



Full-Stack Data Governance

DAVID SOUSA PINTO

Junho de 2023

Full-Stack Data Governance

David Pinto

**Dissertação para obtenção do Grau de Mestre em
Engenharia Informática, Área de Especialização em
Sistemas Computacionais**

Orientador: Professor Paulo Oliveira

Porto, 30 de junho de 2023

Declaração de Integridade

Declaro ter conduzido este trabalho académico com integridade.

Não plagiei ou apliquei qualquer forma de uso indevido de informações ou falsificação de resultados ao longo do processo que levou à sua elaboração.

Portanto, o trabalho apresentado neste documento é original e de minha autoria, não tendo sido utilizado anteriormente para nenhum outro fim.

Declaro ainda que tenho pleno conhecimento do Código de Conduta Ética do P.PORTO.

ISEP, Porto, 30 de junho de 2023

Resumo

Com a modernização da tecnologia existente, assistiu-se a uma evolução constante dos sistemas de informação das mais variadas organizações, criando uma cultura virada para os dados, assumindo estes o papel de um dos bens mais valiosos dos dias de hoje.

A crescente produção de informação induziu a chegada de um paradigma de *too much data*, em que a gestão de informação se tornou insustentável e o processo de tomada de decisão moroso. Um traço definidor deste paradigma é a existência de silos de informação ou, por outras palavras, de uma arquitetura de dados fragmentada. Isto é, a informação existe, muitas vezes, de forma duplicada e/ou tendo por base estruturas e processamentos dispares.

Urge, então, criar soluções para uniformizar a consulta de informação (através de Data Warehouses, Data Lakes, etc.) e ativar estruturas de Governo para a mesma (e.g., Data Owners, Data Stewards, Data Custodians, etc.), para que os consumidores de informação tenham confiança nos dados que estão a trabalhar e consigam agilizar o processo de tomada de decisão.

Neste sentido, esta dissertação estuda o assunto Data Governance e algumas plataformas de Data Governance, desenvolvendo uma solução numa dessas plataformas, Collibra. Deste modo, para se perceber as vantagens da solução preconizada em relação à atual, a solução é analisada e comparada usando métricas para avaliar atributos de qualidade e levantar possíveis melhorias às soluções analisadas.

Deste trabalho resulta a implementação End-to-End de Data Governance em Collibra que proporciona uma maior facilidade de trabalho, num ambiente mais amigável e com melhor desempenho, comprovado pelas métricas avaliadas e a documentação de todo o processo de desenvolvimento da mesma.

Palavras-chave: Data Governance, Collibra, Data Management, Data Stewardship, Meta-data, Data Lineage, Data Catalog

Abstract

With the modernization of existing technology, there has been a constant evolution of the information systems of the most celebrated organizations, creating a culture focused on data, assuming these the role of one of the most valuable assets of today.

The growing production of information induced the arrival of a *too much data* paradigm, in which information management became unsustainable and the decision-making process slow. A defining feature of this paradigm is the existence of information silos or, in other words, a fragmented data architecture. That is, information often exists in duplicate and/or based on disparate structures and processes.

Therefore, it is urgent to create solutions to standardize the query of information (through Data Warehouses, Data Lakes, etc.) and activate Government structures for the same (e.g., Data Owners, Data Stewards, Data Custodians, etc.), so that information consumers have confidence in the data they are working with and are able to streamline the decision-making process.

In this sense, this dissertation studies the subject of Data Governance and some Data Governance platforms, developing a solution on one of these platforms, Collibra. Thus, in order to understand the advantages of the recommended solution in relation to the current one, the solution is analyzed and compared using metrics to evaluate quality attributes and raise possible improvements to the analyzed solutions.

This work resulted in the End-to-End implementation of Data Governance in Collibra, which provides greater ease of work, in a friendlier environment and with better performance, as evidenced by the evaluated metrics and the documentation of the entire development process.

Conteúdo

Lista de Figuras	xiii
Lista de Tabelas	xvii
Lista de Acrónimos	xix
1 Introdução	1
1.1 Contexto	1
1.2 Descrição do Problema	2
1.3 Objetivos	2
1.4 Metodologia	3
1.5 Estrutura do Documento	4
2 Estado da Arte	5
2.1 Conceitos Fundamentais	5
2.1.1 Data	5
2.1.2 Data Management	6
2.1.3 Data Governance	7
2.1.3.1 Data Stewardship	9
2.1.3.2 Data Architecture	11
2.1.3.3 Data Modeling and Design	12
2.1.3.4 Data Security	13
2.1.3.5 Data Warehouse/Data Lake	14
2.1.3.6 Data Quality	16
2.1.3.7 Data Lineage	17
2.2 Plataformas de Data Governance	17
2.2.1 Collibra	17
2.2.2 Alation	18
2.2.3 SAP Master Data Governance	18
2.2.4 Comparação	19
3 Análise	21
3.1 Ambiente Collibra	21
3.1.1 Data Gathering Tool (DGT)	23
3.2 Modelo de Domínio	25
3.3 Engenharia de Requisitos	26
3.3.1 Atores	27
3.3.2 Requisitos	27
3.3.3 Requisitos funcionais	29
3.3.4 Requisitos não funcionais	29
3.4 Casos de Uso	30

4	Análise de Valor	33
4.1	Processo de Inovação	33
4.2	New Concept Development	35
4.2.1	Identificação da Oportunidade	36
4.2.2	Análise de Oportunidade	36
4.2.3	Geração de ideias	38
4.2.4	Seleção de ideias	38
4.3	Valor da Solução	41
4.3.1	Valor	41
4.3.2	Valor para o Cliente	42
4.3.3	House of Quality	43
4.3.4	Valor Percecionado	44
4.3.5	Proposta de Valor	45
5	Design	47
5.1	Arquiterura	47
5.2	Especificação de Funcionalidade	48
5.2.1	Ativação Data Owner/Steward/Custodian R.1, R.2 e R.3	48
5.2.2	Lineage Path R.4	50
5.2.3	Carregamento em massa R.6	51
5.2.4	Data Set Management R.9	52
5.2.5	Customização Landing Page R.8	52
5.2.6	Data Gathering Tool R.16, R.17 e R.18	54
5.2.6.1	Criação da DGT	54
5.2.6.2	Criação de Iniciativas	55
5.2.6.3	Ativação da Equipa	56
5.2.6.4	Abertura e Fecho de Secções	56
5.2.6.5	Criação de Tabelas	57
5.2.6.6	Criação de Colunas	57
5.2.6.7	Adição de utilizadores à Data Access	58
5.2.6.8	Importação de Secções de outras DGTs	58
6	Implementação	59
6.1	Workflows BPMN	59
6.1.1	Ativação Data Owner	60
6.1.2	Carregamento em Massa	64
6.1.3	Lineage Path	65
6.1.4	Apresentação de logs	68
6.1.5	Data Set Management	69
6.1.6	Data Gathering Tool	70
6.1.6.1	Criação da DGT	70
6.1.6.2	Criação de Iniciativas	72
6.1.6.3	Ativação da Equipa	74
6.1.6.4	Abertura e Fecho de Secções	76
6.1.6.5	Criação de Tabelas	78
6.1.6.6	Criação de Colunas	80
6.1.6.7	Adição de Utilizadores a Data Access	82
6.1.6.8	Importação de Business Understanding e Data Understanding	84

6.1.6.9	Importação dos ativos de Controlo de Qualidade	85
6.1.6.10	Importação de Secções de outras DGTs	87
6.1.6.11	End DGT	88
6.2	Costumização da Landing Page	88
7	Avaliação da Solução	91
7.1	Hipóteses de Investigação	91
7.2	Indicadores e Fontes de Informação	92
7.3	Metodologia de Avaliação	92
7.3.1	Quantitative Evaluation Framework	92
7.3.2	Testes de Aceitação do Utilizador (UATs)	93
7.3.2.1	Testes de Integração	93
7.3.2.2	Testes funcionais	93
7.3.2.3	Testes de ponta-a-ponta	94
7.3.2.4	Testes de aceitação	94
7.3.2.5	UAT efetuado	94
8	Conclusão	97
	Bibliografia	99
A	Tipos de Ativos e as suas Relações	103

Lista de Figuras

2.1	The DAMA Wheel	7
2.2	Knowledge Area Context Diagram	9
2.3	Data Governance Program	10
2.4	Sources of Data Security Requirements	14
3.1	Tipos de Ativos e as suas Relações	22
3.2	Arquitetura da DGT	25
3.3	Modelo de Domínio	26
3.4	Diagrama de Casos de Uso	31
4.1	Innovation Process Parts	34
4.2	NCD	35
4.3	SWOT	37
4.4	HoQ	43
4.5	Canvas	46
5.1	Diagrama de Componentes	48
5.2	Workflow de Ativação do <i>Data Owner</i>	49
5.3	Workflow do Lineage Path	51
5.4	Workflow de Ativação <i>Data Owner</i> em Massa	51
5.5	Workflow de Ativação <i>Data Steward</i> em Massa	52
5.6	Workflow do Data Set Management	52
5.7	Landing Page Mockup	54
5.8	Workflow de Criação de DGTs	55
5.9	Workflow de Criação de Iniciativas	55
5.10	Workflow de Ativação da Equipa	56
5.11	Workflow da Abertura e Fecho de Secções	57
5.12	Workflow de Criação de Tabelas	57
5.13	Workflow de Criação de Colunas	58
5.14	Workflow de Adição de Utilizadores em Data Access	58
5.15	Workflow de Importação de Secções de outras DGTs	58
6.1	Variáveis do Formulário	59
6.2	Formulário Email	60
6.3	Excerto de Código - Criação de Ativos	60
6.4	Ativação Data Owner - Botão de execução	61
6.5	Ativação Data Owner - Seleção do utilizador	61
6.6	Ativação Data Owner - Escolha do utilizador	62
6.7	Ativação Data Owner - Notificação aceitação	62
6.8	Ativação Data Owner - Email de notificação	63
6.9	Ativação Data Owner - Pedido justificação	63
6.10	Ativação Data Owner - Seleção novo utilizador	64

6.11 Ativação em Massa - Vista Global	65
6.12 Lineage Path - Botão	66
6.13 Lineage Path - Escolha de Atributos	66
6.14 Lineage Path - Escolha de Não Repetidos	67
6.15 Lineage Path - Escolha de Dados Pessoais	67
6.16 Relatório Final	68
6.17 Vista dos Logs	69
6.18 Imagem data set	70
6.19 Imagem tabela	70
6.20 DGT - Comunidade das DGTs	71
6.21 Criação da DGT - Botão	71
6.22 Criação da DGT - Preenchimento dos Campos	72
6.23 Criação da DGT - DGT Criada	72
6.24 Criação de Iniciativa - Botão	72
6.25 Criação de Iniciativa - Preenchimento dos Campos	73
6.26 Criação de Iniciativa - Iniciativa já Existe	73
6.27 Criação de Iniciativa - Vista das Iniciativas	74
6.28 Ativação da Equipa - Botão	74
6.29 Ativação da Equipa - Seleção da Responsabilidade	75
6.30 Ativação da Equipa - Escolha do Utilizador	75
6.31 Ativação da Equipa - Finalizar ou Continuar	76
6.32 Abertura e Fecho de Secções - Botão	77
6.33 Abertura e Fecho de Secções - Seleccionar Secções	77
6.34 Abertura e Fecho de Secções - Notificação	78
6.35 Abertura e Fecho de Secções - Secção Aberta	78
6.36 Criação de Tabelas - Botão	79
6.37 Criação de Tabelas - Escrita do Nome	79
6.38 Criação de Tabelas - Tabela Criada	80
6.39 Criação de Colunas - Botão	80
6.40 Criação de Colunas - Seleção da tabela	81
6.41 Criação de Colunas - Escrita do Nome	81
6.42 Criação de Colunas - Continuidade	82
6.43 Criação de Colunas - Coluna Criada	82
6.44 Adição de Utilizadores a Data Access - Botão	83
6.45 Adição de Utilizadores a Data Access - Escrita do Nome	83
6.46 Adição de Utilizadores a Data Access - Utilizador Criado	84
6.47 DGT - Botão de Importação de Business Understanding e Data understanding	84
6.48 Importação de <i>Business Understanding</i> e <i>Data Understanding</i> - Secção <i>Business Understanding</i>	85
6.49 Importação de <i>Business Understanding</i> e <i>Data Understanding</i> - Secção <i>Data Understanding</i>	85
6.50 DGT - Botão de Importação dos Ativos de Controlo de Qualidade	86
6.51 DGT - Importação dos Ativos de Controlo de Qualidade	86
6.52 DGT - Importação dos Ativos de Controlo de Qualidade	86
6.53 DGT - Botão de Importação de Secções	87
6.54 DGT - Importação de Secções	87
6.55 DGT - Importação de Secções	88
6.56 DGT - Botão End DGT	88
6.57 Landing Page Widget	89

6.58	Landing Page Image1	89
6.59	Landing Page Image2	90
6.60	Landing Page navbar	90
7.1	Espaço tridimensional	92
7.2	Resumo das UATs	95
A.1	Tipos de Ativos e as suas Relações	104

Lista de Tabelas

1.1	Passos da metodologia DSRM	3
2.1	Data Warehouse versus Data Lake (O'Leary 2014)	16
2.2	Comparação entre as várias Plataformas de Data Governance	19
3.1	Demonstração de Relação	23
4.1	Differences between Fuzzy Front End (FFE) and New Product Development (NPD). Adaptado de (Peter e Koen 2002)	34
4.2	Escala Preferências AHP	39
4.3	Prioridade relativa de cada critério	39
4.4	Prioridade relativa das alternativas segundo o critério - Tempo de Desenvolvimento	40
4.5	Prioridade relativa das alternativas segundo o critério - Complexidade	40
4.6	Prioridade relativa das alternativas segundo o critério - Desempenho	40
4.7	Prioridades relativas das soluções	41

Lista de Acrónimos

AHP	Analytic Hierarchy Process.
BPMN	Business Process Model and Notation.
DA	Data Architecture.
DG	Data Governance.
DGT	Data Gathering Tool.
DL	Data Lineage.
DM	Data Management.
DOS	Data Organisation Structures.
DQ	Data Quality.
DS	Data Security.
DSS	Data Stewardship.
FFE	Fuzzy Front End.
HoQ	House of Quality.
NCD	New Concept Development.
NPD	New Product Development.
QEF	Quantative Evalution Framework.
QFD	Quality Function Deployment.
UI	User Interface.

Capítulo 1

Introdução

Este capítulo introduz o trabalho realizado no âmbito da dissertação de Mestrado, contextualizando e descrevendo o problema a que pretende dar resposta. São também elencados os objetivos e a abordagem seguida para a concretização dos mesmos. Por fim, é apresentada a estrutura do presente documento.

1.1 Contexto

A tecnologia impulsionou a Revolução Industrial 1.0 com fábricas a vapor em 1765; A Revolução Industrial 2.0 com produção e fabricação em massa na década de 1870; A Revolução Industrial 3.0 com o surgimento da eletrónica e automação em 1969; e atualmente, a Revolução Industrial 4.0 com o rápido desenvolvimento de “*smart things*” como inteligência artificial em *machine e deep learning*, robôs inteligentes, *smart devices*, manufatura aditiva inteligente, cidades inteligentes, etc (Governance e Activity 2020). Todas estas revoluções proporcionaram sempre quer desafios, quer oportunidades, para as organizações.

À medida que a Revolução Industrial 4.0 é introduzida na cultura dominante e produz um dilúvio cada vez maior de informações, a soma coletiva dos dados mundiais crescerá de 33 zettabytes (ZB) em 2018 para 175 ZB em 2025, com uma taxa composta de crescimento anual de 61% (Patrizio 2018).

Com recursos de informação tão ricos, que podem ser comparados a “petróleo da era digital”, a governação e gestão eficazes podem trazer crescimento económico e uma forte vantagem competitiva no mercado, no entanto para tal é necessário que as empresas estejam devidamente preparadas para serem capazes de fazer uma escolha inteligente sobre os dados, nomeadamente, quais são importantes, onde os armazenar, a quem pertencem e que informações valiosas se consegue obter deles.

Tendo isto em consideração, uma plataforma onde todos estes processos possam ser efetuados é vital para a gestão de Data Governance (DG). Daí que o cliente, empresa Portuguesa no setor da energia, optou por uma passagem de todos os processos que possui ao dia de hoje e que se encontram propagados ao longo de diversas ferramentas, para a plataforma Collibra¹, de modo a terem tudo centralizado num único local, tornando assim a gestão de DG mais fácil e controlada.

Deste modo, esta dissertação alude esta mesma passagem dos processos de DG para a plataforma Collibra.

¹É uma plataforma de DG voltada para empresas que as capacita a encontrar significado nos seus dados e melhorar as decisões de negócio.

1.2 Descrição do Problema

Atualmente a empresa que propôs este trabalho de mestrado possui uma implementação de DG incompleta, uma vez que não existe uma Data Lineage ² que efetue uma monitorização dos ciclos de vida dos dados, nem um repositório com todas as bases de dados, Data Lakes ³, Data Warehouses ⁴, etc, onde se consiga de forma rápida e eficiente visualizar-se a origem de todos os dados, quem são os responsáveis pelos mesmos, de como é que esses dados estão a ser usado e por quem.

Por outro lado também existe o problema dos silos de informação, isto é, a informação existe, muitas vezes, de forma duplicada e/ou tendo por base estruturas e processamentos dispares, o que faz com que os consumidores de informação não tenham confiança nos dados que estão a trabalhar, tal como não terem o conhecimento de onde é que os dados se encontram para os utilizar.

E arcaica, visto que o processo de criação e desenvolvimento de Data Gathering Tool (DGT) ⁵, ao dia de hoje, é feito em Excel, em que embora haja uma evolução da mesma, sendo adicionado novos componentes ao processo, o mesmo ainda não consegue englobar tudo que eles necessitam.

1.3 Objetivos

Os pilares e objetivos da solução proposta enquadram-se nos seguintes pontos, bem como alguns dos artefactos a entregar, visando suprir lacunas caracterizadas pela existência de silos de informação. Em particular, com este trabalho pretende-se alcançar os seguintes objetivos:

Capacidades completas de Data Lineage:

- Disponibilizar o ciclo de vida de dados completo com capacidades de drill-down;
- Mostrar relações entre atributos e correspondentes endpoints;
- Construir diagramas de linhagem, automatização, e exportação;
- Ilustrar as alterações sofridas durante o ciclo de vida dos dados;
- Integrar com os restantes componentes de Governação de Dados;

Capacidades completas de Data Stewardship e Workflows:

- Caracterizar a gestão de acesso aos dados;
- Promover a cooperação entre as partes interessadas sobre as regras de Governação;
- Automatizar a plataforma através do desenvolvimento de workflows, aproveitando estruturas de código;

²Data Lineage fornece informações históricas de registos de dados e as suas origens.

³Data Lakes é usado para descrever a forma evolutiva dos repositórios de dados.

⁴Data Warehouse é uma coleção de dados orientada a objetos, integrada, variável no tempo e não atualizável, que é usada para apoiar o processo de decisão e negócios inteligentes

⁵Processo desenvolvido pelo cliente para a catalogação de novos ativos, que por sua vez foram criados para fornecer novos dados ou tipos de dados às partes interessadas

- Integrar com artefactos de governo de dados para incremento da cooperação entre os diferentes stakeholders;

Desenvolvimento de Data Gathering Tools:

- Criar e desenvolver DGTs;
- Adicionar todas as pessoas necessárias na participação da DGT;
- Visualizar as DGTs finalizadas;

1.4 Metodologia

O autor optou por utilizar a metodologia Design Science Research Methodology (DSRM). Esta metodologia é direcionada no desenvolvimento de novos artefactos de modo a tratar um tipo generalizado de problema, avaliando a utilidade dos artefactos para resolver problemas desse tipo (Teperi, Gotcheva e Aaltonen 2021).

Para complementar a análise, os passos subentendidos da metodologia em foco encontram-se evidenciados na tabela 1.1.

Tabela 1.1: Passos da metodologia DSRM

Passo	Preocupações	Output para o próximo passo	Ponto partida
1. Identificar problema	- Definir problema - Evidenciar importância	Inferência	Iniciação centrada no problema
2. Definir objetivos de uma solução	O que realizaria um melhor novo artefacto	Teoria	inicialização centrada no objetivo
3. Desenho e desenvolvimento	Artefacto	Conhecimento de como fazer	Inicialização centrada no desenho e desenvolvimento
4. Demonstração	- Encontrar contexto apropriado - Usar artefacto para resolver o problema	Métricas, conhecimento de análises	Cliente / Contexto iniciado
5. Avaliação	- Observar a eficácia - Eficiência e iterar novamente para o desenho	Conhecimento disciplinar	
6. Comunicação	- Publicações académicas - Profissionais		

Conforme exposto na tabela, existe uma relação entre cada passo da metodologia com os vários capítulos do projeto.

O capítulo Estado da Arte 2 não tem relação com nenhum passo da metodologia.

O capítulo da Engenharia de requisitos 3, que é referente ao passo 3. Desenho e desenvolvimento, consiste na exposição dos requisitos, atores e casos de uso.

Depois o capítulo da Análise de Valor 4, que é referente aos passos 1. Identificar problema e 2. Definir objetivos de uma solução, endereça a contextualização do problema, juntamente com os objetivos a cumprir. Para além disso, são também analisados softwares existentes que podem ser utilizados para o desenvolvimento da solução e define-se os requisitos da solução a desenvolver.

O capítulo seguinte, 6, refere-se ao ponto 4. Demonstração, que consiste na implementação da solução, que, por sua vez, engloba a parte de desenvolvimento do ponto 3 e o ponto 4 devido ao facto de o desenvolvimento ocorrer simultaneamente com a demonstração.

Posteriormente à implementação é essencial que haja uma avaliação do que foi desenvolvido, daí que o capítulo seguinte, 7, corresponde ao passo 5. avaliação. Nesta secção é avaliada a solução implementada tendo em consideração o modelo de avaliação escolhido com o objetivo de certificar-se de que todos os requisitos são respondidos, também são elaborados testes que incluam os requisitos funcionais e não funcionais.

O último passo da metodologia, 6. Comunicação refere-se ao documento em si, visto que este passo tem como preocupações as publicações académicas e profissionais e as mesmas são discutidas ao longo do documento.

1.5 Estrutura do Documento

O presente documento encontra-se dividido em oito capítulos:

- **Introdução:** Introduce e enquadra o trabalho de mestrado, contextualizando e descrevendo o problema a que pretende dar resposta. Apresenta os objetivos e a abordagem seguida para a concretização dos mesmos.
- **Estado da Arte:** Apresenta um estudo da arte relativo a conceitos-chaves para o projeto e as melhores plataformas de DG com o objetivo de melhor se entender como devidamente estruturar uma implementação de DG.
- **Análise:** Apresenta o modelo de domínio e uma explicação da solução implementada, tal como os requisitos funcionais e não funcionais ligados aos casos de uso escolhidos para o projeto a desenvolver.
- **Análise de Valor:** Apresenta os diferentes métodos de análise adotados ao projeto com o objetivo de determinar a necessidade do mesmo, tal como a forma como este deve ser desenvolvido.
- **Design:** Detalha o processo de desenho da solução. É apresentada a arquitetura, tal como o diagrama de componentes da solução implementada, usando boas práticas de engenharia de software.
- **Implementação:** Descreve uma abordagem para a implementação das soluções em Colibra, assim como excertos que corroboram o seu desenvolvimento.
- **Avaliação da Solução:** Apresenta o processo de avaliação das soluções implementadas nas diferentes plataformas, seguindo a metodologia proposta e são detalhados os testes efetuados tal como o resultado da solução em relação aos mesmos.
- **Conclusão:** Apresenta um breve sumário do trabalho desenvolvido, os objetivos alcançados e possíveis melhorias para um trabalho futuro.

Capítulo 2

Estado da Arte

Antes de se começar a análise, desenho e implementação de qualquer projeto, é necessário primeiro perceber-se o que já existe sobre o tema.

Deste modo, ao longo deste capítulo são apresentados conceitos relacionados com o tema da dissertação, nomeadamente Data Management (DM), Data Governance (DG) e por último são também apresentadas possíveis ferramentas para o desenvolvimento da solução.

2.1 Conceitos Fundamentais

Para ser possível entender-se o propósito deste projeto é necessário primeiro perceber-se o que é Data Governance e tudo que o compõem. Deste modo, esta secção apresenta um estado da arte sobre Data Governance.

2.1.1 Data

Existem várias definições de dados, no entanto todas elas acentuam que representa factos¹. Desta forma, focando-se no tema de tecnologia, dados também são entendidos como informações que foram armazenadas em formato digital, embora os mesmos não sejam limitados a informações que foram digitalizadas.

Ainda assim, como, ao dia de hoje, conseguimos capturar tanta informação eletronicamente, chamamos a diversas coisas de "dados" que em tempos anteriores não teriam sido chamados de tal - coisas como nomes, moradas, datas de nascimento, o que alguém comeu no jantar de sábado, o livro mais recente que comprou (DAMA 2017).

Estes dados sobre as pessoas podem ser agregados, analisados e usados para gerar dinheiro, melhorar a saúde, influenciar opinião pública e oferecer melhor publicidade.

