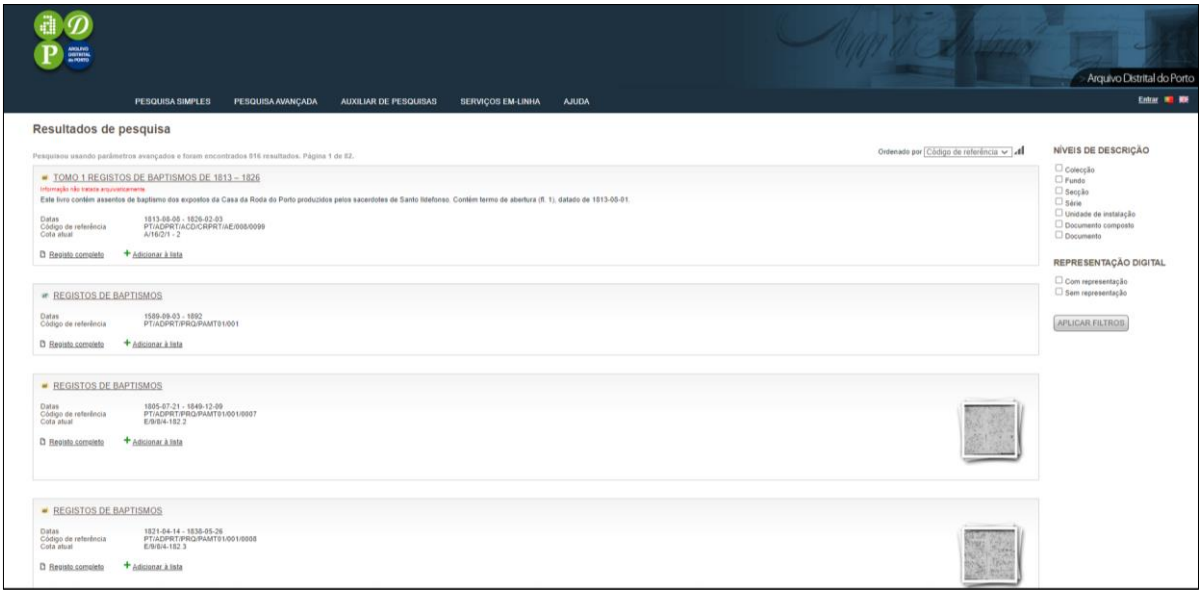


Figura 33 - Resultado da pesquisa avançada ao utilizar "registos de baptismos", com a data completa (Pergunta 5)

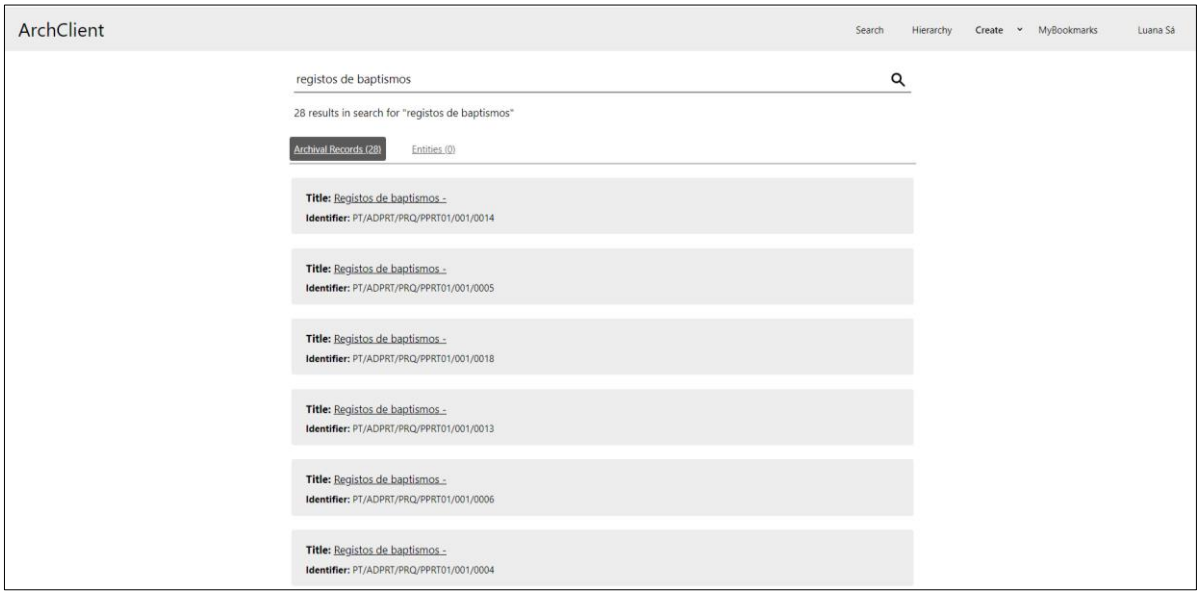


Figura 34 - Resultado da pesquisa no DigitArq+ ao utilizar "registos de baptismos" (Pergunta 6)