Além disso, como a constante evolução da tecnologia, a capacidade de recolher uma maior variedade de eventos e atividades aumentou, desde as repercussões do big bang até aos nossos batimentos cardíacos, e para recolher, guardar e analisar coisas que anteriormente não eram considerados dados para valerem a pena de serem guardados, nomeadamente vídeos, imagens e documentos, está perto de superar a capacidade de sintetizar estes dados em informações utilizáveis.

¹Dicionário New Oxford American define dados como "factos e estatísticas recolhidas juntas para análise" e a American Society for Quality (ASQ) define dados como "um conjunto de fatos recolhidos".

De acordo com Bulao (2023), até o final do ano de 2022, foram produzidos 94 zettabytes de dados, e é estimado que até 2025, infraestruturas cloud recebam mais de 200 zettabytes de dados.

Com isto, um aspeto bastante importante é saber como tirar vantagem do número exorbitante de dados que cada empresa possui e para tal, requer-se práticas de DM extensíveis e confiáveis.

2.1.2 Data Management

DM é essencial para obter eficácia operacional, reduzir custos e melhorar os serviços de troca de dados e informações entre instituições que são conciliadas a partir de múltiplas fontes. DM é o processo de mecanismos de política para controlar, proteger e aumentar o valor dos dados e informações ao longo do ciclo de vida dos mesmos (Institute 2020).

Dados em organizações, como por exemplo, dados de clientes e dados financeiros, podem ser considerados como ativos organizacionais, visto que os mesmos têm valor de negócio para as organizações. Como tal, os dados requerem uma gestão sistemática para processar e monitorizar a sua qualidade com o fim de incentivar as operações de negócio e alcançar os objetivos de negócio (Kaewkamol 2022).

DM deve também colaborar entre os negócios e o papel da tecnologia da informação, sendo que nas organizações visa a (Harwanto e Hidayanto 2022):

- Fornecer compreensão e suporte para informações em toda a organização e aos seus stakeholders.
- Garantir a integridade dos data ativos capturando-os, armazenando-os e protegendo-os.
- Garantir que os dados e informações tenham uma elevada qualidade.
- Garantir que a confidencialidade e privacidade dos dados dos stakeholders são mantidos.
- Impedir o acesso não autorizado à utilização dos dados e informações.
- Garantir que os dados possam oferecer valor à empresa e serem usados efetivamente.

Do mesmo modo que outros processos de gestão, DM deve equilibrar as necessidades estratégicas e operacionais. Este equilíbrio pode ser melhor atingido seguindo um conjunto de princípios que reconhecem características importantes e orientam DM.

De acordo com DAMA (2017), estes princípios são :

- Dados são um ativo com propriedades únicas.
- O valor dos dados pode e deve ser expressado em termos económicos.
- Gerir dados significa gerir a qualidade dos mesmos.
- São necessários metadados para gerir dados.
- É preciso planeamento para gerir dados.
- DM é multifuncional, requer uma série de habilidades e conhecimentos.
- DM requer uma perspectiva empresarial.

- DM deve ter em consideração uma variedade de perspetivas.
- DM é o gestor do ciclo de vida.
- Diferentes tipos de dados têm diferentes características do ciclo de vida.
- Gerir dados inclui a gestão os riscos associados com os dados.
- Os requisitos de DM devem orientar as decisões de tecnologia da informação
- A gestão eficaz dos dados requer um compromisso de liderança.

2.1.3 Data Governance

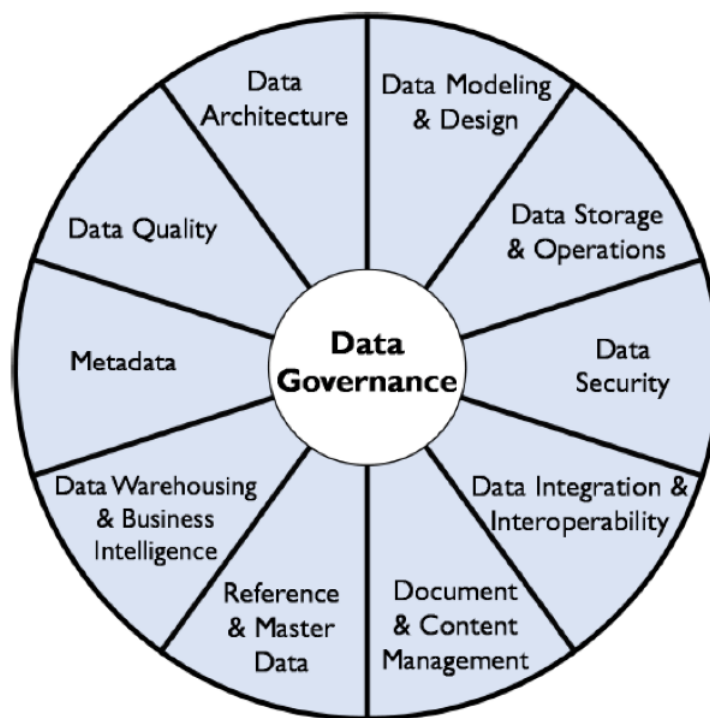


Figura 2.1: The DAMA-DMBOK2 Data Management Framework Wheel
DAMA (2017).

Como é possível observar-se na figura 2.1, DAMA (2017) relaciona DG com nove áreas distintas. E embora todas estas áreas sejam vitais num processo de DG, ao longo deste documento irão ser abordados apenas aquelas que se relacionam com o desenvolvimento da solução.

Existem várias definições de DG. Thomas (2006) indica que DG "refere-se a órgãos organizacionais, regras, diretivos de decisão e responsabilidades das pessoas e dos sistemas de informação à medida que executam processos relacionados com a informação". A autora declara também que "DG define as regras de compromisso que a direção seguirá conforme a organização usa os dados".

Newman e Logan (2006) definem DG como a "coleção de direitos de decisão, processos, padrões, políticas e tecnologias necessárias para gerir, manter e explorar informação como um recurso empresarial".

Smallwood (2014) descreve "Um programa DG deve fazer parte de um programa de governação IT. DG envolve processos e controlos para garantir que as informações ao nível dos dados são verdadeiras, precisas e únicas. Envolve a limpeza dos dados para remover dados corrompidos, imprecisos ou estranhos e retirar os duplicados, para eliminar ocorrências redundantes de dados".

Com base na definição acima, DG é indispensável pois determina as políticas e procedimentos para garantir um DM proativo e eficaz. A implementação de uma framework de DG que também permite a colaboração de vários níveis da organização para gerir dados em toda a empresa e fornece a capacidade de harmonizar os vários programas relacionados com dados com metas corporativas é também vital.

Uma vez que DG é a combinação de tecnologia e gestão ao longo de todo o processo de DM, que normalmente inclui a estrutura, política, sistema, tecnologia, ferramentas, padrão de dados, especificação do processo, supervisão e verificação da organização. Deste modo, segundo Liu (2020), DG visa alcançar as seguintes metas:

- Partilha compreensiva de dados, a não partilha é exceção. Partilha oportuna, atualização oportuna, não oportuna é exceção. Promover a aplicação de dados de elevado valor e quantificar os benefícios da aplicação de dados.
- Gestão unificada da fonte dos dados, coleção unificada de dados de negócio, standardização unificada de padrões de dados, libertação unificada de serviços de dados.
- Realizar todo o ciclo de vida da gestão de dados.
- Para obter a consistência de dados básicos importantes, tal como pessoal e ativos sob o mesmo calibre estatístico.
- A eficácia e a reputação da DG são gradualmente criados.

Dentro de um processo de DG, como em quase tudo, existe uma hierarquia. Esta hierarquia é composta por diferentes títulos com responsabilidades diferentes. Sendo que a seguir se encontram descritos os mais importantes e as suas funções:

- **Data Owner** Tem uma plataforma na qual supervisiona as atividades de governo realizadas nos seus domínios de interesse;
- **Data Governance/Delivery Officer** Dirige a definição e a gestão da estratégia e governo de dados e lidera a implementação de políticas de dados;
- **Data Steward** Define regras de limpeza para os seus conjuntos de dados em um catálogo centralizado e disponível para todos os utilizadores de dados e oferece suporte operacional aos Data Owner com a implementação de políticas e melhorias;
- **Data Custodian** Apoia a identificação da linhagem de dados técnica dos principais sistemas de origem que darão suporte aos requisitos de negócios;
- **Business User** Tem uma visão clara das fontes de dados para os relatórios que pode precisar de criar e garante a adesão às políticas e usa dados para impulsionar o desempenho dos negócios.

O contexto do diagrama de DG inclui definições e metas, consistindo de entradas, atividades, entregas, fornecedores, participantes, clientes, técnicas, ferramentas e métricas, conforme é possível observar-se na figura 2.2.

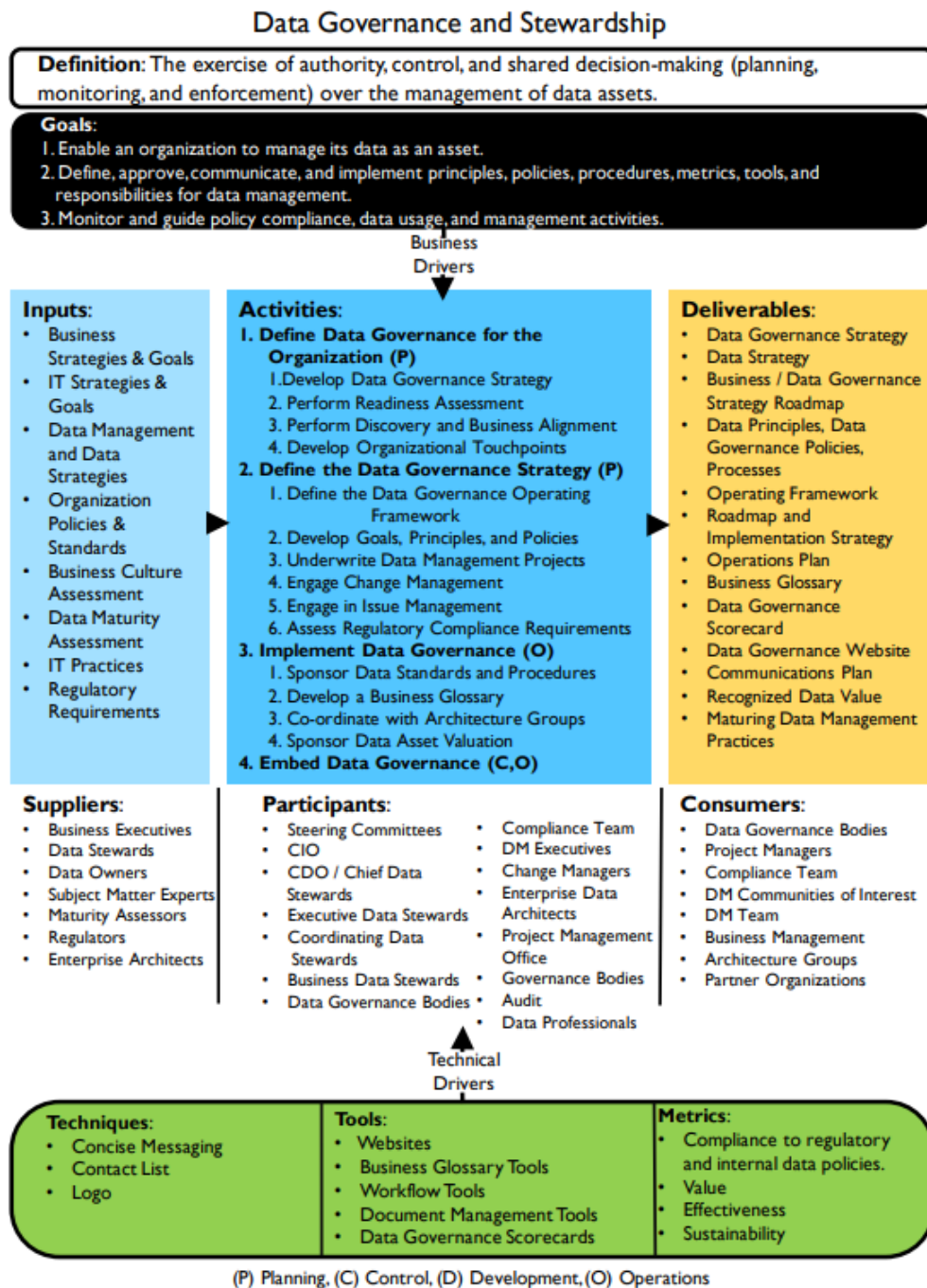


Figura 2.2: Context Diagram: Data Governance and Stewardship DAMA (2017).

Ao longo deste capítulo irão ser abordados conceitos que fazem parte do processo de DG e como é que esses conceitos se relacionam com o mesmo.

2.1.3.1 Data Stewardship

Data Stewardship (DSS) é o rótulo mais comum para descrever responsabilidade por dados e processos que garantem o controlo e o uso efetivos dos data ativos. O Steward pode ser formalizada por meio de cargos e descrições, ou pode ser uma função menos formal conduzida

por pessoas que tentam ajudar uma organização a obter valor de seus dados. Frequentemente, termos como Custodian ou Trustee são sinónimos para aqueles que desempenham funções semelhantes às do Steward.

Enquanto que DG se refere, de forma geral, à organização e gestão das informações dos dados e ativos em toda a organização através das pessoas, processos e tecnologias, como a preocupação em estabelecer qualidade de informação para otimizar uma melhor decisão, segundo (LexisNexis 2023) DSS está mais focada em DM, processos e atividades a fim de apoiar os objetivos de DG.

No que diz respeito ao princípio da organização e gestão dos dados e informações entre as pessoas, processos e tecnologias, é importante que uma boa implementação de DSS levará ao sucesso de um programa de DG em que o governo de dados geralmente se concentra em estabelecer em regras e políticas para garantir a eficácia e confiabilidade dos dados na organização com a ajuda da DSS para gerir as funções operacionais e táticas, responsabilidades e atividades que aderem e apoiam o plano de governo (Practices e Data 2023).

Sem DSS, DG é apenas uma framework. A figura 2.3 ilustra as partes envolvidas em DG e como DSS deve estar envolvido nesse processo para suportar todo o processo.

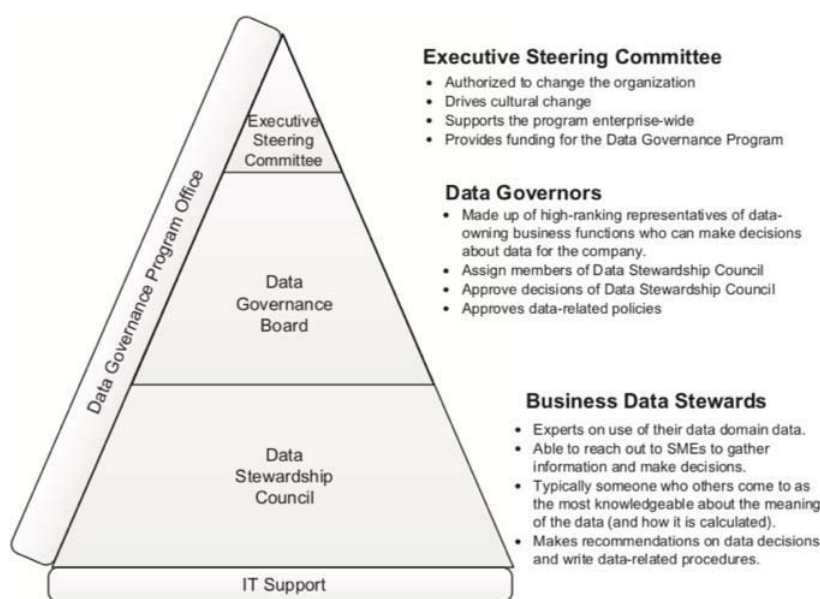


Figura 2.3: Data Governance Program (Wong, Maarop e Samy 2020).

Para desenvolver uma boa DSS, vários objetivos podem ser definidos de seguinte maneira (Wong, Maarop e Samy 2020):

- Um conselho DSS totalmente funcional;
- Assegurar o estabelecimento e cumprimento de procedimentos e políticas;
- Definir funções e responsabilidades claras para todos os Data Stewards;
- Definir Data Stewards comerciais e técnicos;
- A responsabilização da DM;
- Assegurar a formação de todas as pessoas envolvidas na gestão de dados;
- Para manter a qualidade dos dados;

- Garantir que processos e procedimentos escritos sejam aprovados e que podem ser usados.

2.1.3.2 Data Architecture

Data Architecture (DA) descreve data assets lógicos e físicos. Como parte da arquitetura corporativa, a DA define a estrutura geral e os dados relacionados dentro do organização, a DA é fundamental para a gestão (Maulina e Ruldeviyani 2019).

DA nem sempre são necessárias para os projetos. E saber quando e quando não usar é um aspeto importante e normalmente determina o quão bem sucedido um projeto poderá ser. Criar uma arquitetura adequada é geralmente uma função do quanto os arquitetos e designers entendem a organização e as suas necessidades de dados.

Uma arquitetura é muitas vezes necessária quando o projeto tem um ou mais dos seguintes requisitos (Tejada 2018):

- Armazenamento e processamento de dados em volumes muito grandes para bases de dados tradicionais;
- Transformação de dados não estruturados para análise e comunicação;
- Captura, processamento e análise ilimitada de fluxos de dados em tempo real ou com baixa latência.

As DA geralmente incluem alguns dos seguintes componentes, dependendo dos requisitos do projeto bem como a infraestrutura disponível (Agbo, Kupriyanov e Rytsarev 2020):

- Fontes de dados;
- Armazenamento de dados;
- Processamento em lote;
- Ingestão de mensagens em tempo real;
- Processamento de fluxo;
- Armazenamento de dados analíticos;
- Análise e relatórios;
- Automação (Orquestração).

Escolher a arquitetura certa para o projeto em causa tem um caminho longo para reduzir custos, fazer decisões mais rápidas e melhores e facilitar a previsão de necessidades futuras, bem como a criação de novos produtos e serviços. No entanto, segundo Agbo, Kupriyanov e Rytsarev (2020) isto vem com os seus seguintes desafios:

- Segurança: O maior desafio é a privacidade dos utilizadores, visto que os dados podem ser dados pessoais que podem ser identificados e ligados a utilizadores, levantando assim um elevado número de preocupações com a privacidade;
- Escalabilidade: As soluções devem ser projetadas de forma a permitir a fácil escalabilidade;
- Data Quality: Os dados são de alta qualidade se forem adequados para os seus usos pretendidos em operações, tomadas de decisões, análises e planeamentos. Isto significa

que os dados são consistentes, precisos, abrangentes, completos, ordenados, livre de defeitos, relevantes, fáceis de interpretar e auditáveis.

2.1.3.3 Data Modeling and Design

Data Modeling é a primeira etapa no processo de design de uma base de dados. Esta etapa às vezes é considerada uma fase de design abstrato e de alto nível, também chamada de projeto conceitual (WATT 2014).

Data Modeling é o processo de criação de uma representação visual de um sistema de informação inteiro ou de partes dele para comunicar conexões entre pontos de dados e estruturas.

O objetivo é ilustrar os tipos de dados usados e armazenados no sistema, as relações entre esses tipos de dados, as formas como os dados podem ser agrupados e organizados e os seus formatos e atributos (IBM 2023).

Os modelos de dados são construídos em torno das necessidades de negócio. Regras e requisitos são definidos antecipadamente por meio de feedback das partes interessadas do negócio, para que possam ser incorporados no design de um novo sistema ou adaptados na iteração de um existente (IBM 2023).

Existem vários esquemas diferentes usados para representar dados. Os seis esquemas mais comuns são: Relacional (SQL), Dimensional, Orientado a Objetos, Baseado em Fatos, Baseado em Tempo e NoSQL. Os modelos destes esquemas existem em três níveis de detalhe: conceitual, lógico e físico (DAMA 2017). Estes são discutidos em maior detalhe em baixo (IBM 2023):

- **Modelos de Dados Conceituais** Oferecem uma visão geral do que o sistema terá, como será organizado e que regras de negócio estão envolvidas. Os modelos conceituais geralmente são criados como parte do processo de coleção dos requisitos iniciais do projeto. Normalmente, eles incluem classes de entidade (definindo os tipos de coisas que são importantes para o negócio representar no modelo de dados), as suas características e restrições, os relacionamentos entre eles e os requisitos relevantes de segurança e integridade dos dados;
- **Modelos de Dados Lógicos** Estes são menos abstratos e fornecem mais detalhes sobre os conceitos e relações no domínio em consideração. Um dos vários sistemas formais de notação de Data Modeling é seguido. Estes indicam atributos de dados, como tipos de dados e os seus correspondentes comprimentos, e mostram as relações entre as entidades. Os modelos de dados lógicos não especificam nenhum requisito técnico do sistema e podem ser úteis em ambientes de implementação altamente processuais ou para projetos orientados a dados por natureza, como design de Data Warehouse;
- **Modelos de Dados Físicos** Estes fornecem um esquema de como os dados serão armazenados fisicamente numa base de dados. Como tal, eles são os menos abstratos de todos. Eles oferecem um design finalizado que pode ser implementado como uma base de dados relacional, incluindo tabelas associativas que ilustram os relacionamentos entre entidades, bem como as chaves primárias e estrangeiras que serão usadas para manter esses relacionamentos. Os modelos de dados físicos podem incluir propriedades específicas do sistema de gestão de base de dados (DBMS), incluindo ajustes de desempenho.

Data Modeling torna mais fácil para programadores, arquitetos de dados, analistas de negócio e outras partes interessadas de visualizar e entender as relações entre os dados em bases de dados ou Data Warehouses. Além disso, pode (WATT 2014):

- Reduzir erros no desenvolvimento de software e bases de dados;
- Aumentar a consistência na documentação e no design do sistema em toda a empresa;
- Melhorar o desempenho de aplicações e bases de dados;
- Facilitar o mapeamento de dados em toda a organização;
- Melhorar a comunicação entre programadores e equipas de inteligência de negócios;
- Facilitar e acelerar o processo de design de bases de dados nos níveis conceitual, lógico e físico.

2.1.3.4 Data Security

Data Security (DS) é o planeamento, desenvolvimento e execução de políticas e procedimentos de segurança para fornecer autenticação, autorização, acesso e auditoria de dados e informação.

De acordo com (Catro et al. 2021), os requisitos e procedimentos de DS são classificados em quatro grupos, conhecidos como os quatro A's:

- **Acesso** Gerir os privilégios e acessos dos utilizadores para que eles possam acessar aos sistemas em tempo útil. Um exemplo pode ser o acesso ao sistema de gestão de informações da empresa;
- **Autorização** Conceder os privilégios de acesso apropriados às exibições de dados específicas, dependendo das funções do utilizador;
- **Autenticação** Validar se os utilizadores são quem eles dizem ser no momento de acesso aos sistemas. Os sistemas mais comuns são baseados em um duplo contendo um nome de utilizador e uma palavra-passe;
- **Auditoria** Analisar as ações de segurança e a atividade do utilizador para garantir a conformidade com os regulamentos e com as políticas e normas da empresa.

Políticas e procedimentos eficazes de segurança de dados garantem que as pessoas certas possam usar e atualizar os dados da maneira correta, e que todo o acesso e atualização inapropriados são restritos (Ver figura 2.4).

Compreender e cumprir com os interesses e necessidades da privacidade e confidencialidade de todas as partes interessadas é do melhor interesse de cada organização. Os relacionamentos com clientes, fornecedores e constituintes confiam e dependem do uso responsável De dados (Ray 2012).

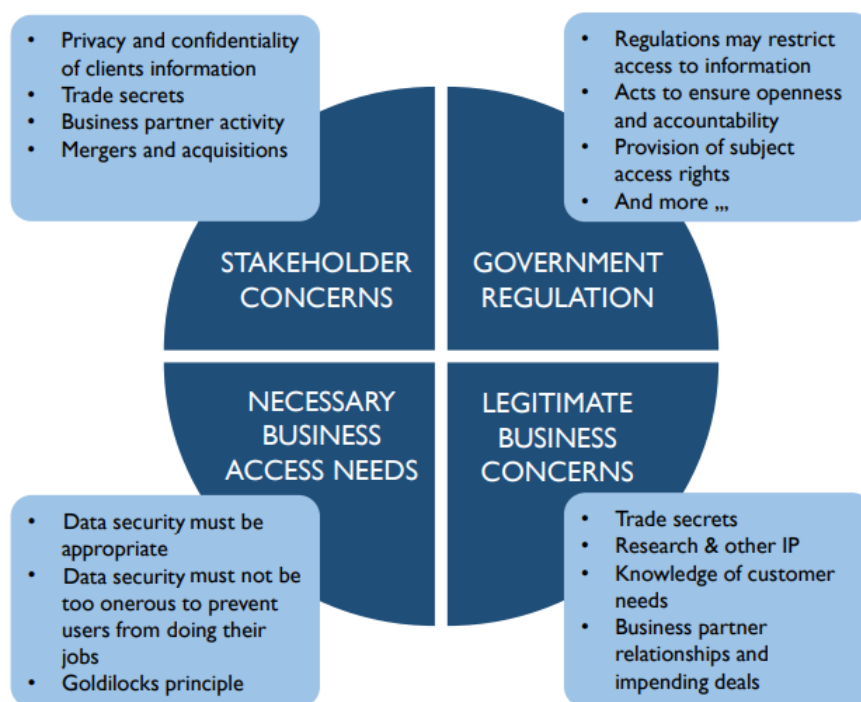


Figura 2.4: Sources of Data Security Requirements (Ray 2012).

A implementação de práticas de DS depende da cultura corporativa, da natureza dos riscos, da sensibilidade dos dados a empresa gere e os tipos de sistemas em vigor. Os componentes do sistema de implementação devem ser guiados por um plano estratégico de segurança e arquitetura de suporte.

2.1.3.5 Data Warehouse/Data Lake

Data Lake é um novo conceito, usado para descrever a forma evolutiva dos repositórios de dados. É uma framework de pensamento, bem como uma solução técnica e define como receber e entregar dados, como separá-los em diferentes repositórios com base nas suas características especiais e como protegê-las.

Além disso, o Data Lake também define como aceder aos dados para relatórios ou exploração e modelagem analítica, e as ferramentas e a maneira de os implementar (Paschalidi 2015).

O conceito do Data Lake trata-se de dados serão entregues uma vez, evitando a recriação de cópias entre departamentos e divisões dentro de uma empresa. Uma combinação de tempo real e entrega em lote de dados existe, um suporte para dados não estruturados é fornecido ao usar novas tecnologias e novas tecnologias permitem trazer a análise para os dados em vez de trazer os dados para a análise (Paschalidi 2015).

Data Warehouse é uma coleção de dados orientada a objetos, integrada, variável no tempo e não atualizável, que é usada para apoiar o processo de decisão e negócios inteligentes.

É orientada a objetos pois é organizada em torno de assuntos chave da empresa, nomeadamente clientes, vendas, compras, entre outros e fornece uma visão simples e concisa sobre assuntos específicos, excluindo dados não úteis no processo de apoio à decisão.

É integrada pois é construída integrando-se várias fontes de dados heterogêneas onde são aplicadas técnicas de limpeza e transformação de dados.

É variável no tempo pois consegue marcar o tempo para mostrar o momento no tempo em que o registro é/era preciso.

Por último é não atualizável, pois o armazenamento físico dos dados transformados está separado do ambiente operacional (Oliveira 2021).

Existem três esquemas populares de Data Warehouse, estrela, floco de neve e constelação/-galáxia (Wamars e Randriatoamanana 2016).

- **Estrela** Esquema de Data Warehouse que consiste numa tabela de fatos e zero ou mais tabelas de dimensão e sem tabelas de sub-dimensão;
- **Floco de Neve** Esquema de Data Warehouse que consiste de uma tabela de fatos e uma ou mais tabelas de dimensão com uma ou mais tabelas de sub-dimensões;
- **Constelação ou Galáxia** Esquema de Data Warehouse onde pelo menos duas tabelas de fatos compartilham uma ou mais tabelas de dimensão e/ou sub-dimensão.

ETL é um processo que consiste em três atividades, Extração, Transformação e Carregamento (Wamars e Randriatoamanana 2016).

- **Extração** Processo para identificar a fonte de dados internos ou externos e definir os procedimentos de extração;
- **Transformação** Processo que inclui sub-atividades como limpar, integrar e consolidar os dados e conversão e manipulação de dados;
- **Carregamento** Processo para carregar dados de OLTP ², para tabelas de Data Warehouse.

As principais diferenças entre os dados tradicionais Data Warehouse e o Data Lake são descritos na tabela 2.1:

²OLTP ou Online Transaction Processing é um tipo de processamento de dados que consiste na execução de uma série de transações que ocorrem simultaneamente, como por exemplo, banco, compras, entrada de pedidos ou envio de mensagens de texto online.

Tabela 2.1: Data Warehouse versus Data Lake (O'Leary 2014)

Caraterísticas	Data Warehouse	Data lake
Complemento ou integrado	Complemento caro	Parte da computação empresarial
Comunidades de utilizadores que utilizam os dados	Um	Várias
Número de fontes de dados	Um (Tipicamente)	Múltiplas
Tipo de perguntas	Roll-up	Várias
Perguntas query conhecidas com antecedência	Sim	Não
Esquema de dados otimizado	Estrela, Floco de Neve	Não
Foco de dados	Tempo de design/armazenamento	Tempo de uso/pesquisa
Dados estruturados ou não estruturados	Estruturados	Não estruturados, semi-estruturados e estruturados
Nível dos dados	Evento	Evento e sub-eventos
Pontualidade dos dados	Intervalo de tempo associado a extrações, transferências e cargas (ETL)	Tempo real

2.1.3.6 Data Quality

No início da década de 1990, os cientistas de computadores começaram a deparar-se com um problema, definir, medir e melhorar a qualidade dos dados eletrónicos armazenados em bases de dados, data warehouses e sistemas legacy (Batini e Scannapieco 2006). Até este ponto, Data Quality (DQ) era apenas falada por estatísticos e investigadores em gestão.

De acordo com Catro et al. (2021), DQ são atividades de planeamento, implementação e controlo que aplicam técnicas de gestão da qualidade para medir, avaliar, melhorar e garantir a aptidão dos dados para a sua utilização.

Um programa de DQ é mais eficiente quando faz parte de um programa de DG, tendo em consideração também que muitas vezes empresas estabelecem DG devido a problemas de DQ. Deste modo, segundo Loshin (2001), uma organização de governação pode acelerar o trabalho de um programa de DQ através de:

- Definindo prioridades;
- Identificando e coordenando o acesso àqueles que devem estar envolvidos em várias decisões e atividades relacionadas com a qualidade dos dados;
- Desenvolvendo e mantendo padrões de DQ;
- Relatando medições relevantes da DQ de toda a empresa;
- Fornecendo orientação que facilite o envolvimento dos trabalhadores;
- Estabelecendo mecanismos de comunicação para a partilha de conhecimentos;
- Desenvolvendo e aplicando políticas de conformidade e DQ;
- Monitorizando e relatando sobre o desempenho;

- Partilhando os resultados da inspeção de DQ para aumentar a consciência, identificar oportunidades para melhorar e construir consenso para melhorias;
- Resolvendo variações e conflitos; fornecendo direção;

2.1.3.7 Data Lineage

Data Lineage (DL) (ou Data Provenance) fornece informações históricas de registos de dados e as suas origens.

DL gerado por transformações complexas é de elevado valor para diferentes aplicações, por exemplo DL com gestão de metadados, data debugging para rastrear o motivo dos erros dos dados, data trust para rapidamente responder a oportunidades de negócio e a desafios regulatórios e data security para colocar controlos nos acessos a query de dados baseados na DL (Glavic 2014).

Segundo Dehghani (2022) DL é entender de onde é que os dados originam e que por transformações é que passaram. Permanece um elemento de compreensão, pois confiar nos dados é necessários para depurar problemas com os dados. Isto é particularmente útil em Machine Learning onde a qualidade do modelo depende em grande parte da qualidade dos dados e do treino.

Embora o uso de abordagens de DL seja uma forma promissora para a gestão de big data, o processo é bastante complexo. Os desafios incluem a escalabilidade do armazenamento da linhagem, tolerância a falhas no armazenamento da linhagem, captura precisa da linhagem para operações black box, entre outros, (Tang et al. 2019).

Estes desafios devem ser abordados com cuidado e os trade-offs entre eles precisam de ser devidamente avaliados para se conseguir produzir um design realista para a captura de DL.

2.2 Plataformas de Data Governance

Como discutido no capítulo 1, a empresa que propôs este projeto possui uma implementação de DG incompleta e arcaica e como tal essa implementação passará a ser gerada através de uma aplicação.

2.2.1 Collibra

De acordo com Billigence (2023), Collibra software é uma plataforma de DG voltada para empresas, para DM e stewardship. Collibra capacita as empresas a encontrar significado nos seus dados e melhorar as decisões de negócio.

Com uma plataforma compartilhada, utilizadores de negócios e de IT podem colaborar para formar uma cultura orientada por dados utilizando o conjunto de produtos da Collibra.

Collibra permite às empresas que, a partir da sua aplicação, consigam atuar nas seguintes áreas:

- **Data Quality & Observability:** Monitorizando a qualidade dos dados e a confiabilidade de data pipelines para ajudar a corrigir anomalias rapidamente utilizando regras de validação automáticas;

- **Data Catalog:** Fornecendo a possibilidade de se criar um catálogo de dados que permite uma maior facilidade em se encontrar e perceber os dados oriundos de várias fontes;
- **Data Governance:** Oferece a possibilidade de existir um único local para encontrar, entender e criar uma linguagem partilhada em torno dos dados para todos os utilizadores. Possível também a operacionalização de workflows e processos de DG para fornecer dados confiáveis;
- **Data Lineage:** Permite que os utilizadores mapeiem relações entre sistemas, aplicações e relatórios para fornecer uma visão rica em contexto de toda a empresa, aprimorando qualquer dados relevantes;
- **Proteção:** Minimizando os riscos e acelerando a governação dos acessos aos dados descobrindo, definindo e protegendo dados de uma plataforma unificada;
- **Data Privacy:** Possível dar aos dados privacidade através da centralização, automação e orientação dos workflows que incentivam a colaboração, operacionalizam a privacidade e atendem aos requisitos regulatórios globais.

2.2.2 Alation

De acordo com Datameer (2023), Alation é uma ferramenta de data catalog que ajuda organizações identificar, entender e administrar os seus data assets. Possibilita as organizações a catalogar os data assets, capturar e adicionar informações de nível técnico e de negócio sobre eles, administrar e governar estes ativos.

Também ajuda os stakeholder a perceber que data assets é que existem, do que são feitos, como estão sendo utilizados e a gerir os riscos, privacidade e conformidade dos mesmos.

Alation fornece as seguintes características aos utilizadores:

- **Policy Center:** As políticas são convenientemente centralizadas aqui de modo a não criarem confusão, permite a criação de políticas e fornece uma vista global de como é que as políticas estão mapeadas para os dados;
- **Governance Workflow:** Permite aos utilizadores a criação e configuração de governance workflows através da aplicação de DG da Alation, sem o uso de programação ou da invocação de outras aplicações, garantindo também a transparência e auditabilidade dos processos de aprovação ao longo do catálogo de dados;
- **Stewardship Workbench:** Utilização de Inteligência Artificial e Machine Learning para automatizar a descoberto de potenciais data stewards baseado em quem é que realmente está a usar os dados;
- **Governance Dashboards:** Os líderes podem medir ativamente o progresso de tarefas abertas, uso de políticas, e muito mais, visualizando apenas o dashboard.

2.2.3 SAP Master Data Governance

Segundo SAP (2023) o SAP Master Data Governance, é uma solução nativa da nuvem que facilita o início rápido de uma iniciativa de DM usando uma abordagem modular de expansão sob demanda.

SAP Master fornece as seguintes características aos utilizadores:

- **Consolidação dos Dados:** Envolve a criação dos melhores registos com mapeamento de chaves entre duplicados e consolidação com gestão central para dados mestres de alta qualidade;
- **Data Governance Centralizada:** Fornece um processo de negócios centralizado que ajuda na gestão de dados mestres. A consolidação e o controlo centralizado dos dados ajudam a reduzir os custos gerais;
- **Processamento em Massa:** Vários registos de dados mestre podem ser atualizados ao mesmo tempo. Deste modo, é um processo de mudança em massa altamente eficaz e alterações em massa podem ser feitas em dados de parceiros de negócios, clientes, fornecedores e materiais;
- **Modelos de dados pré-construídos, workflows e interfaces de utilizador:** Fornece modelos de dados padrão pré-construídos, interfaces de utilizador, regras e definições de workflows para dados mestres financeiros, parceiros de negócios, dados mestres de materiais, clientes e fornecedores.

2.2.4 Comparação

Foi feita uma avaliação às características Free trial, documentação da plataforma, user friendly e potencialidades, efetuando assim uma avaliação a cada uma delas.

Tabela 2.2: Comparação entre as várias Plataformas de Data Governance

	Collibra	Alation	SAP Master Data Gov.
Free Trial	Sim	Sim	Sim
Documentação	Boa	Boa	Boa
User Friendly	4/5	4/5	3/5
Potencialidades	4/5	2/5	4/5

As três opções possuem resultados bastante semelhantes, no entanto é de notar que a plataforma Collibra, de forma geral, possui uma melhor avaliação que a sua concorrência, visto que, possuindo o free trial e a documentação igual às outras plataformas, porém possui 4/5 quer em user friendly, quer em potencialidade, enquanto que as outras opções possuem 4/5, 2/5 e 3/5, 4/5.

Capítulo 3

Análise

O objetivo deste capítulo é apresentar e descrever o ambiente da solução, com o auxílio ao modelo de domínio, a análise de requisitos funcionais e não-funcionais ligados aos casos de uso escolhidos para a prova de conceito a desenvolver, tal como apresentar alguns diagramas para o auxílio na compreensão da solução implementada.

3.1 Ambiente Collibra

O ambiente Collibra é composto por comunidades, domínios, ativos, relações entre os mesmos e vistas, sendo que tudo se encontra conectado. Diferentes domínios representam diferentes ativos e/ou áreas de negócio, como por exemplo do domínio "new data sets", possui apenas ativos do tipo data set, enquanto que um domínio "common systems" possui ativos do tipo sistema, estes domínios podem ou não encontrar-se dentro do mesmo sistema. As vistas são locais onde podem ou não existir múltiplos tipos de ativos para se poder visualizar, no entanto, uma vista pode simplesmente ser a "página" de um domínio onde se consegue observar os ativos que se encontram dentro do mesmo.

Quando um utilizador entra no ambiente Collibra, a primeira vista é a "Landing Page", sendo que é a partir desta que os utilizadores conseguem navegar para as comunidades, domínios ou vistas necessárias para efetuar os mesmo trabalhos diários.

Para auxiliarem nos trabalhos diários dos utilizadores, existem também workflows, desenvolvidos especialmente para um propósito, qualquer que o mesmo seja. Estes workflows podem possuir diversos "triggers", nomeadamente criação, modificação ou remoção de ativos, criação ou remoção de domínios ou comunidades, no entanto, o mais comum e que é visível para os utilizadores é o simples botão, que se encontra na vista do ativo ou na vista do domínio caso se tenha selecionado um ativo. Imagens que representam todos estes cenários são apresentadas ao longo deste e do próximo capítulo.

Tal como referido anteriormente, tudo está conectado, e os workflows não são exceção. Os workflows de ativação, quer normal ou em massa, ativam utilizadores que posteriormente, no âmbito das DGTs, vão ser esses utilizadores ativos que as vão catalogar, e é através das DGTs que alguns tipos de ativos são catalogados no ambiente Collibra, com o uso do workflow "End DGT".

Os workflows de "Lineage Path" e "Data Set Management" necessitam da linhagem presente no Collibra, através de diversas relações entre os ativos, para executarem corretamente, uma vez que o "Lineage Path" necessita das relações entre data set e coluna e coluna com mais colunas, e o "Data Set Management" necessita do "caminho" criado pelas relações para obter

o sistema a partir da tabela, uma vez que apenas o sistema que origina do *Data Lake* do cliente é que é utilizado neste workflow.

Na figura 3.1 é apresentado os diversos tipos de ativos existentes no Collibra e as relações criadas entre os mesmos. Grande parte destes tipos de ativos e relações não entram no âmbito desta dissertação, devendo-se dar uma maior ênfase aos três blocos centrais, denominados de *Technology Assets*, *Physical Data Dictionary* e *Data Usage Registry*, que é onde se encontram os tipos de ativos mais utilizando no âmbito desta dissertação.

Os nomes destes três blocos centrais, são tipos de domínios existentes em Collibra. Cada ativo, para ser criado, tem de ser criado num domínio, que por sua vez, um domínio para ser criado tem de ser numa comunidade.

Deste modo, domínios são agrupamentos lógicos de ativos, de acordo com as suas funções, projetos ou áreas de conhecimento, têm um tipo de domínios, nomeadamente, *Technology Assets*, *Physical Data Dictionary* e *Data Usage Registry* por exemplo, e têm de ter um nome único dentro da mesma comunidade.

Uma comunidade pode ser um grupo de comunidades e domínios com nomes únicos, sendo que cada comunidade pode contar qualquer número de sub-comunidades, nesta solução as comunidades correspondem a divisões funcionais do cliente.

É possível observar-se a linhagem entre os tipos de ativos, nomeadamente a linhagem do sistema -> base de dados -> schema -> table -> data set -> coluna, que é a linhagem mais usada, pois permite ao utilizador observar de onde é que a tabela, coluna ou data set que o mesmo está a visualizar originou e as relações entre data set e coluna e coluna com coluna, são as relações utilizadas no workflow do *Lineage Path* apresentado mais à frente na subsecção 5.2.2.

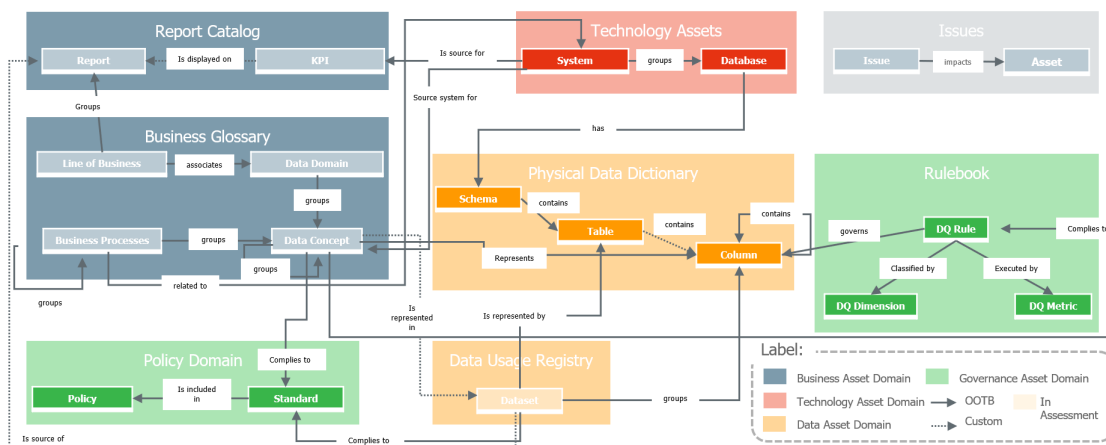


Figura 3.1: Tipos de Ativos e as suas Relações

Uma relação é uma instância de exatamente um tipo de relação, que especifica as direções dessa relação, um ativo pode ter várias ocorrências do mesmo tipo de relação. Para se entender melhor o que é uma relação entre ativos e como é que a mesma é visível em Collibra, usa-se o seguinte exemplo:

Tabela 3.1: Demonstração de Relação

Tipos Ativo	Ativos	Tipo de Relação	Relação
Data Concept	Portugal	Data Concept has a Code Value	Portugal has code PT
Code value	PT	Code Value is code for Data Concept	PT is code for Portugal

3.1.1 Data Gathering Tool (DGT)

Data Gathering Tool ou DGT, conforme discutido no capítulo 1 é um processo que o cliente possui para centralizar toda a informação que possui sobre novos ativos. O processo criado pelos mesmos é efetuado na ferramenta Excel, e o objetivo deste requisito é substituir esse método, utilizando a ferramenta Colibra para migrar o processo das DGTs para a plataforma Colibra ¹.

A DGT detalha, funcionalmente e tecnicamente, os dados que suportam as necessidades do negócio e identifica e contextualiza os dados através da disponibilização de metadados a consolidar no catálogo de dados corporativos. Uma DGT é a saída do pedido dos utilizadores de negócios com todos os conjuntos de dados e as suas informações que são criadas no *Data Lake*.

Dada a importância da DGT e para que esta seja devidamente aproveitada, existe um processo que garante a sua utilização efetiva.

Uma DGT é composta por 4 fases (*Business Understanding*, *Data Understanding*, *Data Modeling and Validation*), sendo que cada fase tem várias secções.

A fase **Business Understanding** possui as secções *Functional (Data Model)*, *Functional (Data Model Detail)*, *Functional (Queries)* e *Data Access*. Nesta secção é definido as especificações funcionais de cada conjunto de dados e campo.

A fase **Data Understanding** possui as secções *Sources (Data Model)*, *Sources (Data Model Detail)*, *Sources (Data Model Design)*, *Sources (Queries)* *Sources Auxiliary Tables* e *Sources (Samples)*. Nesta secção é definido as especificações técnicas do sistema de origem, amostras e queries.

A fase **Data Modeling** possui as secções *Destination (Data Model)*, *Destination (Data Model Detail)* e *Destination (Data Model Design)*. Nesta secção é definido as especificações técnicas do modelo de dados final.

Por último a fase **Validation** possui as secções *Quality Guidelines*, *Quality Controls (catalogue)*, *Quality Controls (Outputs) Acceptance* e *Tamr*. Nesta secção é definido o catálogo de controlos antes da ingestão, que ocorre até ao início dos testes de aceitação após a fase de ingestão de dados.

A fase **Business Understanding** é catalogada por *Business Users* e *Data Stewards*, a fase **Data Understanding** é catalogada por *Data Custodian*, as fases **Data Modeling** e **Validation** são catalogadas por *Data Officers*.

Entre as diversas fases, existem secções com nomes semelhantes, nomeadamente fases que possuam *Data Model*, *Data Model Detail* e *Queries* no nome.

¹Por motivos de confidencialidade, não é possível apresentar o processo em Excel do cliente, no entanto, todos os passos existentes para o funcionamento correto do processo serão apresentados ao longo desta secção.

As secções *Functional (Data Model)*, *Functional (Data Model Detail)*, *Sources (Data Model)*, *Sources (Data Model Detail)*, *Destination (Data Model)* e *Destination (Data Model Detail)* retratam os mesmos ativos, isto é, os dados a serem catalogados são os mesmos, sendo que o que se altera é quem é o responsável por as catalogar, tal como discutido anteriormente. É com esta situação em mente que se implementou os requisitos retratados na subsecção 6.1.6.8.

A fase **Validation** é a mais distinta dentro das quatro, pois esta é referente a *Data Quality*, uma vez que é feita a validação das três fases anteriores, tal como se os campos estão corretamente preenchidos, se existe coerência e coesão entre fases, entre outras. De modo, a validar-se se toda a catalogação ² está a ser feita de forma correta e seguindo as boas práticas do processo.

Quando o processo de catalogação de uma DGT termina, começa o processo de criação e/ou catalogação dos ativos que são retratados ao longo da DGT. Isto é, ao longo de uma DGT são retratados ativos do tipo data concept, tabela, coluna e data set, os nomes e atributos dos mesmo são campos que são preenchidos pelos utilizadores ao longo do desenvolvimento da DGT e no final, todos estes campos catalogados têm de ser utilizados para popular o ambiente Collibra.

Como tal, o processo do *End DGT*, referido anteriormente por "processo de criação e/ou catalogação dos ativos", verifica se, os data concepts, tabelas, colunas e data sets já existem no ambiente Collibra e se já se encontram catalogados.

Caso já existam e todos os atributos catalogados na DGT já se encontram catalogados no ativo, então nada é realizado, no entanto, caso existam atributos não catalogados nos ativos, os mesmo são catalogados com as informações oriundas da DGT e caso não existam os ativos referenciados na DGT (A verificação é feita através do nome, uma vez que não é possível existirem ativos do mesmo tipo com o nome igual), os mesmos são criados, com a exceção dos data sets, que como referidos na subsecção 5.2.4 são apenas criados através do workflow *Data Set Management*.

No caso da figura 3.2, a DGT apresentada já se encontra finalizada, pois todas as secções encontram-se fechadas e o "Status" encontra-se em "CV", que significa "Completed Validation" ou "Validação Completa".

²A catalogação é o processo de descoberta e compreensão de dados que oferece ao cliente uma melhor compreensão e visibilidade dos dados e a sua estrutura.

DGT View ▾

> Delete Move Validate

☐	Name ↑	Proceed to details	Status
☐	010_ALLGRO_AUG_PART_01	010_ALLGRO_AUG_PART_01	CV
☐	0. Register	010_ALLGRO_AUG_PART_01	-
☐	1. Business Understandi...	Business Understanding	Closed
☐	1.1 Functional (Data ...	Functional (Data Model)	Closed
☐	1.2 Functional (Data ...	Functional (Data Model Detail)	Closed
☐	1.3 Functional (Queri...	Functional (Queries)	Closed
☐	1.4 Data Access	Data Access	Closed
☐	2. Data Understanding	Data Understanding	Closed
☐	2.1 Sources (Data M...	Sources (Data Model)	Closed
☐	2.2 Sources (Data M...	Sources (Data Model Design)	Closed
☐	2.3 Sources (Data M...	Sources (Data Model Detail)	Closed
☐	2.4 Sources (Samples)	Sources (Samples)	Closed
☐	2.5 Sources (Queries)	Sources (Queries)	Closed
☐	2.6 Source Auxiliary ...	Source Auxiliary Tables	Closed
☐	3. Data Modelling	Data Modelling	Closed
☐	3.1 Destination (Dat...	Destination (Data Model)	Closed
☐	3.2 Destination (Dat...	Destination (Data Model Detail)	Closed
☐	3.3 Destination (Dat...	Destination (Data Model Desig...	Closed
☐	4. Data Validation	Data Validation	Closed
☐	4.1 Quality Guidelines	Quality Guidelines	Closed
☐	4.2 Quality Controls ...	Quality Controls (Catalogue)	Closed
☐	4.3 Quality Controls ...	Quality Controls (Outputs)	Closed
☐	4.4 Acceptance	Acceptance	Closed
☐	4.5 Tamr	Tamr	Closed

Figura 3.2: Estrutura da DGT

3.2 Modelo de Domínio

Foi utilizado o modelo de domínio, representado na figura 3.3 para se representar as diversas partes que compõem a plataforma.