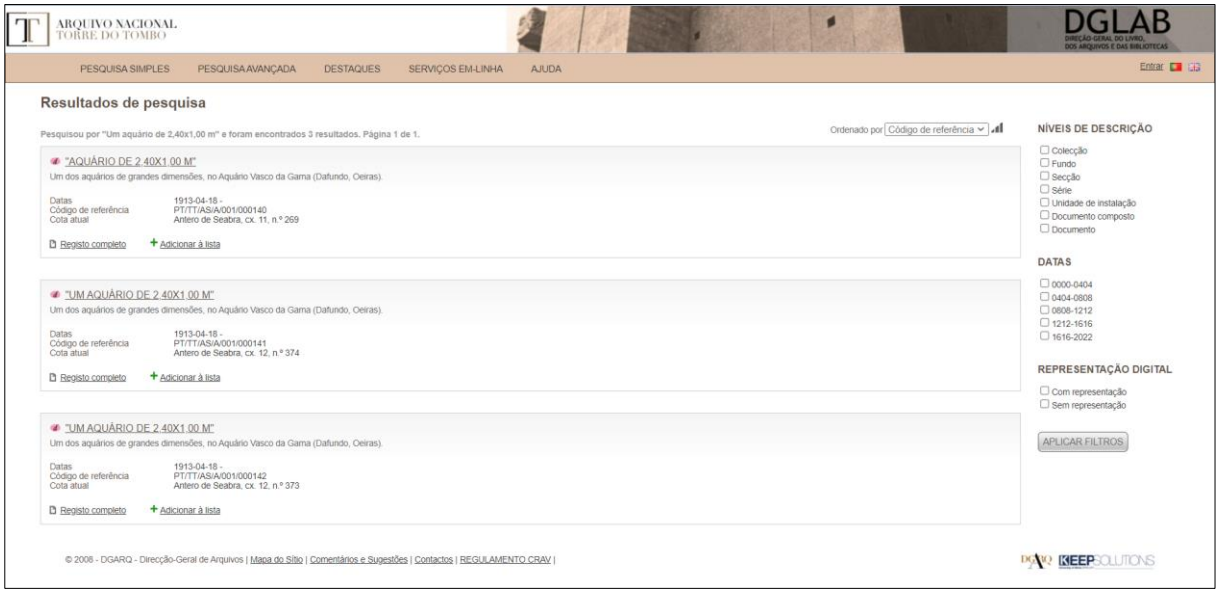


Figura 35 - Resultado da pesquisa simples ao utilizar "Um aquário de 2,40x1,00 m" (Pergunta 6)

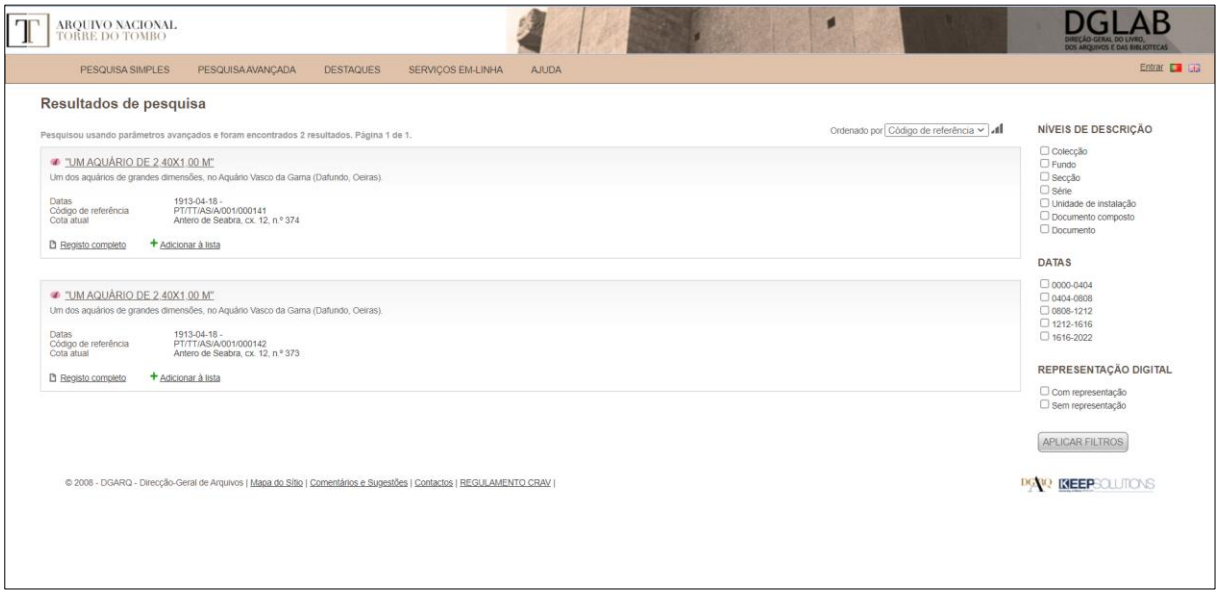


Figura 36 - Resultado da pesquisa avançada ao utilizar "Um aquário de 2,40x1,00 m" (Pergunta 6)