Deste modo, temos o utilizador, que pode manipular várias Data Gathering Tools (DGTs) e várias DGTs podem ser manipuladas por vários utilizadores; O mesmo pode também executar vários workflows e os mesmos podem ser executados por vários utilizadores; Múltiplos

utilizadores podem interagir com diversas comunidades, domínios e ativos, que por sua vez, os mesmos podem ser interagidos por múltiplos utilizadores.

As DGTs possuem múltiplos ativos, domínios e comunidades, no entanto um ativo, um domínio e uma comunidade apenas se pode encontrar numa DGT.

Múltiplos workflows podem manipular, isto é, criar, modificar ou eliminar, diversos ativos, domínios e comunidades e os mesmos podem ser manipulados por diferentes workflows.

Uma comunidade pode englobar vários domínios, tal como um domínio pode englobar vários ativos, no entanto um domínio apenas se pode encontrar numa comunidade, tal como um ativo apenas se pode encontrar num domínio.

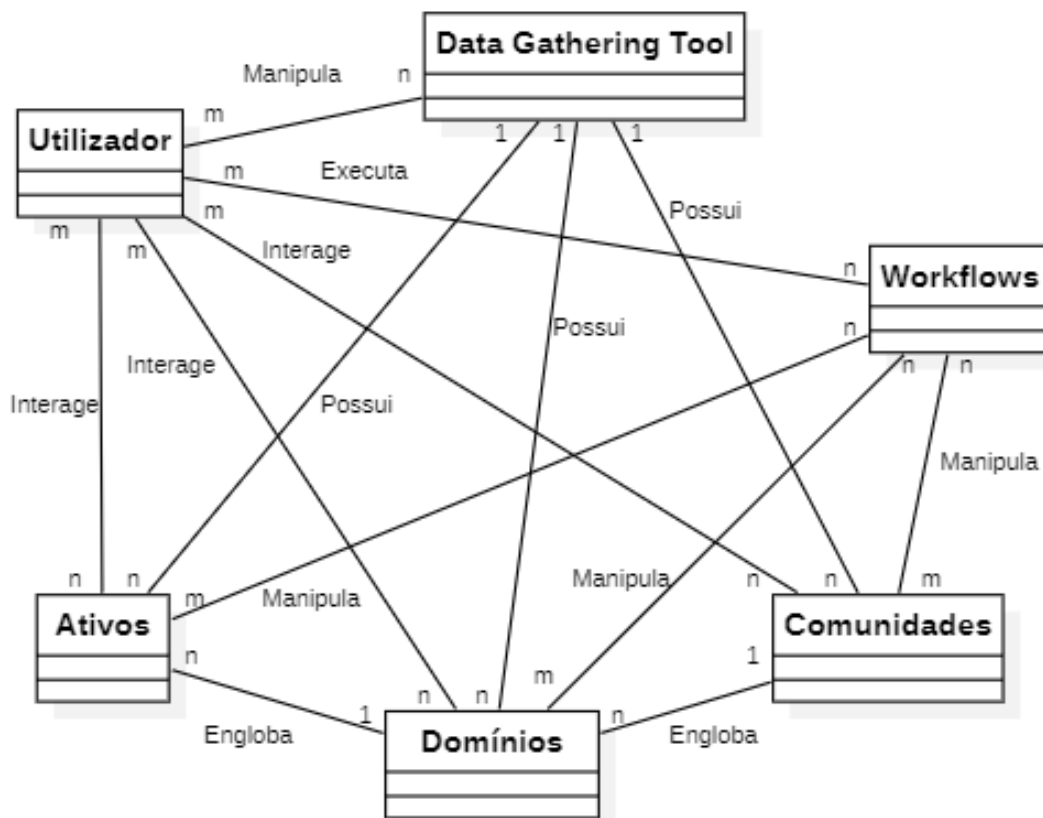


Figura 3.3: Modelo de Domínio

3.3 Engenharia de Requisitos

A engenharia de requisitos é uma disciplina que engloba todas as tarefas e técnicas que resultam no entendimento dos requisitos de um sistema. Identifica os objetivos do sistema a desenvolver e a operacionalização de tais objetivos em serviços e restrições (Lamsweerde 2000).

É o primeiro processo no desenvolvimento de software e é de extrema importância para perceber o que o cliente pretende, analisando as suas necessidades de modo a averiguar a viabilidade, negociar uma solução, especificá-la e, por fim, validá-la. É composto por 7 fases distintas: arranque, elicitação, elaboração, negociação, especificação, validação e gestão (Lamsweerde 2000).

3.3.1 Atores

Como referido no capítulo 2 os atores aqui presentes correspondem a posições dentro do que é uma estrutura normal dentro de um departamento que trabalha com dados de uma empresa.

Os atores encontram-se ordenados pela hierarquia dentro do alcance do que são as suas responsabilidades e permissões dentro do Collibra. Todos estes atores têm acesso aos ativos existentes no Collibra, no entanto as ações que os mesmos podem efetuar variam.

Administrador - Utilizador com maior responsabilidade uma vez que este utilizador tem acesso a tudo que existe no Collibra, aplicado aos programadores.

Data Officer - Utilizador responsável por ativar *Data Owner*, *Stewards* e *Custodians*. Dentro do processo de uma DGT é o responsável pela mesma, tendo acesso a todas as fases e secções.

Data Owner - Utilizador responsável por um ou mais ativos do tipo data domain.

Data Custodian - Utilizador responsável por um ou mais ativos do tipo sistemas. Dentro do processo de uma DGT terá acesso aos campos que necessita de preencher e aos que foram preenchidos pelo *Business User* e pelo *Data Steward*.

Data Steward - Utilizador responsável por um ou mais ativos do tipo data sets. Dentro do processo de uma DGT terá acesso aos campos que necessita de preencher e aos que foram preenchidos pelo *Business User*.

Business User - Utilizador com menor responsabilidade. Terá acesso apenas aos campos que necessita de preencher dentro do processo de uma DGT.

CoE DAAI - Utilizador do grupo responsável pelo Data Lake do cliente.

Activation Manager - Utilizador responsável pela ativação de *Data Owner*, *Stewards* e *Custodians*. É um elemento pertencente ao grupo de *Data Officer*.

3.3.2 Requisitos

Um requisito é uma condição ou capacidade possuída pelo software que tenta resolver um problema do mundo real. Estes problemas podem relacionar-se com a necessidade de automatizar partes de um sistema, corrigir deficiências ou implementar novas funcionalidades. IEEE define um requisito como (Thakur 2013):

- Capacidade ou condição necessária para um utilizador resolver um problema ou atingir um determinado objetivo;
- Capacidade ou condição que deve ser atendida ou possuída por um sistema para satisfazer um determinado contrato, padrão, especificação ou qualquer outro documento imposto;
- Uma representação documentada de uma condição ou capacidade referida nos dois pontos anteriores.

Os requisitos descrevem como um sistema se deve comportar e como tal diferem de um cliente para outro bem como de um processo de negócio para outro (Thakur 2013).

R.1 Como *Data Officer*, eu quero gerir todos os processos de Data Stewardship numa plataforma, mais concretamente o processo de ativação do *Data Owner*.

R.2 Como *Data Officer*, eu quero gerir todos os processos de Data Stewardship numa plataforma, mais concretamente o processo de ativação do Data Steward.

R.3 Como *Data Officer*, eu quero gerir todos os processos de Data Stewardship numa plataforma, mais concretamente o processo de ativação do Data Custodian.

R.4 Como Product Owner, quero ver as características dos ativos refletidas no Lineage Path, para que o Colibra seja sustentável e autossuficiente.

R.5 Como Product Owner, quero monitorizar as operações regulares da plataforma, fornecendo logs de execução para diferentes funções, para que os utilizadores não dependam da consola do Colibra.

R.6 Como *Data Officer*, quero poder carregar em massa *Data Owner*, *Stewards* e *Custodians* no Colibra.

R.7 Como *Data Officer*, quero integrar o Colibra com o data lake do cliente para a utilização dos seus dados.

R.8 Como Product Owner, quero customizar a landing page do Colibra para que ela reflita as informações mais críticas que os diferentes utilizadores necessitam de saber quando fazem login.

R.9 Como *Data Officer*, quero que seja criado um data set para cada tabela que seja ingerida do *Data Lake*.

R.10 Como *Data Officer*, pretendo integrar o Colibra com a ferramenta de Master Data Management do cliente para que as regras de qualidade e os *Golden Records* ³ se reflitam na framework de governo.

R.11 Como *Data Officer*, pretendo a criação de um dashboard para mostrar o uso dos workflows, mostrar as instâncias dos mesmos sem atendimento e acompanhar que personalizações é que estão a igualar as expectativas dos utilizadores.

R.12 Como *Data Officer*, pretendo automatizar o processo de limpeza de acessos obsoletos aos datasets do data lake em caso de saídas de utilizadores da empresa.

R.13 Como *Data Officer*, pretendo automatizar o Colibra como fonte para o Power BI dashboard, para que não seja necessário intervenção manual para atualizar as informações.

R.14 Como *Data Officer*, pretendo ter o processo Data Domain Activation no Colibra automatizado, para que todos os passos sejam realizados na plataforma.

R.15 Como *Data Officer*, quero ter acesso a um relatório detalhado sobre o estado da plataforma, para poder acompanhar as práticas de governo do cliente.

R.16 Como *Data Officer*, pretendo a criação de um repositório central para a criação e desenvolvimento de DGTs.

R.17 Como *Data Officer* quero ter a possibilidade de adicionar utilizadores ao desenvolvimento da DGT no Colibra.

³São registos ou dados de alta qualidade, confiáveis e considerados como fontes autoritativas e precisas, sugere a ideia de algo valioso, confiável e de referência.

R.18 Como *Data Officer*, pretendo o desenvolvimento de workflows para o auxílio na criação e desenvolvimento de DGTs.

3.3.3 Requisitos funcionais

Requisitos funcionais capturam o funcionamento esperado do sistema. Descrevem serviços, tarefas ou funções que o sistema deve executar (Malan e al 2001).

Dos requisitos listados anteriormente estes foram os que se consideraram como requisitos funcionais, **R.1, R.2, R.3, R.6, R.13, R.16 e R.18**.

3.3.4 Requisitos não funcionais

A apresentação dos requisitos não funcionais a seguir será conduzida com base na metodologia FURPS+, que é amplamente reconhecida na área de engenharia de software.

FURPS+ é uma técnica para validar os requisitos priorizados após uma compreensão das necessidades e exigências do cliente. A sigla FURPS representa Funcionalidade, Usabilidade, Confiabilidade, Desempenho e Suportabilidade. Com o passar do tempo e a necessidade crescente de ver a solução a partir de diferentes perspectivas, surgiu o FURPS+ (Hyderabad 2004).

Estes requisitos não funcionais têm um papel crítico durante o desenvolvimento do sistema pois servem como critérios de seleção na escolha de tecnologias ou arquiteturas (Chung 2012).

Funcionalidade - Representa as principais funcionalidades do produto que se enquadram no domínio do problema. Algumas das preocupações que se encaixam nesta categoria são, auditoria, localização, segurança, e-mail, entre outras (Coepd 2014). Dentro do contexto deste projeto foram identificados os requisitos **R.4, R.12, R.14 e R.17** como requisitos de funcionalidade.

Usabilidade - Consiste em observar, capturar e declarar requisitos com base em questões da interface gráfica do utilizador. Estas questões podem remeter para a acessibilidade, estética e consistência da interface gráfica (Coepd 2014). No contexto deste projeto foram identificados os requisitos **R.8, R.11 e R.15** como requisitos de usabilidade.

Confiabilidade - Refere-se à integridade, conformidade e interoperabilidade do software. Alguns dos requisitos que podem ser considerados são: frequência e gravidade da falha; tempo médio entre falhas; possibilidade de recuperação de dados; entre outros (Carlos Vazquez 2016). No contexto deste projeto foi identificado o requisito **R.9** como requisito de confiabilidade.

Desempenho - Preocupa-se em avaliar o desempenho do sistema e envolve questões como a taxa de transferência de informação, tempo de resposta, consumo de memória, entre outros (Carlos Vazquez 2016). No contexto deste projeto não foram identificados requisitos de desempenho.

Suportabilidade - Diz respeito a preocupações como a testabilidade, manutibilidade, compatibilidade, escalabilidade, entre outras (Coepd 2014). No contexto deste projeto foi identificado o requisito **R.5** como requisito de suportabilidade.

+ **(Outros)** - Abrange restrições que não se enquadram nas categorias previamente mencionadas e que se encontram de seguida enunciadas:

- **Restrições de desenho** - Como o próprio nome indica, limita o desenho do sistema como por exemplo ao restringir-se o tipo de base de dados a utilizar (Coepd 2014). No âmbito deste projeto não foram identificados requisitos que se enquadram nesta categoria
- **Restrições de implementação** - Requisitos que obriguem o uso de uma determinada linguagem de programação, biblioteca ou de determinados padrões (Coepd 2014). No âmbito deste projeto não foram identificados requisitos que se enquadram nesta categoria
- **Restrições de interface** - Restrições que indicam os mecanismos de interação com sistemas externos ou internos (Coepd 2014). No âmbito deste projeto os requisitos **R.7 e R.10** enquadram-se nesta categoria.
- **Restrições físicas** - Limitações físicas que afetam o hardware que alberga o sistema, como por exemplo o tamanho, forma e peso (Coepd 2014). No âmbito deste projeto não foram identificados requisitos que se enquadram nesta categoria.

3.4 Casos de Uso

Uma forma de capturar requisitos funcionais são os casos de uso. Cada caso de uso define uma interação entre um ator externo e o sistema de modo a chegar a um determinado objetivo. Um caso de uso deve usar linguagem do domínio não ambígua e descrever todo o fluxo de interação entre o ator e o sistema, incluindo cenários alternativos e de erro (Malan e al 2001).

Na figura 3.4 está presente o diagrama de caso de uso que representa as funcionalidades da solução implementada. Pode-se primeiro observar a diferença entre utilizadores autenticados e não autenticados, pois os últimos apenas conseguem efetuar log in na plataforma.

De seguida pode-se observar que, dentro dos utilizadores autenticados, existem vários atores, alguns conseguem executar os mesmo processos, mas existem diferenças entre todos eles.

De primeira vista, consegue-se observar que o *Data Officer* é o ator que mais funções possui, sendo que o mesmo consegue ativar *Data Owners*, *Data Stewards* e *Data Custodians*, quer utilizando os workflows de ativação ou o de carregamento em massa, consequentemente, é fundamental que o mesmo consiga executar workflows para tal. No âmbito das DGTs, uma vez que o mesmo é o responsável pelas mesmas, o *Data Officer* consegue criar, desenvolver e visualizar DGTs.

O admin ou administrador, é o responsável pela criação de workflows, para que os outros atores consigam efetuar os seus trabalhos. Como tal, é necessário que o admin consiga criar e utilizar os workflows.

O *Activation Manager*, uma vez que é um elemento no processo de ativação de *Data Owners*, *Data Stewards* e *Data Custodians*, é vital que o mesmo os consiga ativar, e como referido em cima, para tal é essencial que consiga executar workflows.

Business Users, *Data Stewards* e *Data Custodians* têm o mesmo trabalho. Trabalho esse que é desenvolver ou catalogar as DGTs, mesmo que atuem em secções distintas, como será referido mais à frente, é necessário que os mesmos consigam desenvolver e visualizar DGTs, e para tal necessitam de conseguir utilizar workflows.

Por último, o ator *CoE DAAI*, apenas necessita de conseguir visualizar as DGTs, pois, uma vez que o mesmo é um responsável pelo *Data Lake* do cliente, este necessita de validar alguns atributos no processo de validação da mesma.

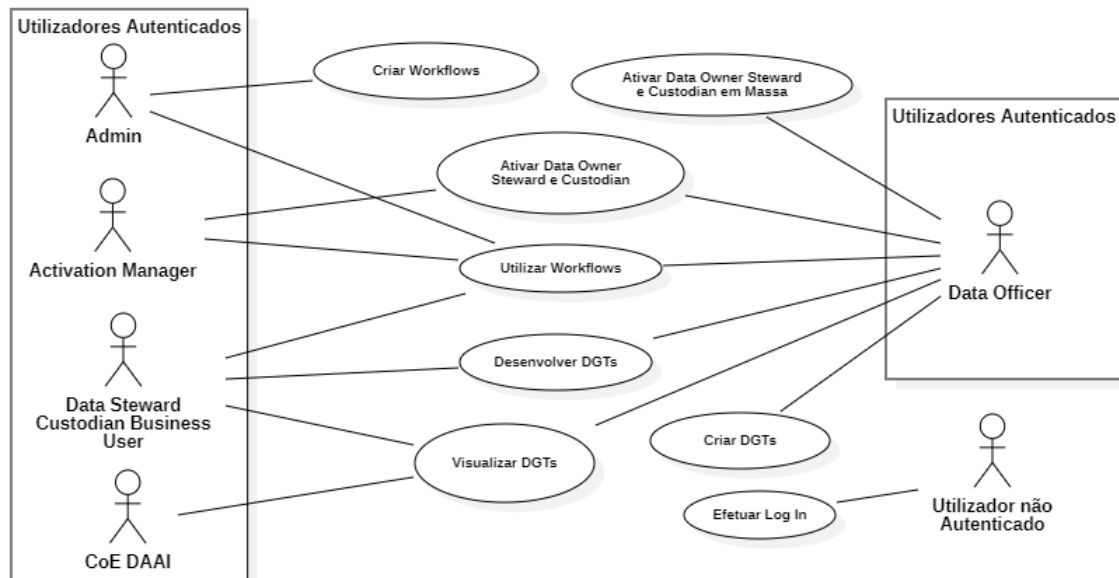


Figura 3.4: Diagrama de Casos de Uso

Capítulo 4

Análise de Valor

Neste capítulo é efetuado uma análise de valor ao projeto, que por sua vez, inclui a identificação das forças, fraquezas, oportunidades e ameaças, tal como uma avaliação do valor percebido.

A criação de valor e a inovação são alguns dos aspetos mais importantes quando se avalia se um negócio é bem sucedido ou não.

O processo de decisão antes de se investir num produto novo pode ser bastante difícil para a maioria das empresas. Deste modo, uma análise de valor poderá ser feita para ajudar nesse processo.

A metodologia mencionada anteriormente, análise de valor, visa identificar, quantificar e retificar fraquezas em processos e produtos e ao mesmo tempo fornecer um conjunto de funções a um custo mínimo. Os princípios subjacentes são os seguintes (Kermode, Sivaloganathan e Shahin 2000):

- Os clientes focam-se nas propriedades funcionais dos produtos.
- No desenho de um produto todos os custos incidem no provimento de funcionalidades.

A análise de valor também examina as vantagens e os custos para reduzir as despesas sem diminuir a qualidade do produto final, por outras palavras, tem como objetivo atingir o valor máximo possível de uma maneira sustentável.

Este capítulo descreve a análise de valor deste projeto para avaliar o seu potencial valor tendo em consideração o contexto do problema a ser resolvido.

Também são descritos os conceitos que estão associados a este tópico, incluindo o Processo de Inovação, o New Concept Development (NCD) e o Valor da Solução.

4.1 Processo de Inovação

Inovação pode ser considerada como a introdução de algo novo, por exemplo, bens, produtos, ou métodos, cujo resultado final visa evoluir e manter uma estratégia de negócios adaptados às necessidades dos clientes, pois eles estão sempre a variar.

De acordo com Peter e Koen (2002), o processo de inovação pode ser dividido em três áreas: Fuzzy Front End (FFE), New Product Development (NPD) e Comercialização como pode ser observado na figura 4.1.

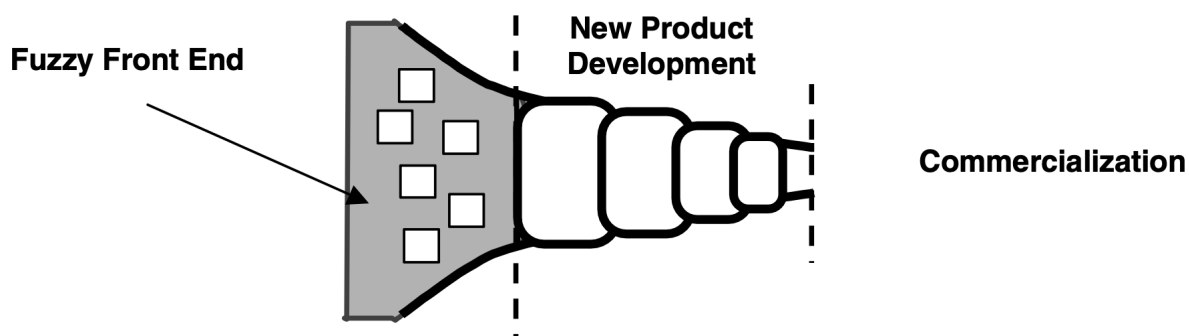


Figura 4.1: Innovation Process Parts (Peter e Koen 2002)

O FFE é frequentemente visto como "uma das maiores oportunidades para melhorar o processo geral de inovação", uma vez que a parte do processo do NPD já foi aprimorada de acordo com (Peter e Koen 2002).

(Peter e Koen 2002) também menciona que "muitas das práticas que auxiliaram a parte do NPD não se aplicam ao FFE", visto que alguns aspetos, como a natureza do trabalho, data de comercialização, nível de financiamento, receitas expectáveis, atividades e medidas do progresso são muito diferentes, como mostrado na tabela 4.1.

Tabela 4.1: Differences between FFE and NPD. Adaptado de (Peter e Koen 2002)

Aspetos	FFE	NPD
Natureza do trabalho	Experimental muitas vezes caótico. Momentos de "Eureka". Pode agendar trabalho - mas não invenção	Disciplinado e orientado para os objetivos com um plano do projeto
Comercialização na data de financiamento	Imprevisível ou incerto. Variável - nas fases iniciais muitos projetos podem ser "bootlegged", enquanto que outros irão precisar de financiamento para prosseguir	Alto grau de certeza. Orçamentado
Expectativas de receita	Muitas vezes incerto, com elevado nível de especulação	Previsível, com elevada certeza, análise e documentação à medida que a data do lançamento do produto se aproxima
Atividade	Pesquisa elaborada individualmente e em equipas para minimizar o risco e otimizar o potencial	Produto multifuncional e/ou equipa de desenvolvimento de processos
Medidas do progresso	Conceitos fortalecidos	Conquista de marcos

Devido à falta de termos e condições para elementos-chave do FFE em empresas, Peter e Koen (2002) desenvolveu o modelo NPD que tenta fornecer uma terminologia comum para

esses elementos-chave.

4.2 New Concept Development

Para melhor se entender o modelo NCD, (Peter e Koen 2002) fornece a definição das seguintes palavras:

- **Oportunidade** - "Uma lacuna de negócios ou tecnologia, que uma empresa ou indivíduo percebe que existe entre a situação atual e um futuro imaginário para capturar uma vantagem competitiva, responder a uma ameaça, resolver um problema ou melhorar uma dificuldade";
- **Ideia** - "A forma mais embrionária de um novo produto ou serviço. Muitas vezes consiste numa visão de alto nível da solução idealizada para o problema identificado pela empresa";
- **Conceito** - "Tem uma forma bem definida, incluindo uma descrição escrita e visual, que inclui os seus principais recursos e benefícios para o cliente combinados com um vasto entendimento da tecnologia necessária".

Como é observável na figura 4.2, o NCD é uma relação e não um modelo linear que é composto por três tópicos cruciais (Peter e Koen 2002).

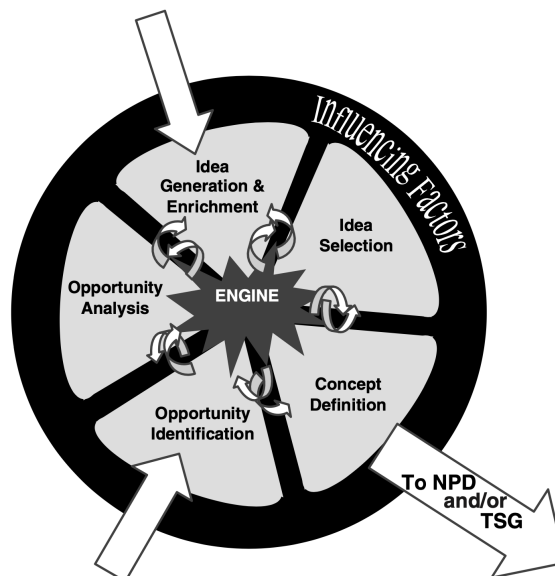


Figura 4.2: New Concept Development (NCD) (Peter e Koen 2002)

- **Motor** - Representa a liderança, cultura e estratégias de negócio da organização que estimula os cinco elementos-chave controláveis pela organização.
- **Elementos-Chave** - Cinco elementos de atividade controláveis do FFE, nomeadamente identificação e análise da oportunidade, geração e seleções de ideias e por último definição do conceito.

- **Fatores de influência** - Estes fatores podem ser as capacidades da organização, o mundo exterior (leis, política governamental, concorrentes, etc) e as ciências capacitadoras (internas ou externas) que possam estar envolvidas. Estes fatores influenciam todo o processo de inovação e não podem ser controlados pelas organizações.

Na mesma figura, as setas apontadas para o centro representam onde tudo começa - identificação da oportunidade ou geração de ideias - e a seta a apontar para fora mostra que os conceitos saem do modelo NCD e entram no NPD.

4.2.1 Identificação da Oportunidade

Conforme analisado no modelo NCD, um dos cinco elementos-chave é a identificação de oportunidades. Este é um ponto de partida para construir um ou mais conceitos que, eventualmente, podem levar a alguns tipos de inovação.

A identificação de oportunidades tem, na sua essência, o objetivo de identificar o que a organização quer perseguir. Geralmente é impulsionado por metas de negócios que podem levar a uma nova direção ou uma atualização para um produto existente. A identificação geral da oportunidade define o mercado ou a área tecnológica da qual a empresa pode querer participar.

Alguns exemplos destas oportunidades podem ser, a resposta de curto prazo a uma ameaça competitiva, uma possibilidade de capturar vantagem competitiva ou um meio de reduzir os custos operacionais (Peter e Koen 2002). Este elemento é tipicamente motivado pelos objetivos do negócio.

Tendo em consideração o âmbito deste projeto e os objetivos do cliente, as oportunidades identificadas são:

- Necessidade de uma nova solução para automatização dos processos de DGT;
- Necessidade de uma solução com um bom design que permita a evolução do sistema de forma sustentável e escalável;
- Necessidade de uma solução que seja flexível e confiável, evite erros e promova a visibilidade e controlo da gestão dos diversos processos de DG.

4.2.2 Análise de Oportunidade

Na análise de oportunidade, são avaliadas as oportunidades identificadas e se valem a pena prosseguir-se.

Antes do estabelecimento de oportunidades específicas de negócio e de tecnologias, informações adicionais são necessárias. Muitas vezes, essas informações são obtidas com avaliações do mercado e experiências científicas.

O esforço despendido pela organização para tentar obter esta informação deve ser sempre considerado e depende do valor obtido ao reduzir as incertezas sobre a atratividade da oportunidade (Peter e Koen 2002).

Para este efeito, escolheu-se efetuar uma análise SWOT ao projeto. A análise SWOT (abreviação de forças, fraquezas, oportunidades e ameaças) é uma ferramenta de estratégia

de negócio usada para avaliar como uma organização se compara à concorrência, esta análise inclui fatores internos e externos à organização (Teoli, Sanvictores e An 2021).

De seguida encontra-se definida a terminologia sustentada pela ferramenta (Sarsby 2016):

- **Fatores internos:** Fatores que a pessoa ou a organização tem controlo sobre;
- **Fatores externos:** Fatores que a pessoa ou a organização tem pouco ou nenhum controlo sobre;
- **Fatores benéficos:** Fatores que ajudam a alcançar o sucesso;
- **Fatores prejudiciais:** Fatores que bloqueiam ou impedem o sucesso.

Por último, tem-se as 4 quadrantes da análise SWOT que são representadas pelas letras do acrónimo (Sarsby 2016):

- **S (Forças):** São fatores internos e benéficos que contribuem uma oportunidade ou que ultrapassam uma dada ameaça. As forças podem ser de natureza económica, tecnológica, pessoal, entre outras;
- **W (Fraquezas):** São fatores internos e prejudiciais que significam as incapacidades de tirar proveito de uma oportunidade ou as vulnerabilidades a uma dada ameaça;
- **O (Oportunidades):** São fatores externos e benéficos que podem resultar de avanços tecnológicos, tendências sociais, entre outros;
- **T (Ameaças):** São fatores externos e prejudiciais que podem derivar de novos competidores no mesmo segmento de mercado, regulações restritivas, entre outros.

A figura 4.3 representa as quatro quadrantes da análise SWOT no contexto deste projeto.

Forças 1. Automatização dos processos de Data Governance 2. Centralização de todos os processos de Data Governance	Fraquezas 1. Prazos e pressões 2. Manutenção do desenvolvimento
Oportunidades 1. Primeiro projeto português a ter um processo End-to-End de Data Governance em Colibra 2. Aumento da procura	Ameaças 1. Dependência de uma plataforma 2. Elevado e rápido crescimento dos processos a ser implementados

Figura 4.3: Análise SWOT

Após efetuada a análise SWOT a oportunidade associada a este projeto é considerada como válida. As fraquezas e ameaças identificadas não são impeditivas nem superam as forças e oportunidades.

4.2.3 Geração de ideias

Uma vez analisada e validada a oportunidade o próximo passo reflete-se na geração de potenciais soluções possíveis de se implementar e capazes de alcançar os objetivos associados à oportunidade identificada.

A geração de ideias é o elemento no modelo NCD que lida com as primeiras ideias sobre como resolver o problema identificado pela oportunidade.

Espera-se que essas ideias tenham várias iterações e que novas possam surgir ao longo do processo de discussão e brainstorming.

Tendo em consideração o âmbito do projeto, surgiram as seguintes ideias para cumprir o principal objetivo do trabalho:

- Utilização de workflows Business Process Model and Notation (BPMN) para o desenvolvimento de atividades para os utilizadores dentro da plataforma Colibra;
- Utilização da API disponibilizada para o desenvolvimento de operações "under the hood" automatizadas, isto é, o desenvolvimento de operações que funcionam de forma automática sem necessitarem de input humano;
- Criação de comunidades e domínios para suportar todos os processos de DG do cliente.

4.2.4 Seleção de ideias

A seleção de ideias tem como objetivo selecionar as ideias a perseguir no sentido de obter o maior valor de negócio. Muitas vezes, o problema não é ter ideias novas, mas sim decidir que ideias é que devem ser seguidas de forma a alcançar um maior valor comercial. Efetuar uma boa seleção de ideias é um passo vital para assegurar o sucesso prolongado de um determinado negócio (Peter e Koen 2002).

De modo a atingir este objetivo, o autor decidiu empregar o método de apoio à decisão Analytic Hierarchy Process (AHP).

O AHP consiste na geração de um peso para cada critério de avaliação, neste caso para cada funcionalidade, de acordo com o seu respetivo valor para o decisor. Quanto maior o peso, mais importante é o correspondente critério (R. Saaty 1987).

Este método pode ser dividido em 3 partes distintas, Divisão Hierárquica, Definição de Prioridades e Avaliação da Consistência (T. L. Saaty 1988).

A primeira parte, Divisão Hierárquica, centra-se principalmente em 3 pontos chaves: o objetivo, os critérios e as alternativas. O objetivo relaciona-se com a escolha da melhor abordagem de implementação do sistema. Os critérios que serão empregues às alternativas encontram-se de seguida enumerados:

- **Tempo de desenvolvimento** - Este critério diz respeito ao tempo de desenvolvimento do sistema. Visto que projeto está a ser realizado num ambiente agile, com sprints de tempo definido, existem prazos que devem ser cumpridos. Para este efeito, considera-se uma alternativa superior a outra se o tempo despendido na fase de implementação for menor;
- **Complexidade** - Este critério compreende a dificuldade associada à implementação dos componentes de cada uma das alternativas como também os custos ligados aos

processos de manutenção. Para este efeito, considera-se uma alternativa superior a outra se a dificuldade e os custos de implementar e manter a solução for menor;

- **Desempenho** - Este critério preocupa-se com questões de desempenho da alternativa, tais como a velocidade na pesquisa de informação. Para este efeito, considera-se uma alternativa superior a outra se os tempos de resposta forem menores e se os dados apresentados forem mais recentes.

A segunda parte do método AHP, Definição de Prioridades, consiste em definir prioridades para cada um dos critérios previamente estabelecidos com recurso à escala apresentada na tabela 4.2 [93].

Tabela 4.2: Escala Preferências AHP

Classificação Numérica	Definição
9	Extremamente preferido
8	Muito fortemente a extremamente preferido
7	Muito fortemente preferido
6	Fortemente a muito fortemente preferido
5	Fortemente preferido
4	Moderadamente a fortemente preferido
3	Moderadamente preferido
2	Igualmente a moderadamente preferido
1	Mesma preferência

A tabela 4.3 apresenta a matriz de prioridades tendo por base os critérios de **Complexidade (A1)**, **Desempenho (A2)** e **Tempo de desenvolvimento (A3)**. A mesma evidencia a ordem de prioridade relativa a cada um dos critérios, que neste caso é **Desempenho** com maior prioridade, seguido da **Complexidade** e por último o **Tempo de Desenvolvimento**.

Tabela 4.3: Prioridade relativa de cada critério

	A1	A2	A3	Prioridade Relativa
A1	1	1/2	4	0.33
A2	2	1	5	0.57
A3	1/4	1/5	1	0.10

É de notar que o critério complexidade é quatro vezes mais preferencial do que o critério tempo de desenvolvimento, o critério desempenho é duas vezes mais preferencial do que o critério complexidade e o critério desempenho é cinco vezes mais preferencial do que o critério tempo de desenvolvimento.

O próximo passo traduz-se na priorização das alternativas com base nestes critérios. Para tal, é apresentada uma tabela de prioridades relativas para cada critério tendo por base as alternativas propostas. De maneira a aumentar o grau de clareza das tabelas, as alternativas têm a seguinte nomenclatura **Utilização de Workflows BPMN (B1)**, **Utilização da API (B2)** e **Criação de Comunidades e Domínios (B3)**.

Deste modo, apresenta-se a tabela 4.4 que prioriza as alternativas pelo critério de Tempo de Desenvolvimento

Tabela 4.4: Prioridade relativa das alternativas segundo o critério - Tempo de Desenvolvimento

	B1	B2	B3	Prioridade Relativa
B1	1	1/4	1/7	0.08
B2	4	1	1/3	0.26
B3	7	3	1	0.66

É de notar que a alternativa Utilização da API é quatro vezes mais preferencial do que a alternativa Utilização de Workflows BPMN, a alternativa Criação de Comunidades e Domínios é sete vezes mais preferencial que a alternativa Utilização de Workflows e três vezes mais preferencial que a alternativa Utilização da API.

Logo após, apresenta-se a tabela 4.5 que prioriza as alternativas pelo critério de Complexidade

Tabela 4.5: Prioridade relativa das alternativas segundo o critério - Complexidade

	B1	B2	B3	Prioridade Relativa
B1	1	2	5	0.57
B2	1/2	1	4	0.33
B3	1/5	1/4	1	0.10

É de notar que a alternativa Utilização de Workflows BPMN é duas vezes mais preferencial do que a alternativa Utilização da API e cinco vezes mais preferencial do que a alternativa Criação de Comunidades e Domínios e a alternativa Utilização da API é quatro vezes mais preferencial do que a alternativa Criação de Comunidades e Domínios.

Para finalizar, apresenta-se a tabela 4.6 que prioriza as alternativas pelo critério de Desempenho

Tabela 4.6: Prioridade relativa das alternativas segundo o critério - Desempenho

	B1	B2	B3	Prioridade Relativa
B1	1	3	6	0.63
B2	1/3	1	5	0.29
B3	1/6	1/5	1	0.08

É de notar que a alternativa Utilização de Workflows BPMN é três vezes mais preferencial do que a alternativa Utilização da API e seis vezes mais preferencial do que a alternativa Criação de Comunidades e Domínios e a alternativa Utilização da API é cinco vezes mais preferencial do que a alternativa Criação de Comunidades e Domínios.

Uma vez efetuada a comparação dos critérios com as alternativas, o próximo passo remete na identificação da alternativa que melhor responde ao objetivo inicialmente definido. Para tal, elaborou-se a matriz da tabela 4.7 que relaciona todas as alternativas com todos os critérios.

Os valores que se encontram na coluna mais à direita da tabela, que apresenta a prioridade relativa final de cada solução, foram obtidos a partir da multiplicação da matriz de prioridades das soluções com o vetor de pesos do critério calculado na tabela 4.3 que se encontra

Tabela 4.7: Prioridades relativas das soluções

	A1	A2	A3	Prioridade Relativa
B1	0.57	0.63	0.08	0.56
B2	0.33	0.29	0.26	0.30
B3	0.10	0.08	0.66	0.14
Prioridade Relativa	0.33	0.57	0.10	

representado na última linha da tabela 4.7. Deste modo, é possível inferir que a melhor alternativa é a B1, Utilização de Workflows BPMN, que, como é possível observar-se a partir dos resultados obtidos, obteve uma prioridade relativa de 0.56.

4.3 Valor da Solução

Esta secção fornece as definições de alguns termos fundamentais, como valor, valor para o cliente e valor percebido. É também nesta secção que a proposta de valor da solução proposta para este projeto é analisada.

4.3.1 Valor

O conceito de valor pode ser definido como uma medida expressa em dinheiro ou esforço que reflete o desejo de se reter ou obter um determinado produto ou serviço. Outra definição de valor, indica que este é a troca equivalente de serviço ou bens que um proprietário ou consumidor recebem em troca de dinheiro (Neap e Celik 1999). No entanto também pode ter diferentes significados dependendo se é do ponto de visto do consumidor ou do produtor.

Do lado do produtor, o valor está tipicamente relacionado com os seus clientes e se eles são “leais” para a empresa em termos de perspectiva de negócios, basicamente se eles compram os produtos ou serviços da empresa. Este conceito indica que o valor que um cliente atribui a um dado produto varia conforme a sua percepção pessoal (Neap e Celik 1999).

Do ponto de vista do consumidor, muitas vezes o valor está relacionado com a qualidade dos produtos ou serviços adquiridos de uma empresa, e a satisfação geral e relacionamento com o produtor.

Ao pensar em como agregar valor a um produto, alguns aspetos em ter em consideração são (Rich e Holweg 2000):

- **Valor do uso** - Preocupações sobre a utilidade/funcionalidade do produto;
- **Valor estimado** - Vem da propriedade do produto e das características que lhe são atribuídas pelo consumidor. Estas características são coisas mais relacionadas com a estética e com o valor subjetivo dado pelo cliente ao produto ou serviço, não estando diretamente ligadas à sua utilidade.

Outro aspeto que deve ser considerado é o valor de mercado. Este é um elemento importante que pode ser traduzido como a soma do valor de uso e do valor estimado. O seu resultado final representa o que o mercado/consumidores pagariam pelo produto ou serviço.

No contexto deste projeto, os clientes/consumidores finais correspondem ao cliente que propôs o projeto. Considera-se como o maior benefício da solução o desenvolvimento da

solução para a possibilidade de se efetuar as DGT. Apesar disso, estima-se que o maior sacrifício envolvido na utilização deste produto remete-se para a fase inicial onde os colaboradores ainda não se encontram familiarizados com o produto.

4.3.2 Valor para o Cliente

O valor para o cliente pode ser visto como a compreensão geral do que os clientes realmente desejam e valorizam num produto ou serviço, e o que pode ser feito pelo produtor para atender as necessidades dos consumidores para alcançar uma vantagem competitiva sobre outras empresas (Graf e Maas 2008).

Durante a fase de desenvolvimento de software em grande escala, a gestão eficaz e eficiente dos requisitos do cliente é um dos problemas mais prevalentes e também um dos menos compreendidos. Os problemas encontrados nos requisitos tipicamente não são reconhecidos até ao final do processo de desenvolvimento, onde os impactos negativos são substanciais (Karlsson 2004).

Tendo isto em consideração, é necessária a utilização de métodos ou ferramentas que ajudem os programadores a gerir melhor os requisitos e as respetivas necessidades do cliente. Um desses métodos é o Quality Function Deployment (QFD), que foi o método aplicado ao presente projeto.

QFD foi desenvolvido no Japão no final da década de 1960 e começou a ser usado rapidamente durante a década de 1980 nos Estados Unidos da América. Inicialmente foi proposto para permitir o desenvolvimento de produtos com maior qualidade já que a metodologia recolhe e analisa a "voz do cliente" e concentra-se nas necessidades do cliente (Chan e Wu 2002).

De acordo com Chan e Wu (2002) o QFD é um conceito que fornece um meio de traduzir requisitos do cliente para requisitos técnicos.

Este processo compreende quatro fases onde cada uma delas é composta por atividades que devem ser executadas no decorrer do desenvolvimento do produto (Quality-One 2017). No âmbito deste projeto vai-se analisar e aplicar apenas a primeira fase, denominada de Definição do Produto, pois foi a que se considerou mais apropriada.

A fase supracitada consiste na tradução dos desejos e necessidades do cliente em especificações do produto. Também envolve uma componente competitiva no sentido em que avalia a eficácia de produtos concorrentes que satisfaçam as necessidades do cliente (Quality-One 2017).

De maneira a se aplicar a fase de definição do produto é necessário definir os seguintes parâmetros:

- Necessidades do cliente;
- Importância das necessidades do cliente;
- Características de Engenharia;
- Escalas de correlação;
- Concorrentes.

Com base nos objetivos deste projeto, descritos no capítulo 1, as necessidades do cliente são as seguintes:

- Flexibilidade;
- Centralização;
- Visibilidade e Controle;
- Menos Erros;
- Facilidade de Utilização.

Para atender às necessidades do cliente, foram definidos os seguintes requisitos técnicos:

- Documentação dos Workflows;
- Testes Unitários;
- Qualidade do Código;
- Usabilidade.

4.3.3 House of Quality

House of Quality (HoQ) é uma ferramenta que é utilizada na primeira fase (Definição do Produto) da metodologia QFD, e “demonstra a relação entre os desejos do cliente ou “Whats” e os parâmetros de projeto ou “Hows”. Permite capturar uma grande quantidade de informações em um só lugar usando um formato de matriz (Quality-One 2017).

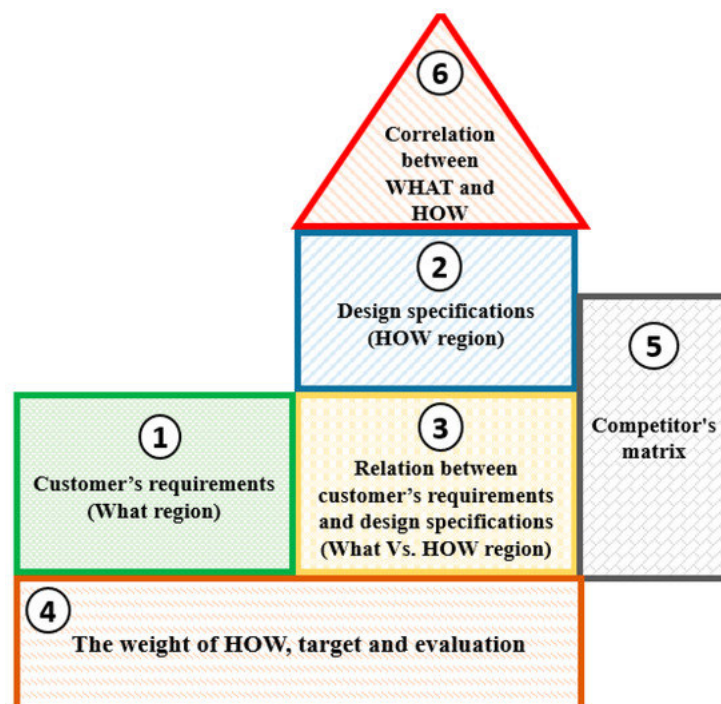


Figura 4.4: Estrutura da House of Quality. (Abdel-Basset 2019)

Segundo Abdel-Basset (2019) os principais componentes do HoQ, e visíveis na figura 4.4 são:

- **Área 1** - Local onde são listados os requisitos do cliente ("whats"), e também onde uma importância é dada a cada um deles de acordo com as prioridades daqueles clientes;
- **Área 2** - Região onde os requisitos técnicos (especificações de projeto - "hows") são colocados;
- **Área 3** - Área onde é definido o relacionamento entre o cliente e os requisitos técnicos ("whats" vs. "hows");
- **Área 4** - Zona onde é feita a soma, linha a linha, de cada prioridade dos requisitos dos clientes multiplicados pelo valor do relacionamento conforme a legenda do modelo;
- **Área 5** - Área onde é feita a comparação do produto com os concorrentes, e a avaliação do quanto satisfazem as necessidades do cliente;
- **Área 6** - Região do modelo onde é feita uma comparação de cada requisito técnico e o quanto as suas importâncias podem afetar uns aos outros.

4.3.4 Valor Percecionado

Valor percebido é o valor que os clientes atribuem a um produto/serviço conforme as suas necessidades, variando de cliente para cliente, ou variando mesmo no valor atribuído pelo próprio fornecedor em relação à percepção que o cliente tem do mesmo (Woodall 2003).

Esta diferença significa que muitas vezes o valor atribuído pelos clientes não depende tanto do valor monetário do produto/serviço, mas sim da capacidade do mesmo de corresponder às necessidades do mesmo e expectativas definidas (Woodall 2003).

(Ravald e Grönroos 1996) define valor percebido pelo cliente como a razão entre os benefícios percebidos e o sacrifício percebido.

(Ulag e Chacour 2001) consideram que o valor percebido pelo cliente é um trade-off entre benefícios e sacrifícios percebidos pelo cliente na oferta de um fornecedor.

Para percebermos melhor a definição de valor percebido, é necessário entender o que os autores consideram por benefícios e sacrifícios.

De acordo com Ravald e Grönroos (1996), benefícios e sacrifícios são:

- **Benefícios** - Combinação de atributos físicos, atributos de serviço e suporte técnico disponível em relação ao uso específico do produto;
- **Sacrifícios** - Todos os custos que o comprador enfrenta ao fazer uma compra: preço da compra, custo de aquisição, transporte, instalação, reparos e manutenção, risco de falha ou mau desempenho.

No âmbito deste projeto, o valor percebido pode ser definido como a razão entre os benefícios, isto é, criar uma solução que traga valor ao cliente, desenvolvendo funcionalidades que tragam melhorias aos atuais processos de DG, desenvolvendo-as de acordo com os requisitos levantados, construindo um sistema que facilite as operações de DG, de uma forma mais automática, completa, mais inteligente e intuitiva para os seus utilizadores, constituindo uma mais valia para os seus processos organizacionais.

E os sacrifícios, nomeadamente, custos associados à implementação e manutenção da solução proposta, como salários dos programadores e software/hardware necessários.

4.3.5 Proposta de Valor

A proposta de valor pode ser definida como os benefícios que são oferecidos por uma empresa aos seus clientes. Osterwalder e Pigneur (2003) descrevem proposta de valor como a definição de como os itens de valor, tal como produtos e serviços, bem como serviços de valor agregado complementar, são agregados e oferecidos para atender às necessidades dos clientes.

Para melhor avaliar a proposta de valor de uma empresa para o seu segmento de clientes, uma ferramenta gráfica denominada de Value Proposition Canvas, que foi proposta por Osterwalder, Pigneur e al. (2015), pode ser usada.

Esta ferramenta divide a proposta de valor num mapa de valor que descreve os recursos de uma proposta de valor específica (produtos e serviços, limitadores de dor e criadores de ganhos) e um perfil do cliente que detalha um segmento de cliente específico (tarefas, dores e ganhos do cliente) (Osterwalder, Pigneur e al. 2015):

- **Produtos e Serviços** - Lista de todos os itens de valor que a empresa tem para oferecer aos seus clientes;
- **Limitadores de dor** - Descrição de como os produtos/serviços aliviam dores dos clientes;
- **Criadores de Ganhos** - Descrição de como é que os produtos/serviços criam ganhos para os clientes;
- **Trabalhos** - Descrição do que é que os clientes estão a tentar alcançar ou finalizar nos seus trabalhos/vida;
- **Dores** - Maus resultados, riscos e obstáculos relacionados com os trabalhos dos clientes;
- **Ganhos** - Descrição dos resultados/benefícios que os clientes querem alcançar.

A Value Proposition Canvas deste projeto é apresentada na Figura 4.5. A parte do mapa de valor do modelo está relacionada com a passagem de todos os processos de DG do cliente para o Collibra. Por outro lado, o segmento de clientes é o cliente que propôs o projeto e os seus stakeholders de negócios.

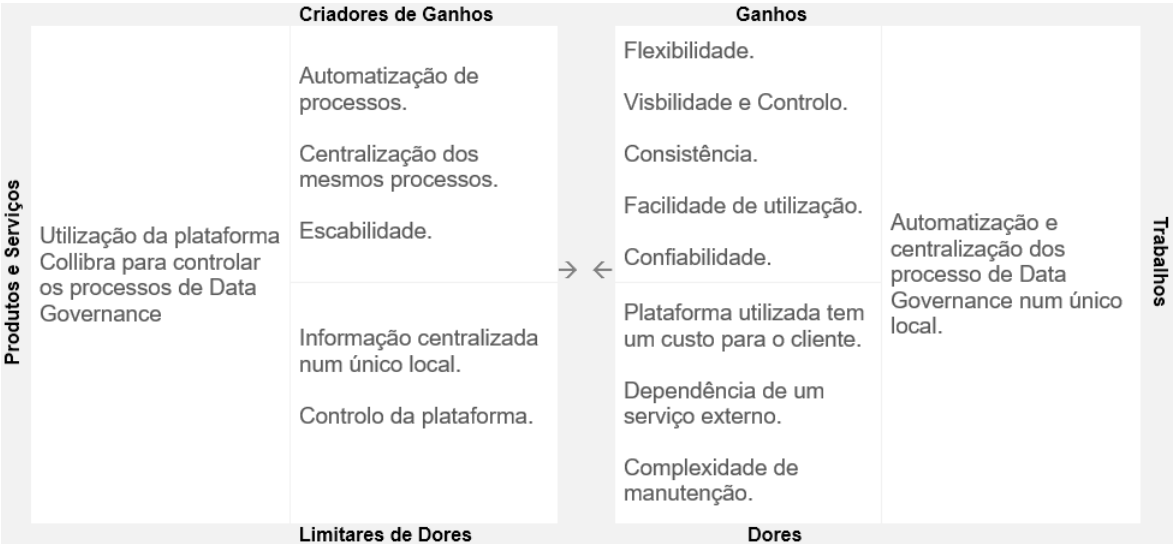


Figura 4.5: Value Proposition Canvas

Capítulo 5

Design

O presente capítulo descreve o desenho e a arquitetura da solução. É apresentada a arquitetura proposta para a solução usando boas práticas de engenharia de software, tal como uma especificação de funcionalidades, uma vez que a maioria destes requisitos são feitos utilizando a metodologia BPMN (1997).