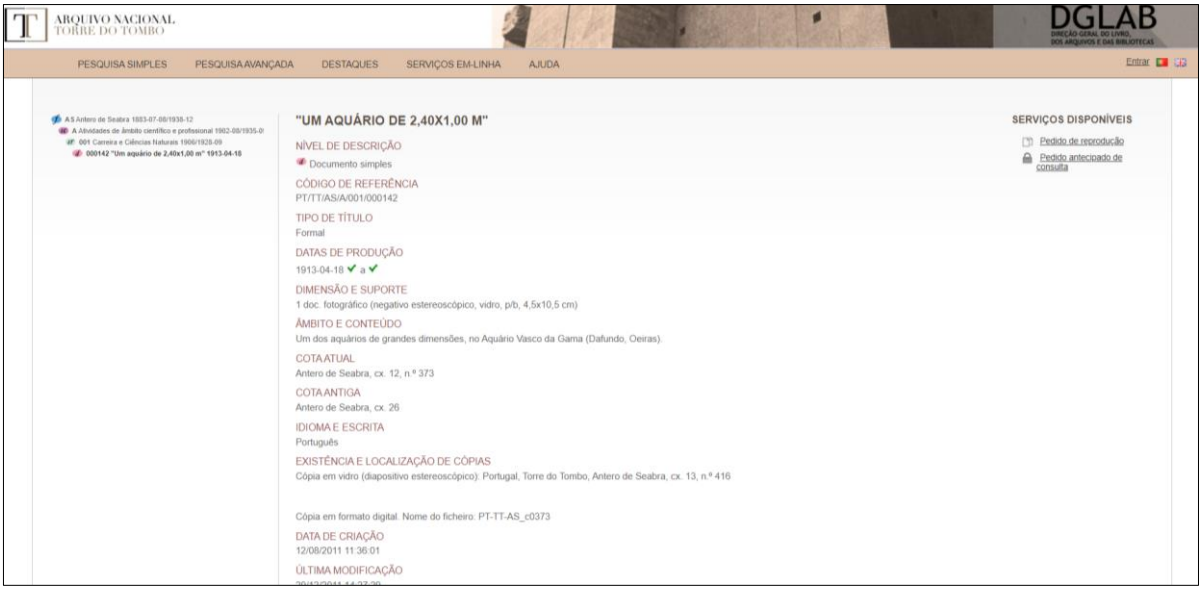


Figura 37 - Resultado da pesquisa avançada ao utilizar o código de referência "PT/TT/AS/A/001/000142" (Pergunta 6)

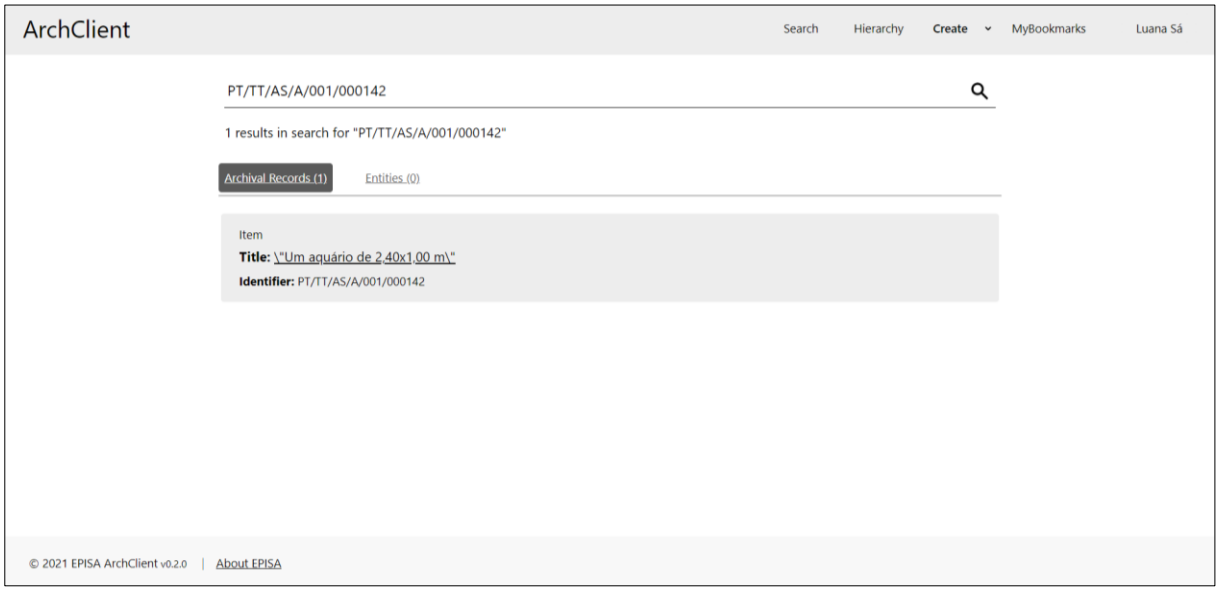


Figura 38 - Resultado da pesquisa no DigitArq+ ao utilizar o código de referência "PT/TT/AS/A/001/000142" (Pergunta 6)