Ao longo deste capítulo, as imagens encontram-se em formato vetorial, pelo que é possível fazer-se zoom para se conseguir visualizar as imagens. Esta decisão foi tomada devido ao facto de existirem vários workflows demasiado grandes e como o número do mesmo não é reduzido, o autor partilha da opinião que o documento fica melhor apresentado sem ter múltiplas páginas horizontais apenas com uma imagem.

5.1 Arquiterura

Foi utilizado o diagrama de componentes representado na figura 5.1 para se representar as diversas partes que compõem a plataforma.

O diagrama de componentes encontra-se dividido em "Ambiente Interno" e "Ambiente Externo", sendo que o primeiro se refere ao ambiente Colibra. Estão também representados dois utilizadores, o administrador e o *End-User* que engloba todos os outros descritos no capítulo 3.

É possível observar que o componente Workflows, possui dois componentes no seu interior, Criação e Utilização Front-End. Este cenário foi representado desta maneira para expressar as duas vertentes dos workflows, a vertente de criação e a vertente de utilização, uma vez que os workflows para serem utilizados têm de ser criados, e estas duas situações distintas são efetuadas por utilizadores díspares em momentos diferentes.

O administrador, através do Eclipse ¹, consegue criar workflows, utilizando a Colibra API. Por sua vez, estes workflows (através do Front-End), vão ser executados pelos outros utilizadores (*End-Users*), estas execuções modificam o Colibra Front-End, utilizando, mais uma vez, o Colibra API,. Estas alterações são efetuadas através do Colibra Back-End. A utilização dos workflows, com uso da Colibra Email API, consegue enviar emails.

¹O Eclipse é uma plataforma de desenvolvimento integrado (IDE) de código aberto amplamente utilizado. É um ambiente de programação que fornece ferramentas e recursos para facilitar o desenvolvimento de software em várias linguagens de programação, nesta situação é utilizado para desenvolver os workflows, através de metodologia BPMN com o auxílio da linguagem Groovy ² para desenvolver os scripts.

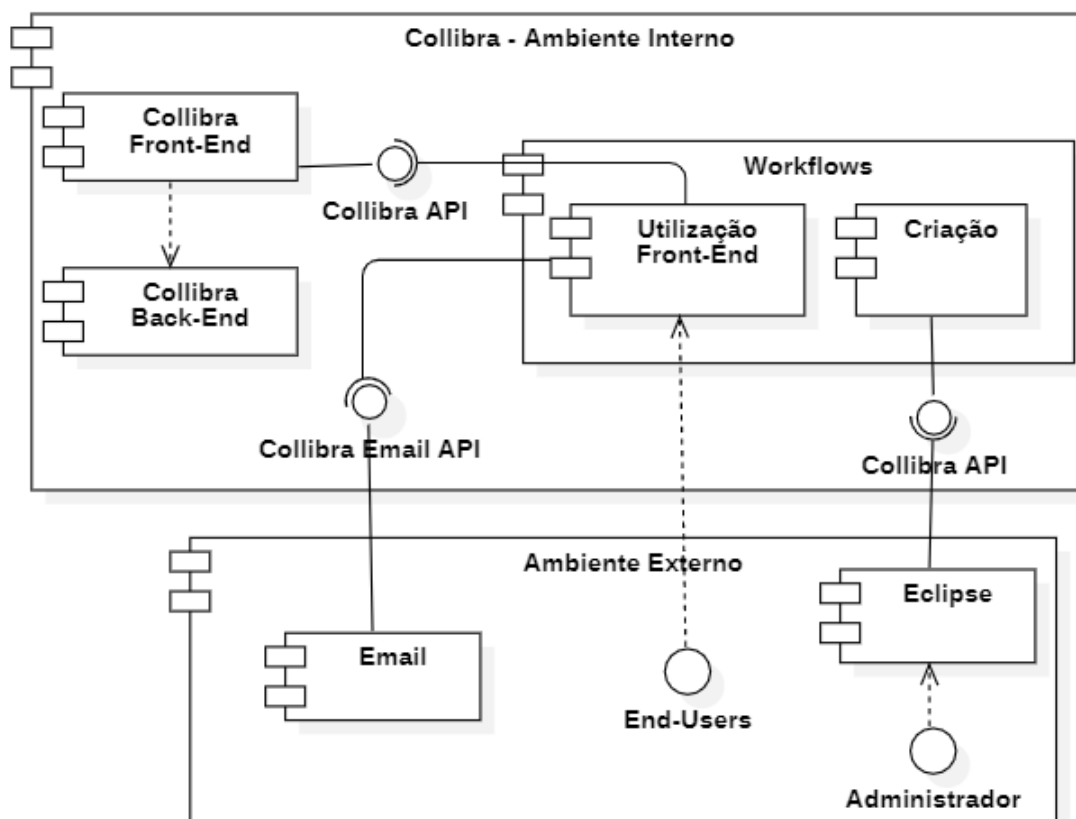


Figura 5.1: Diagrama de Componentes

5.2 Especificação de Funcionalidade

Nesta secção vai ser apresentada a arquitetura quer do sistema, quer dos workflows que atuam no sistema.

5.2.1 Ativação Data Owner/Steward/Custodian R.1, R.2 e R.3

Os requisitos de ativação de *Data Owner*, *Data Steward* e *Data Custodian* são muito semelhantes, por isso, o workflow dos três é igual no que diz respeito ao aspeto do mesmo.

Como não existe apenas um workflow para os três processos, existe, então, alterações de nomes (*Data Owner*, *Steward* e *Custodian*), tipos de ativos (*Data Owner* aplica-se a ativos do tipo data domain, *Data Steward* aplica-se a ativos do tipo data set e *Data Custodian* aplica-se a ativos do tipo sistema), comunidades e domínios (os três tipos de ativos distintos, encontram-se dentro de comunidades e domínios distintos). No entanto, os processos são análogos, e por essa razão, será, apenas discutido o workflow referente à ativação do *Data Owner*.

Um requisito, que se aplica aos três workflows de ativação, é a necessidade de existir uma notificação do estado onde o processo se encontra ao Activation Manager. Esta notificação é esperada acontecer ao quinto e nono dia, sendo que após o nono dia deverá acontecer todos os dias. O email deverá ter informação sobre o estado em que se encontra o processo, isto é, em que local do processo é que se encontra, e em que pessoa é que se encontra. Este requisito é visível na última pista, denominada de "*Notification Lane*".

Nos três processos existe um cargo denominado de *Activation Manager*, este elemento é selecionado aleatoriamente a partir de todos os possíveis, que tal como discutido na secção 4, o *Activation Manager* é um *Data Officer*.

O esperado deste requisito é que se ative um *Data Owner* num ativo do tipo Data Domain em Colibra, este tipo de ativo representa o domínio em que uma tabela ou coluna se encontram, este domínio pode ser "Cliente", "Produtos e Serviços Finais", "Ativos de Retalho", "Pessoas", "Materiais e Serviços" e "Fornecedores". Para tal, é necessário que o utilizador que fizer esta ativação seja um *Data Officer*, desde modo, a possibilidade de se executar o Workflow apenas se encontra disponível em ativos do tipo data domain e para utilizadores que forem *Data Officers*.

Após o *Data Officer* executar o workflow, é esperado que o mesmo escolha o utilizador que deverá ficar como *Data Owner* e que insira a ata em que foi feita a escolha do possível elemento para o cargo de *Data Owner*. Após esta seleção, é despoletado o loop de emails, referido anteriormente, tal como uma task para o *Data Owner* selecionado, com a opção de aceitar ou rejeitar o cargo, desta task existem dois possíveis caminhos.

Caso o utilizador rejeite, entra-se num loop, em que é-lhe pedido que forneça uma justificação, que será enviada por email para as partes interessadas. É também enviado uma task para o *Activation Manager*, este é escolhido aleatoriamente dentro de todos os utilizadores com este cargo, para escolher um novo elemento para o cargo de *Data Owner*, depois procede-se novamente para a task em que o utilizador selecionado para o cargo de *Data Owner* pode aceitar ou rejeitar o cargo. Este loop é executado até que um utilizador aceite o cargo.

Caso o utilizador aceite, é enviado um email e uma notificação no Colibra para as partes interessadas a comunicar o novo cargo, e o cargo é adicionado ao utilizador em questão no Colibra.

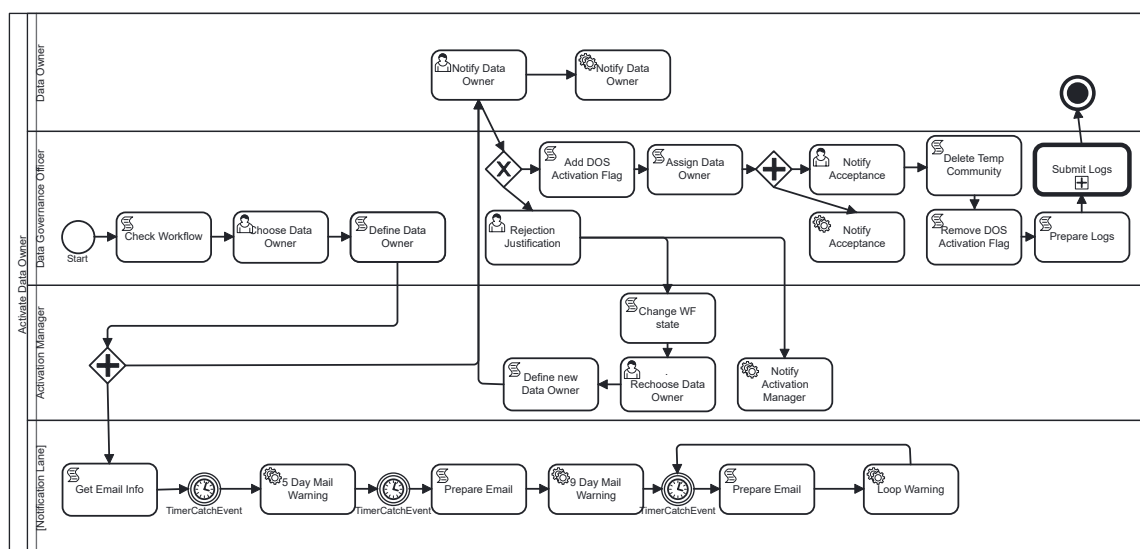


Figura 5.2: Workflow de Ativação do *Data Owner*

5.2.2 Lineage Path R.4

Data Sets, os tipos de ativos discutidos neste requisito, em Collibra, são usados apenas como representações das tabelas que são ingeridas a partir do Data Lake do cliente. Esta criação é gerada pelo Workflow "Data Set Management", que é apresentado mais à frente e por essa razão a explicação sobre o porque de o processo acontecer desta forma é discutido nesse requisito. No entanto para este requisito é importante perceber o objetivo do mesmo.

O esperado deste workflow é importar características de uma ou várias colunas, que têm relação com as colunas que estão relacionadas com o Data Set. Isto é, um data set tem uma relação para uma ou mais colunas, que por sua vez cada uma dessas colunas pode também ter relações para outras colunas.

Cada tipo de ativos, nomeadamente tabela ou coluna, têm atributos ou características próprias, tal como uma descrição, se é dado pessoal ou não, entre outros. Deste modo, o objetivo deste workflow é disponibilizar todos os atributos possíveis, dando a escolher ao *Data Steward* (do data set em questão) e posteriormente enviar um relatório com todos os atributos que foram importados.

Quer para o data set, quer para as colunas que estão relacionadas com o mesmo é necessário que estas possuam um *Data Steward*, não havendo impedimento de ser o mesmo utilizador para ambos. Isto é necessário pois este workflow apenas está disponível para ser executado em Data Sets e por *Data Stewards*.

Deste modo, quando o *Data Steward* executa o workflow, será despoletado uma task para o mesmo, nessa task, é evidenciado todos os atributos catalogados, os seus valores e de onde é que os mesmos vêm, sendo que o *Data Steward* terá de selecionar quais quer importar e quais não quer, tendo a possibilidade de selecionar todos.

Em virtude de haver certos atributos em que cada ativo apenas pode ter um valor, é crucial certificar-se de que não são importados múltiplos atributos do mesmo tipo. Com isto em mente, sempre que seja selecionado mais do que um atributo do mesmo tipo, é despoletada uma task, evidenciado apenas os atributos repetidos, de forma a que o *Data Steward* selecione apenas um valor por atributo. Este processo é um loop que finaliza quando não houver atributos repetidos.

Com o RGPD, o cuidado com dados pessoais nunca foi tão importante, como tal, é possível que alguma das tabelas fontes sejam um dado pessoal, isto é determinado tendo em consideração o valor do atributo que corresponde a isso, sendo que se o atributo tiver o valor a verdadeiro, quer dizer que a tabela fonte representa dados pessoais, caso esteja a falso, quer dizer que não.

Este é um ponto importante, pois para se importar este valor para o Data Set é necessário que o *Data Steward* da coluna que representa dados pessoais, decida se a coluna do data set em questão, deverá ser considerada como dado pessoal ou não. Desta forma, semelhante às tasks anteriores é enviado uma ao *Data Steward* dos ativos que possuírem dados pessoais para fazerem essa seleção.

Por último deve ser enviado um relatório, via email, com todos os atributos importados e para que colunas é que foram importados. Todos os ativos apresentados no email devem ser um link para o mesmo dentro da plataforma Collibra.

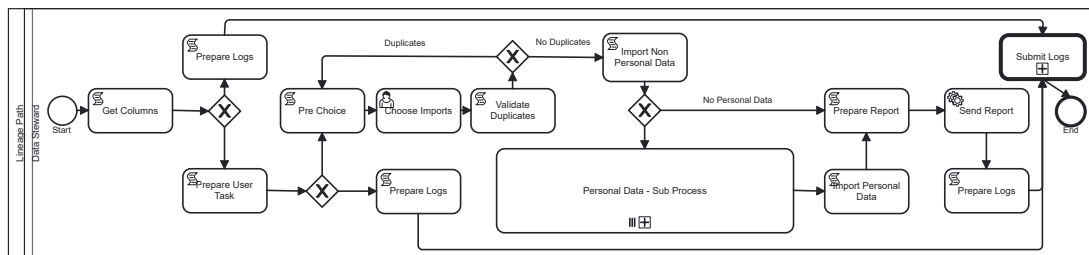


Figura 5.3: Workflow do Lineage Path

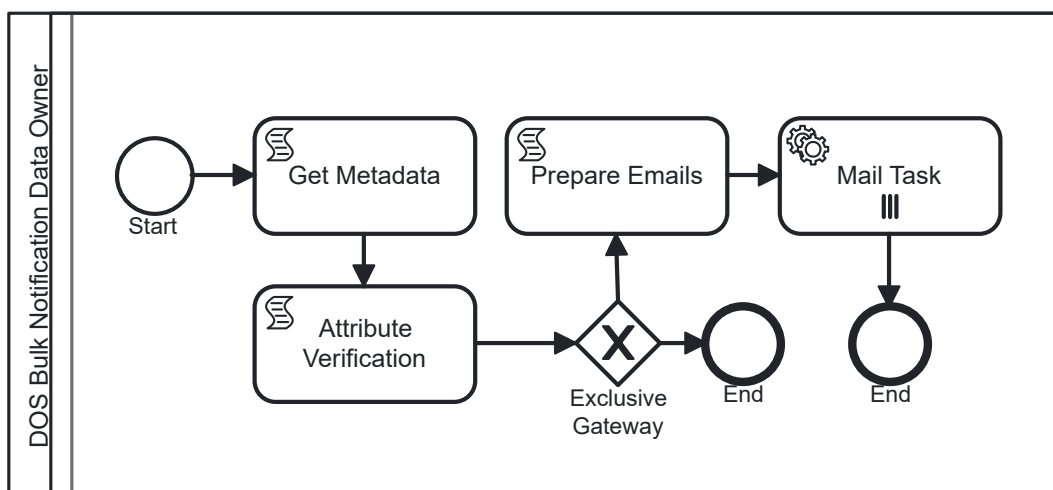
5.2.3 Carregamento em massa R.6

A ativação de Data Organisation Structures (DOS) faz parte do dia-a-dia do Gabinete de Dados Corporativo do cliente. Desta forma, o carregamento em massa de DOS vem como um complemento aos workflows de ativação de DOS. Esta feature permite a atribuição de funções de recursos, como *Data Owner*, *Data Steward* e *Data Custodian* em massa para um determinado conjunto de ativos.

Sempre que alguma ativação é efetuado, é necessário notificar-se os elementos dos grupos de *Data Officers*, e *CoE DAAI*.

O processo de ativação em massa de DOS é semelhante às três roles, no entanto no de ativação de *Data Steward* e de *Data Custodian* é necessário efetuar-se mais um passo do que o do *Data Owner*, pois, no do *Data Steward* é necessário propagar a responsabilidade de data sets para as tabelas ligadas ao data set em questão e no do *Data Custodian* é necessário propagar-se para as bases de dados ligadas ao sistema em questão, daí estes workflows, possuírem mais uma tarefa no workflow do que o do *Data Owner*.

Para tal, foram desenvolvidos três workflows semelhantes, um para *Data Owner*, outro para *Data Steward* e o último para *Data Custodian*, para efetuar a ativação de cada tipo de função. No entanto, os workflows de *Data Steward* e *Data Custodian* são iguais, apenas é apresentado o referente ao *Data Steward*.

Figura 5.4: Workflow de Ativação *Data Owner* em Massa

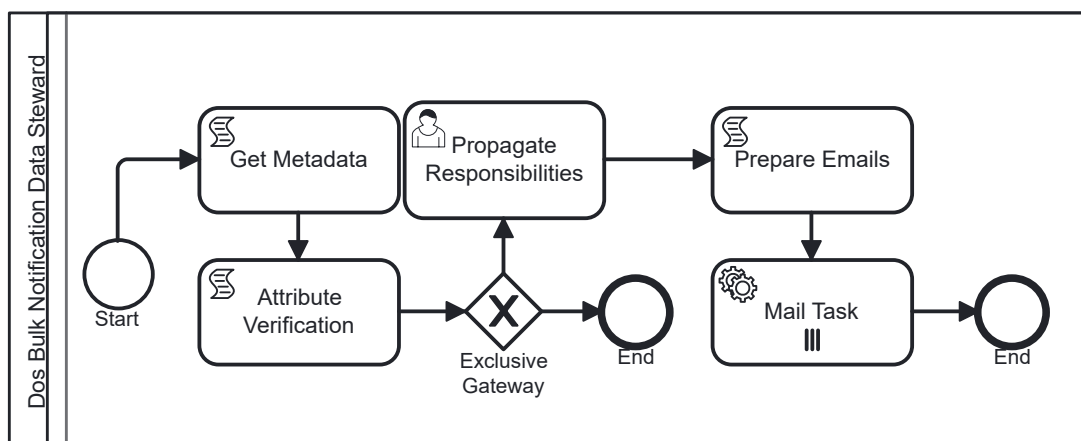


Figura 5.5: Workflow de Ativação *Data Steward* em Massa

5.2.4 Data Set Management R.9

O objetivo deste requisito é que caso se verifique que exista alguma tabela, originada do *Data Lake*, sem data set correspondente, seja automaticamente criado um data set, para servir como um "clone" da tabela, com os atributos devidamente catalogados automaticamente.

Como tal, este workflow é automaticamente executado todos os dias percorrendo a comunidade onde as tabelas oriundas do data lake existem, efetuando a verificação de que todas as tabelas têm um data set correspondente. Caso o mesmo não exista, é criado um novo.

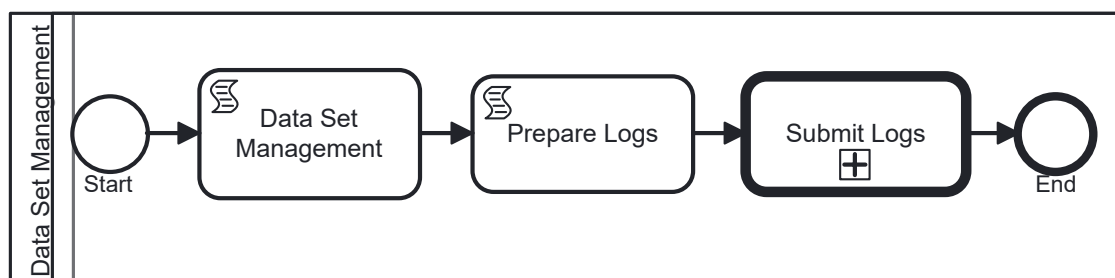


Figura 5.6: Workflow do Data Set Management

5.2.5 Customização Landing Page R.8

Para o desenvolvimento do requisito referente à criação de uma Landing Page para o Collibra, foi fornecido o mockup presente na imagem 5.7.

Existem partes, quer do mockup na imagem 5.7, quer da Landing Page implementada no capítulo 6 que se encontram ocultas, como é possível observar-se os retângulos verdes a taparem os vídeos. Esta ocultação foi efetuada para se proteger o cliente de ser identificado, no entanto foi feito de forma a se conseguir visualizar tudo que foi implementado.

O esperado deste requisito era o desenvolvimento de uma página web estática, com a User Interface (UI) igual à do mockup, a criação de vários botões que redirecionassem para diversas páginas e uma estrutura que fosse capaz de rodar entre múltiplos vídeos.

Ao todo existem onze botões que redirecionam para diversos locais, sendo que os mesmos eram o botão no canto superior denominado de "New here?", este redirecionava para uma página do sharepoint do cliente onde existem documentos que auxiliavam os novos utilizadores.

Os próximos sete botões encontram-se dentro do retângulo branco, nesta divisória, cada ícone é um botão que redireciona para uma página dentro do Colibra, onde os utilizadores conseguem visualizar o correspondente ao nome do botão, nomeadamente o botão denominado de "*Global Views*" redireciona para a página dentro do Colibra onde os utilizadores conseguem visualizar todas as *Global Views* existentes.

Os próximos três botões encontram-se na divisória denominada de "*Explore more materials*", todos estes botões redirecionam para uma página do sharepoint do cliente, no entanto, cada um deles redireciona para uma pasta diferentes com informações diferentes.

Deste modo o primeiro botão "*Did you know...?*" redireciona o utilizador para uma pasta com documentos que possuem informações interessantes, que os utilizadores devem estar familiarizados, sobre a plataforma Colibra. O segundo botão "*Tips and Tricks*" redireciona o utilizador para uma pasta onde é possível observar-se documentos com orientações e sugestões para os utilizadores. O terceiro botão "*What's new?*" redireciona o utilizador para uma pasta onde existe documentos com as últimas novidades sobre plataforma Colibra.

O último botão, que se encontra na divisão denominada por "*Share with us your feedback!*" redireciona os utilizadores para um formulário do Google, criado pelo cliente, onde os utilizadores conseguem partilhar feedback e informações de melhoria sobre a plataforma Colibra.

Por último existem mais dois botões, no entanto estes não redirecionam para nenhuma página, servem apenas para o utilizador navegar no carrossel dos vídeos, podendo circular para a esquerda ou direita visualizando novos vídeos.



Figura 5.7: Landing Page Mockup

5.2.6 Data Gathering Tool R.16, R.17 e R.18

Uma vez que estes três requisitos se referem ao mesmo tópico, a DGT, optou-se por apresentá-los na mesma secção.

5.2.6.1 Criação da DGT

O primeiro passo é então criar-se a DGT em si, e para tal acontecer é necessário obter-se algumas informações do utilizador que a está a criar para se poder criar o nome da DGT. Um nota importante é que apenas utilizadores pertencentes ao grupo de *Data Officers* é que podem criar novas DGTs, e os mesmo são também o corpo responsável pelas mesmas.

As informações necessárias para a criação de uma nova DGT são:

- Nome da iniciativa - A iniciativa engloba as DGTs dentro de grupos, podendo o mesmo ser um tópico, um cliente, um assunto ou uma área de negócio. Caso ainda não exista uma iniciativa, o mesmo deverá ser notificado ao utilizador;
- O sistema de onde os dados presentes nesta DGT originam;

- O Dataset Layer Type em que a DGT se encontra. Para esta seleção existe um número de elementos para se escolher, nomeadamente se origina de tabelas refinadas, aumentadas ou refinadas e aumentadas;
- Tipo de modelo da DGT.

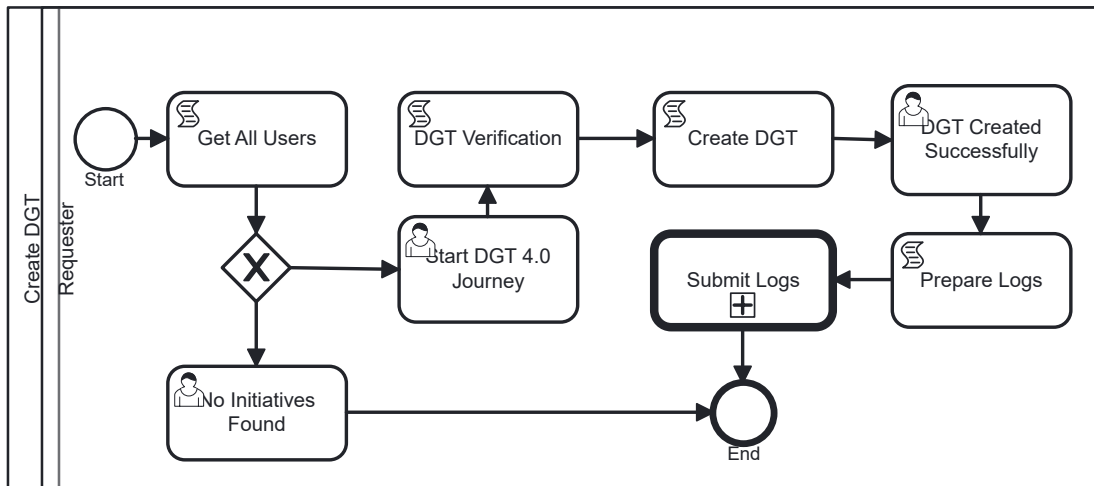


Figura 5.8: Workflow de Criação de DGTs

5.2.6.2 Criação de Iniciativas

Tal como discutido anteriormente, iniciativas são onde as DGTs são agrupadas e como tal é necessário que primeiro elas existam. Deste modo é necessário existir maneira para os elementos do grupo *Data Officers* as criarem.

Para se criar uma iniciativa, é necessário que o utilizador preencha o nome e o id da mesma, não podendo haver duas com o mesmo nome e o mesmo id.

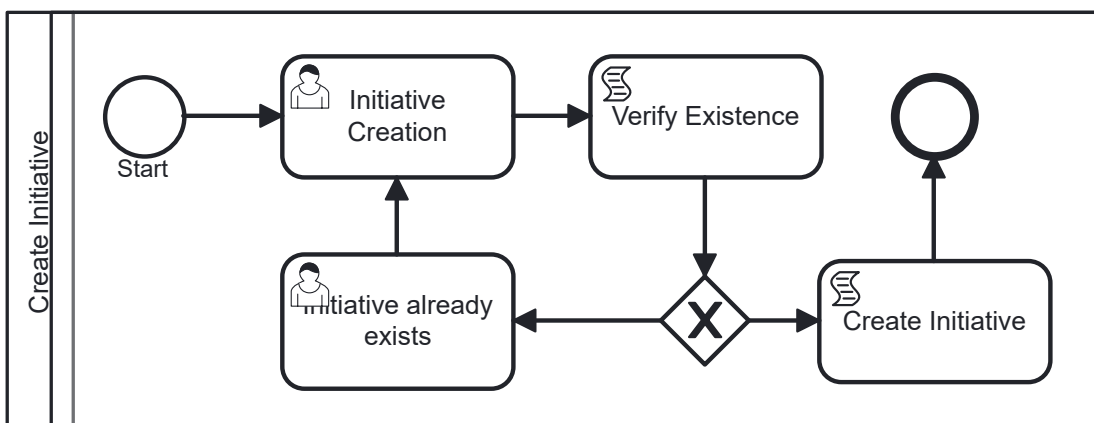


Figura 5.9: Workflow de Criação de Iniciativas

5.2.6.3 Ativação da Equipa

No desenvolvimento da DGT os ativos criados serão catalogados por várias pessoas de vários cargos, e para tal é necessário nomear quem vão ser essas pessoas.

Deste modo é necessário que a pessoa que criou a DGT consiga adicionar pessoas à mesma. Como tal o utilizador executa o workflow, que consiste em três passos: Selecionar a responsabilidade da pessoa; a pessoa a adicionar, sendo que existe a possibilidade destes dois passos entrarem em loop até que toda a equipa esteja selecionada; e, por último, finalizar o processo e ativar a equipa selecionada.

Os grupos possíveis são:

- **Business User** - Apresentado todos os utilizadores existentes na plataforma;
- **Data Steward** - Apresentado todos os elementos que são *Data Stewards* de data sets;
- **Data Custodian** - Apresentado todos os elementos que são *Data Custodians* de sistemas;
- **Data Officer** - Apresentado todos os elementos que pertencem ao grupo de Data Officer;
- **CoE DAAI** - Apresentado todos os elementos que pertencem ao grupo de CoE DAAI;
- **Requester** - Apresentado todos os utilizadores existentes na plataforma.

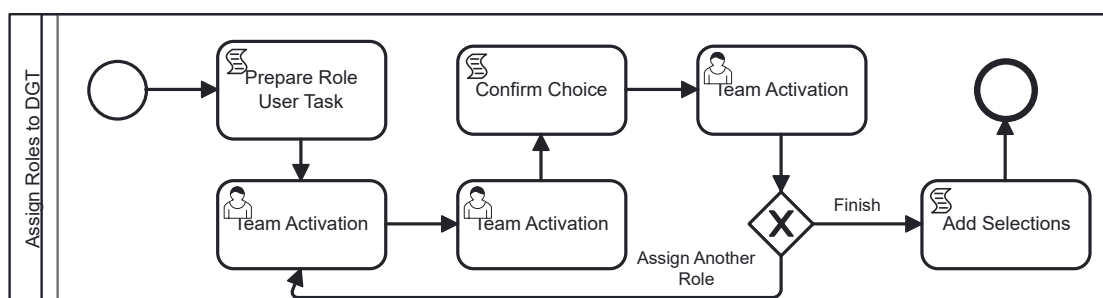


Figura 5.10: Workflow de Ativação da Equipa

5.2.6.4 Abertura e Fecho de Secções

Como foi possível observar-se há secções dentro de diferentes fases que têm nomes semelhantes, isto é devido ao facto de as mesmas representarem os mesmo ativos, nomeadamente as secções com *Data Model* e *Data Model Detail* no nome.

No entanto as mesmas são catalogadas por pessoas de cargos diferentes, sendo que as secções de *Functional* são catalogadas pelos utilizadores do cargo de *Business User*, as secções de *Sources* são catalogadas por utilizadores do cargo de *Data Custodian* e as secções de *Destination* são catalogadas por utilizadores do cargo de *Data Officers*.

O objetivo deste workflow é possibilitar o utilizador que o acionou, de selecionar que secções da DGT é que pretende abrir, visto que primeiro a DGT começa com todas as secções fechadas, e para se avançar no processo é necessário abrir-se as secções que são para se catalogar naquele momento, visto que como existe uma ordem de catalogação das mesmas, estas não são abertas todas ao mesmo tempo.

A ordem utilizada é **Business Understanding**, seguida de **Data Understanding**, seguida de **Data Modeling** e por último **Validation**.

Sempre que uma secção é aberta, deve ser enviado uma notificação para todos os elementos responsáveis pela catalogação da secção em questão. Isto é, caso a seja aberta a secção *Functional (Data Model)*, o utilizador que tem a responsabilidade de *Business User* da DGT em questão deverá receber uma notificação a informá-lo de que necessita de catalogar aquela secção daquela DGT, isto pois, como já referido anteriormente, é o *Business User* que cataloga a fase **Business Understanding**, que é onde se encontra a secção *Functional (Data Model)*.

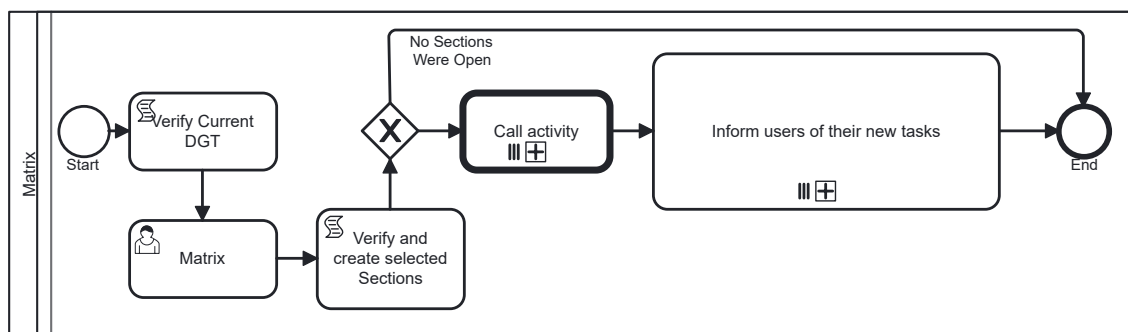


Figura 5.11: Workflow da Abertura e Fecho de Secções

5.2.6.5 Criação de Tabelas

As secções que têm no nome *Data Model* são representadas através de ativos do tipo "DGT Table", e são criadas através do workflow apresentado na figura 5.12. O utilizador que pretende criar tabelas, deve poder criar múltiplas ao mesmo tempo, e não deve existir tabelas com o mesmo nome.

A criação de ativos para as secções *Functional (Queries)*, *Sources (Queries)*, *Sources (Samples)*, *Source Auxiliary Tables*, *Quality Controls (Catalogue)*, *Acceptance* e *Tamr* é igual ao processo de criação de tabelas, por essa razão não será apresentado o workflow das mesmas.

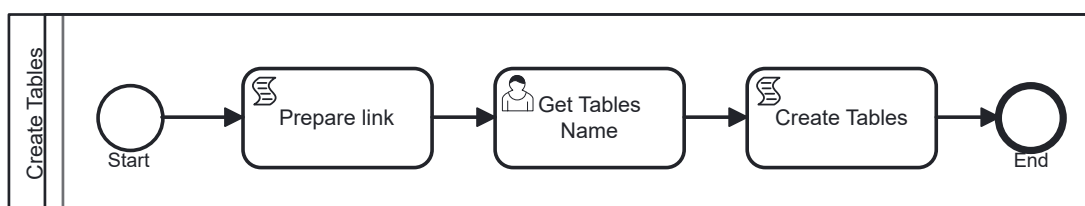


Figura 5.12: Workflow de Criação de Tabelas

5.2.6.6 Criação de Colunas

As secções que têm no nome *Data Model Detail* são representadas através de ativos do tipo "DGT Column", e são criadas através do workflow apresentado na figura 5.13. O utilizador que pretende criar colunas, deve poder criar múltiplas ao mesmo tempo, não deve existir tabelas com o mesmo nome e todas as colunas que o mesmo crie devem estar ligadas a uma

tabela da secção anterior, como tal é necessário primeiro criar-se tabelas e só depois é que deve ser possível criar-se colunas.

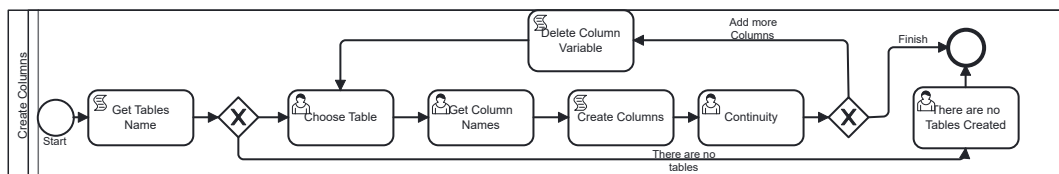


Figura 5.13: Workflow de Criação de Colunas

5.2.6.7 Adição de utilizadores à Data Access

Na secção de *Data Access* em vez de serem representados ativos, são adicionadas pessoas, como tal, é necessário que a pessoa que ative o workflow, consiga adicionar o número de pessoas que pretender de forma eficiente.

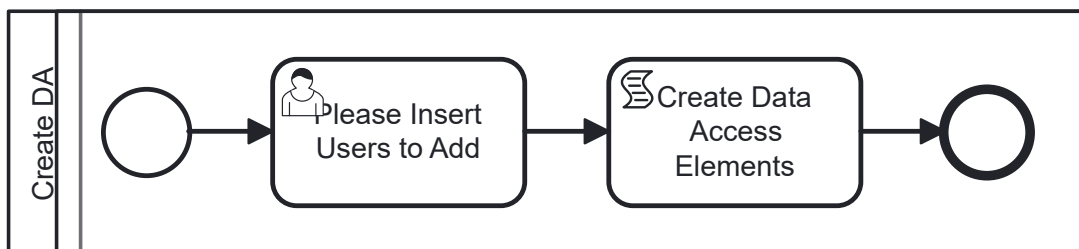


Figura 5.14: Workflow de Adição de Utilizadores em Data Access

5.2.6.8 Importação de Secções de outras DGTs

Ocasionalmente existem DGTs que têm secções semelhantes ou iguais a secções de outras DGTs, deste modo, o workflow representado na figura 6.53 permite ao utilizador, numa dada secção, de uma dada DGT, seleccionar uma outra DGT já finalizada e importar os dados de uma secção, para a secção correspondente da DGT atual.

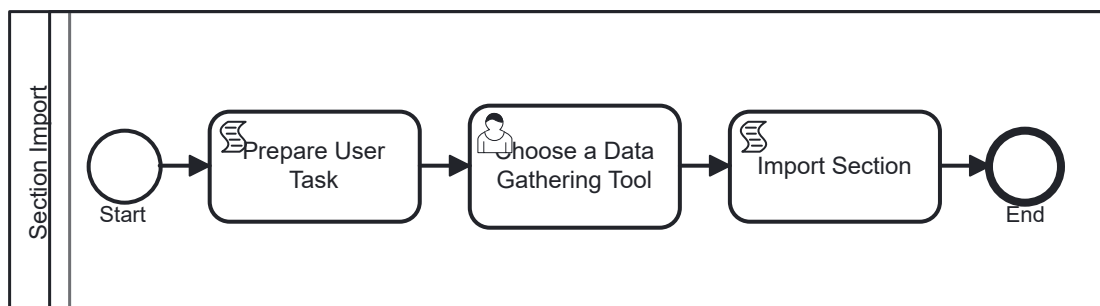


Figura 5.15: Workflow de Importação de Secções de outras DGTs

Capítulo 6

Implementação

O objetivo deste capítulo é apresentar e descrever como a solução escolhida para este projeto no capítulo 5 foi implementada. Todo o desenvolvimento foi predominantemente feito utilizando a metodologia BPMN, com o uso da linguagem groovy para o desenvolvimento dos workflows, sendo tudo desenvolvido dentro da plataforma Collibra.

6.1 Workflows BPMN

Nesta secção, é apresentada a implementação utilizando workflows BPMN. É explorado como é que os workflows foram aplicados para automatizar e otimizar os processos de negócio do cliente.

De modo a se entender melhor como funciona a criação de workflows, é dada uma breve explicação sobre a criação de formulários nas tasks, o preenchimento do formulário para o envio de email e um excerto de código para se melhor perceber o funcionamento da Collibra API.

Na figura 6.1 é possível observar-se a criação de um formulário, que será posteriormente apresentado ao utilizador no decorrer da execução do workflow em causa, o exemplo fornecido é referente ao workflow do "Lineage Path", mais propriamente aos passos presentes nas figuras 6.13, 6.14 e 6.15, uma vez que com a utilização de uma variável, tal como a "checkBoxOptions" presente na coluna denominada de "Form Values" do tipo "dynamicCheckBox", é possível reutilizar-se formulários, sendo apenas necessário alterar-se os valores da variável utilizada.

É também possível observar-se as variáveis denominadas "Finish Selection" e "Import Everything", que ao observar as figuras referidas anteriormente, determina-se que as mesmas se referem aos botões presentes nas imagens, tal como indica o tipo das variáveis que é "button".

General	Form properties:							
Main config								
Documentation								
Form	Id	Name	Type	Default	Required	Readable	Writeable	Form values
Listeners	results	Results	dynamicCheckbox		false	true	true	dynamicValuesExpression:\${checkBoxOptions}
Multi instance	finishSelection	Finish Selection	button	false	false	true	true	
	selectAll	Import Everything	button	false	false	true	true	

Figura 6.1: Variáveis do Formulário

Ao longo deste documento é referido um envio de emails. Esta ação é concretizada preenchendo-se o formulário presente na figura 6.2, onde é necessário, ou pelo menos aconselhado, preencher-se os campos referentes a quem se quer enviar os emails, o assunto e o corpo do email.

Como é possível observar-se na imagem, existe a possibilidade de se preencher estes campos com variáveis, que são populadas dentro do código, possibilitando assim, emails dinâmicos dependendo das situações.

Figura 6.2: Formulário Email

De forma a se estabelecer uma relação visual com a Collibra API, é apresentado o excerto de código presente na figura 6.3, onde é observável uma função que recebe como parâmetros o id de um domínio, o nome do ativo e o tipo do mesmo e com estas variáveis é criado o ativo (asset) da forma apresentada na figura.

```

1538 // Create asset
1539 def createAsset(domain, name, type){
1540     asset = assetApi.addAsset(AddAssetRequest.builder()
1541         .domainId(string2Uuid(domain.toString()))
1542         .name(name)
1543         .typeId(string2Uuid(type))
1544         .build()
1545     )
1546     return asset
1547 }
1548

```

Figura 6.3: Excerto de Código - Criação de Ativos

6.1.1 Ativação Data Owner

Este workflow apenas é disponibilizado a *Data Officers*, e os mesmo têm de se encontrar num ativo do tipo data domain que possua o status a "Entity Associated", o status serve para se conseguir distinguir quais são os domínios que devem ter *Data Owners* e quais não, quer por não representarem domínios ou por ainda não estarem prontos para tal.

No caso de ativação de *Data Steward* o tipo de ativo é um data set e no caso do *Data Custodian* é num ativo do tipo sistema.

Para ativar o workflow, é necessário carregar no botão *Activate Data Owner* como é observável na imagem 6.4.

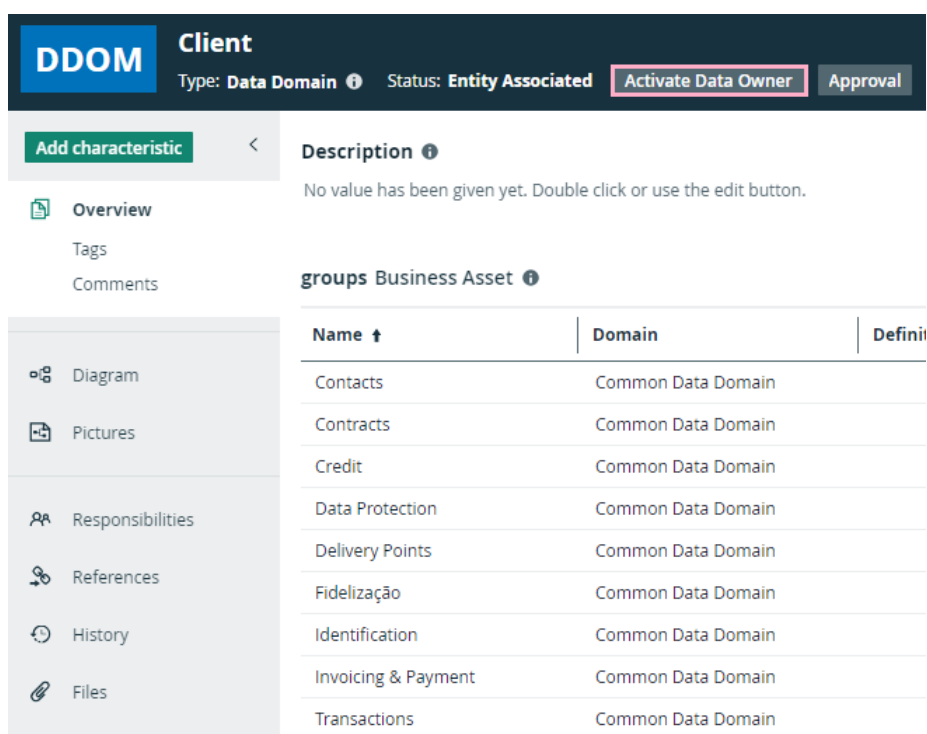


Figura 6.4: Ativação Data Owner - Botão de execução

Após executar o workflow, será enviado uma task para o *Data Officer*, que deverá escolher o utilizador a quem será atribuída a responsabilidade de *Data Owner* para o data domain selecionado. Nesta etapa, o utilizador que interagir com o workflow terá de fazer upload de um ficheiro, referente à ata da reunião de ativação do *Data Owner*, onde foi previamente discutido quem é que deveria ser selecionado para *Data Owner* do data domain em questão. A imagem 6.5 mostra o que o *Data Officer* que acionou o processo visualizará no Collibra:

Due: 7/28/2023

Choose Data Owner

Related to

DDOM Client

Please choose a user to assign the Data Owner role to.

Select ▼

Inserted File

Click here or drag a file to upload

Submit

Figura 6.5: Ativação Data Owner - Seleção do utilizador

O utilizador selecionado pelo *Data Officer*, recebe uma task do Collibra informando que o

mesmo foi selecionado para ser o *Data Owner* de um determinado data domain, juntamente com um link para um ficheiro que contém informações sobre a nova responsabilidade e iniciativas associadas à função. Após a leitura, o utilizador deve aceitar ou rejeitar o pedido de ativação, como se pode verificar na figura 6.6;

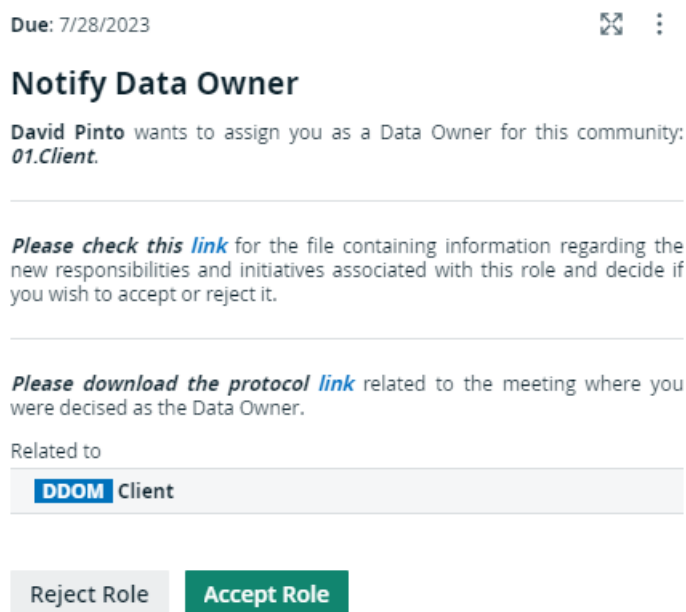


Figura 6.6: Ativação Data Owner - Escolha do utilizador

Caso o utilizador aceite o cargo, todos os elementos do grupo *Data Officer* do Colibra serão notificados, quer via task no Colibra como é observável na figura 6.7, quer via email como se pode verificar na figura 6.8 e o workflow termina;

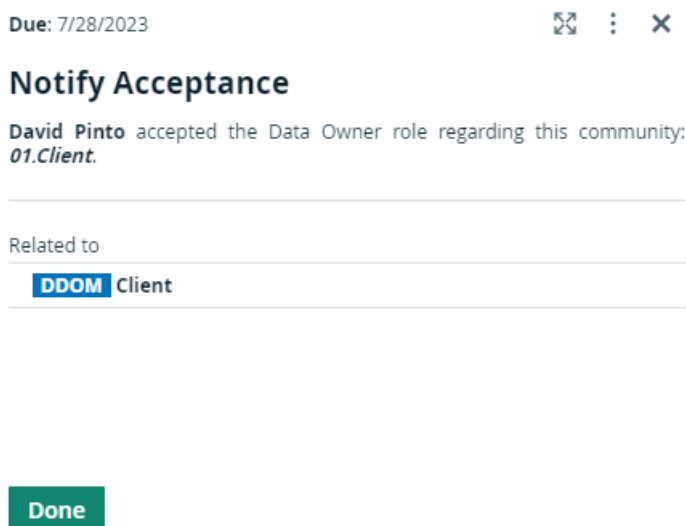


Figura 6.7: Ativação Data Owner - Notificação aceitação

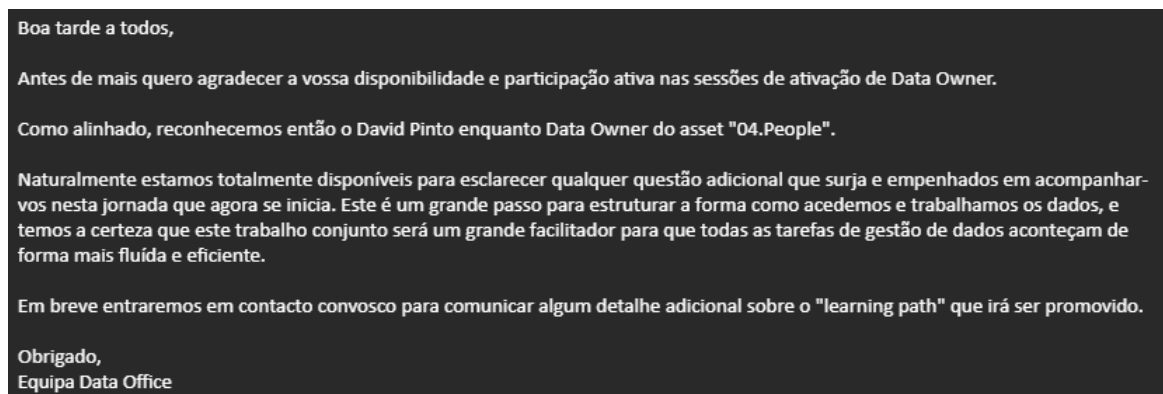


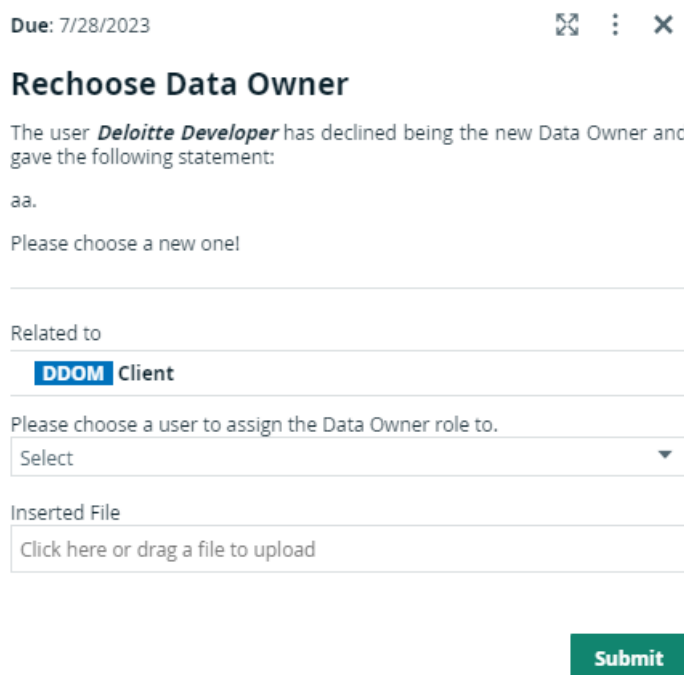
Figura 6.8: Ativação Data Owner - Email de notificação

Caso o utilizador rejeite o cargo, ser-lhe-á enviado uma nova task a pedir uma justificação para a sua rejeição, como é observável na figura 6.9.

A screenshot of a web form titled "Rejection Justification". At the top left, it says "Due: 7/28/2023" with a calendar icon. Below the title, there is a text prompt: "Please justify why you rejected the Data Owner role regarding this community: **01.Client**." followed by another prompt: "Please recommend a new person for the position." Below these prompts is a section labeled "Related to" containing a button labeled "DDOM Client". Underneath is a "Justification" section with a rich text editor toolbar showing options for Paragraph, Bold, Italic, Underline, Strikethrough, Text Color, and Background Color. The editor area is empty. At the bottom right of the form is a green "Submit" button.

Figura 6.9: Ativação Data Owner - Pedido justificação

Quando a task anterior é finalizada, a justificação é enviada para o *Activation Manager*, que por sua vez, terá de selecionar um novo elemento para o cargo de *Data Owner*, tal como enviar a ata da reunião onde o novo elemento foi escolhido, como é observável na figura 6.10.



Due: 7/28/2023

Rechoose Data Owner

The user **Deloitte Developer** has declined being the new Data Owner and gave the following statement:

aa.

Please choose a new one!

Related to

DDOM Client

Please choose a user to assign the Data Owner role to.

Select

Inserted File

Click here or drag a file to upload

Submit

Figura 6.10: Ativação Data Owner - Seleção novo utilizador

As últimas três tasks, formam um loop, até que um utilizador aceite o cargo.

Para além de todas as notificações e task enviadas para os envolvidos no processo, em paralelo com o processo, estará um loop que envia emails. O primeiro email é enviado ao quinto dia do processo com o status do mesmo para o Activation Manager, ao nono dia, será enviado um email com o status do processo para o Activation Manager e para o elemento responsável pela task na qual o processo se encontra. Após o décimo dia, é enviado um email a cada cinco dias para o Activation Manager com o status do processo.

Este loop de email, é um requisito comum aos três processo de ativação, *Data Owner*, *Data Custodian* e *Data Steward*, sendo que está implementada de igual forma nos três workflows.

Tal como discutido na secção 5, os processos de ativação das três responsabilidades são iguais, alterando apenas a responsabilidade que é ativada, no entanto as pessoas envolvidas no processo, para além do utilizador que for escolhido para a ativação, são as mesmas. Como tal não existe a necessidade de se repetir informação.

6.1.2 Carregamento em Massa

Apenas elementos do grupo *Data Office* é que conseguem carregar em massa *Data Owners*, *Stewards* e *Custodians*, e para tal é necessário o mesmo encontrar-se na vista global observável na figura 6.11. Nesta vista pode-se observar quem é o *Data Owner*, *Data Custodian* e *Data Steward* de um ativo específico, qual é o tipo do ativo e quando é que a ativação foi efetuada.

The screenshot shows the 'Activated DOS' interface with a table of data assets. The table has columns for Data Owner, Data Steward, Data Custodian, Asset #, Asset Type, and DOS Activation Date. The interface includes a sidebar with filters and a top bar with a 'Revert to original' button.

Data Owner	Data Steward	Data Custodian	Asset #	Asset Type	DOS Activation Date
			SmartMeters ES	System	
		Business User DGT	Scriptor	System	2023-02-13 12:53:17
			SCADA	System	
			SAP SuccessFactors	System	2023-06-08 13:48:08
			SAP S4	System	
		Business User DGT	SAP R3	System	2023-02-14 16:21:29
			SAP BW54	System	
			SAP BW4	System	
			SAP BW3	System	
		David Pinto	Salesforce	System	2023-06-08 13:47:04
			SABi	System	
			S4G	System	
		David Pinto	rft_cm_obj_status_j	Data Set	2023-01-03 16:15:06
			rft_cm_obj_j	Data Set	
			rft_cm_voucher	Data Set	
			rft_cm_ave_load_mri	Data Set	
			rft_cm_ave_load	Data Set	
			rft_cm_glan_contern	Data Set	
			rft_cm_glan_char	Data Set	
		David Pinto	rft_cm_glan	Data Set	2023-05-24 09:02:00
			rft_cm_glow	Data Set	
			rft_cm_login_authorization	Data Set	
			rft_cm_cspoc	Data Set	

Figura 6.11: Ativação em Massa - Vista Global

Para se ativar em massa é necessário então selecionar-se qual o utilizador pretendido, na responsabilidade desejada, como exemplo temos o caso da segunda linha, em que o *Business User DGT* foi selecionado para ser o *Data Custodian* do ativo denominado "Scriptor", que por sua vez é um sistema, e consegue-se observar que esta ativação foi efetuada no dia 13/02/2023.

A ativação em si, é feita em background através dos workflows discutidos na secção 5 e observáveis nas figuras 5.4 e 5.5.

Estes workflows são acionados sempre que um utilizador for selecionado para cada um das três responsabilidades, sendo que, dependendo da responsabilidade, apenas o workflow destinado a essa responsabilidade é que é iniciado. Após iniciado os workflows fazem as validações do tipo de ativo, pois, como foi falado previamente, *Data Owners* são ativos apenas em ativos do tipo data domain, *Data Custodians* são ativados em ativos do tipo sistema e *Data Stewards* em ativos do tipo data set.

Para além da verificação se o tipo de ativo corresponde à responsabilidade, o workflow de *Data Custodian* propaga a responsabilidade para as bases de dados oriundas desse sistema, uma vez que está definido de que o *Data Custodian* de um sistema é também o *Data Custodian* das bases de dados desse sistema, e o workflow do *Data Custodian* propaga a responsabilidade para a tabela do onde esse data set originou.

6.1.3 Lineage Path

Como discutido no capítulo 5, o objetivo deste requisito é fornecer todos os atributos possíveis, provenientes de colunas relacionadas com as colunas de um certo data set, ao *Data Steward* do mesmo.

Quando essa seleção for finalizada, os dados selecionados são importados para a coluna do data set selecionado e é enviado um email do relatório com todos os atributos que o *Data Steward* optou por importar e para qual das diversas colunas foram importados, como se pode verificar na figura 6.16. Outro requisito, visível na imagem através do nomes com

a cor roxa, é a necessidade de existir um link para o ativo para onde os atributos foram importados.

Para se executar o workflow, é necessário que o utilizador seja um *Data Steward* e que se encontre num data set, conforme observável na figura 6.12.

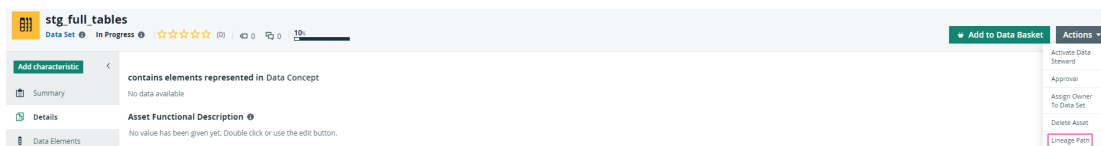


Figura 6.12: Lineage Path - Botão

Após iniciar o workflow, é despoletado a task presente na figura 6.13 para o *Data Steward* do data set onde o workflow foi iniciado. Nesta task o utilizador necessita de selecionar quais os atributos, status e relações é que o mesmo pretende trazer para a coluna diretamente relacionada com o data set. Caso pretenda, o mesmo pode selecionar todos.

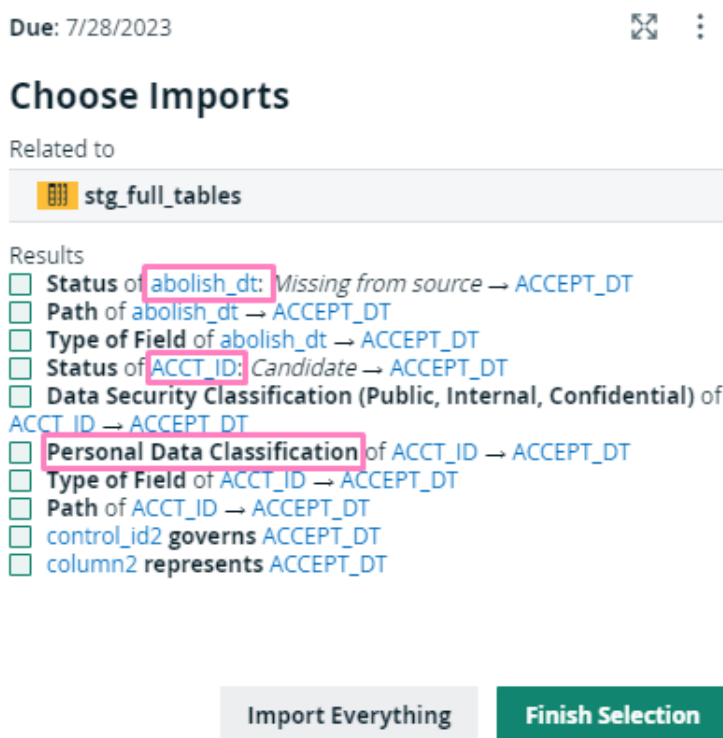


Figura 6.13: Lineage Path - Escolha de Atributos

Após a task anterior existem três caminhos, um para o caso de terem sido selecionados atributos ou status duplicados, outro para o caso de ter sido selecionado o atributo de dado pessoal e o terceiro para a situação em que nenhum dos dois casos anteriores aconteceu, sendo que nesta situação, o processo é terminado e para além de serem importados os atributos selecionados, é enviado um email com o relatório do que foi importado para as partes interessadas, conforme observável na figura 6.16.

Caso tenham sido selecionados atributos ou status repetidos, é despoletada a task presente na figura 6.14, para que o utilizador selecione apenas um valor para cada atributo, devido ao facto de que os atributos e status apenas podem ter um valor.

Due: 7/28/2023

Select Non-Repeated

Related to

stg_full_tables

Results

- ☐ Status of abolish_dt: Missing from source → ACCEPT_DT
- ☒ Status of ACCT_ID: Candidate → ACCEPT_DT
- ☒ Path of abolish_dt → ACCEPT_DT
- ☐ Path of ACCT_ID → ACCEPT_DT
- ☒ Type of Field of abolish_dt → ACCEPT_DT
- ☐ Type of Field of ACCT_ID → ACCEPT_DT

Figura 6.14: Lineage Path - Escolha de Não Repetidos

Caso tenha sido selecionado o atributo referente aos dados pessoais, é então despoletado a task presente na figura 6.15, destinada ao *Data Steward* da coluna que possui este valor, ao contrário das tasks anteriores, que eram destinadas ao *Data Steward* do data set em questão, para que o mesmo decida se a coluna do data set é também dado pessoal ou não.

Due: 7/28/2023

Personal Data Clarification

Related to

stg_full_tables

Choose columns which have personal data:

- ☐ Personally Identifiable Information of ACCT_ID → ACCEPT_DT

Figura 6.15: Lineage Path - Escolha de Dados Pessoais

No final do processo, como referido anteriormente é enviado o relatório presente na figura 6.16.

To whom it may concern,

On the process of the Lineage Path concerning the Column [Client](#), the following characteristics were imported:

Characteristics imported per each column of the given asset:	
Column Imported To	Characteristics
Person	Maximum Number of Characters (Attribute)
Person	Number of Decimal Places (Attribute)
Person	Path (Attribute)
Country	Mandatory (Attribute)
Country	is part of CM_T_TPA_OBJ_RES (Relation)

Thank you,
David Pinto

Figura 6.16: Relatório Final

6.1.4 Apresentação de logs

A solução de apresentação de logs do Collibra depende da consola do Collibra, que é uma instância separada da do cliente. O acesso à mesma deve ser controlado, dado o carácter sensível das informações que circulam nesta interface.

Deste modo, e com especial foco na instância Collibra do cliente, foi desenvolvido um repositório central de registo de informação. Este visa acomodar aspetos cruciais de implementações personalizadas feitas ao longo do tempo. A solução visa ser agnóstica em relação aos desenvolvimentos anteriores, dando espaço para interagir com informações que fluem de empreendimentos futuros.

Os workflows que devem abastecer este repositório central são:

- Ativação *Data Owner*;
- Ativação *Data Steward*;
- Ativação *Data Custodian*;
- Lineage Path;
- Data Set Management;

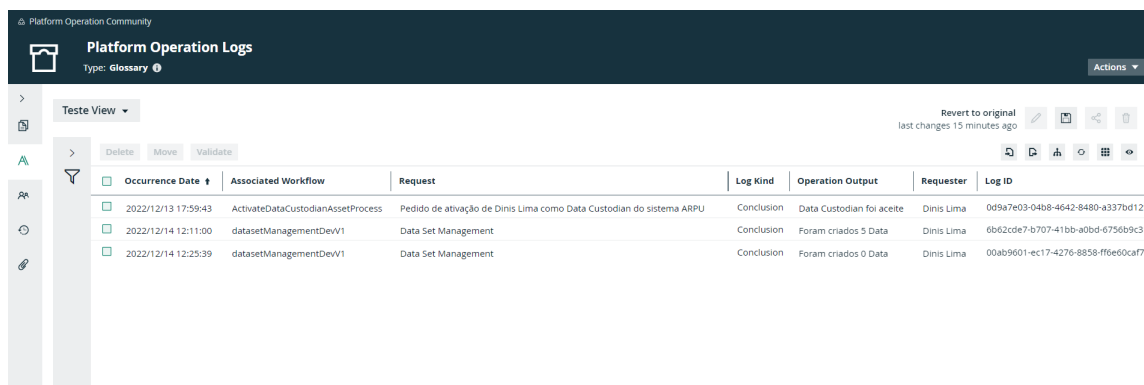
Como foi possível observar-se no capítulo 5, existem vários workflows com referência a logs, essas tasks nos workflows presentes nesse capítulo estão relacionados com a implementação deste requisito.

Este processo funciona utilizando uma *Call Activity*, que é um elemento utilizado para representar a chamada de um subprocesso num processo maior. Esta atividade é usada quando um subprocesso já foi implementado separadamente e precisa de ser incluído num outro workflow. Ao executar o processo, quando a *Call Activity* é alcançada, o controlo é transferido para o subprocesso correspondente. O subprocesso é executado independentemente do processo principal e, quando concluído, o controlo regressa para o workflow principal.

Na solução, a *Call Activity* é o workflow que gera os logs, que é incluído nos workflows de ativação *Data Owner*, ativação *Data Steward*, ativação *Data Custodian*, lineage path, data set management e criação de DGTs, como é possível observar-se na figuras 5.2, 5.3, 5.6 e 5.8 do capítulo 5.

Para permitir aos utilizadores uma observação dos logs dentro da plataforma Collibra, foi criado uma vista onde os utilizados conseguem observar todos os logs, como é possível visualizar-se na figura 6.17.

Esta vista é populada automaticamente através das task referidas anteriormente com informações provenientes do workflow que foi executado, fornecendo dados como, a data, que workflow é que foi executado, onde foi executado, a conclusão do mesmo, isto é, se executou até ao fim ou se ocorreu um erro no processo, o resultado do workflow, quem é que o executou e o Id do log em si.



Occurrence Date	Associated Workflow	Request	Log Kind	Operation Output	Requester	Log ID
2022/12/13 17:59:43	ActivateDataCustodianAssetProcess	Pedido de ativação de Dinis Lima como Data Custodian do sistema ARPU	Conclusion	Data Custodian foi aceite	Dinis Lima	0d9a7e03-0408-4642-8480-a337bd12f
2022/12/14 12:11:00	datasetManagementDevV1	Data Set Management	Conclusion	Foram criados 5 Data	Dinis Lima	6b62cde7-b707-41bb-a0bd-6756b9c31
2022/12/14 12:25:39	datasetManagementDevV1	Data Set Management	Conclusion	Foram criados 0 Data	Dinis Lima	00ab9601-ec17-4276-8858-ffe60ca7f

Figura 6.17: Vista dos Logs

6.1.5 Data Set Management

O processo de pedido de acesso a ativos em Collibra é projetado para garantir a segurança, a governação e o controlo adequados sobre os dados e recursos geridos pela plataforma. O mesmo ajuda a garantir que apenas os utilizadores autorizados têm acesso aos ativos apropriados, de acordo com as políticas e regulamentações da organização.

Em Collibra, pedir acesso a ativos refere-se ao processo de pedir permissão para aceder e utilizar determinados ativos de dados. Estes ativos apenas podem ser do tipo data set, esta é uma política da empresa que já se encontrava definida previamente ao início deste projeto. No entanto o objeto é que só deve ser possível pedir-se acesso a "tabelas" ingeridas através do *Data Lake* do cliente.

Deste modo utilizaram-se data sets para representar essas tabelas, fazendo então com que só fosse possível pedir-se acesso a data sets, pois os mesmos apenas representam tabelas oriundas do *Data Lake* do cliente.

Deste modo foi necessário implementar-se uma maneira de se criar data sets automaticamente sempre que uma nova tabela era ingerida através do *data lake*. E é isso que, o workflow discutido no capítulo 5 e observável na figura 5.6, efetua.

Este é executado diariamente, procura por tabelas que tenham sido ingeridas através do *Data Lake* e verifica se existe um data set com o mesmo nome, caso não haja, é criado um, juntamente com uma relação para a tabela, de forma a representar os atributos da mesma.

Na figura 6.18 é possível observar-se um data set, que foi criado a partir da tabela observável na figura 6.19, tal como a relação entre data set e tabela, denominada de "is clone of Table", e a partir dessa relação consegue-se observar múltiplos atributos da tabela.

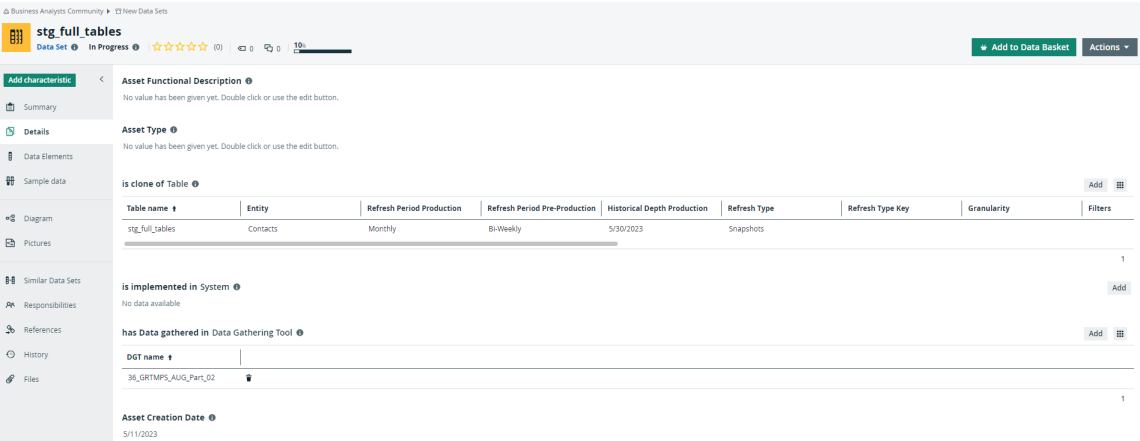


Figura 6.18: Imagem data set

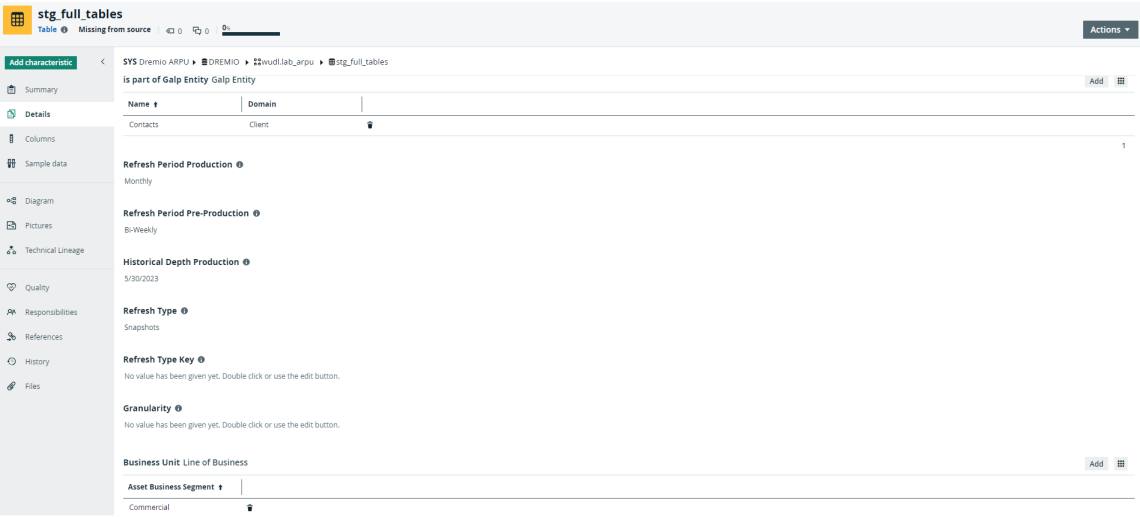


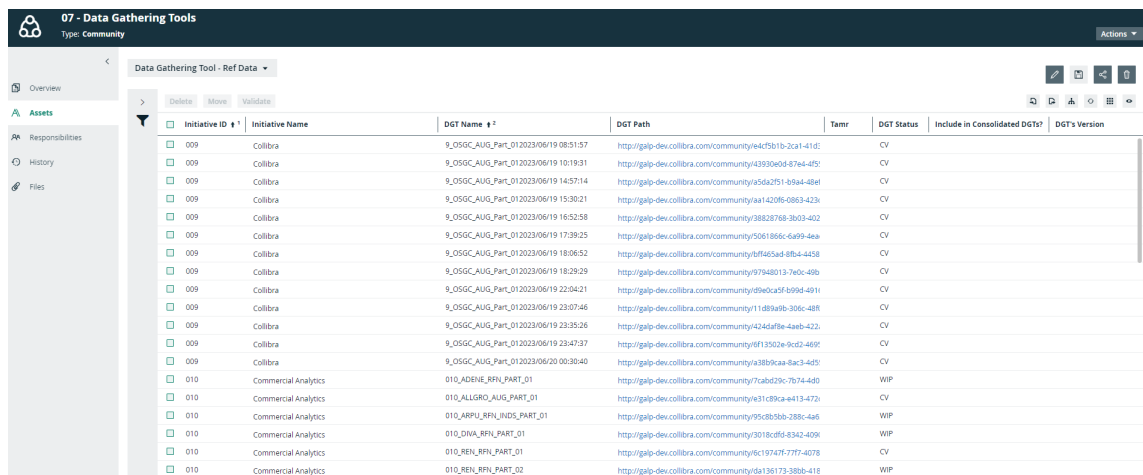
Figura 6.19: Imagem tabela

6.1.6 Data Gathering Tool

No contexto do cliente, uma DGT é uma ferramenta utilizada para recolher e reunir dados de várias fontes. Criando-se novos ativos a partir dessa recolha, para que os mesmo possam ser utilizados em contextos de negócio.

6.1.6.1 Criação da DGT

Para se executar este workflow é necessário que o utilizador seja um elemento do *Data Office* e que se encontre na comunidade das DGTs, esta é observável na figura 6.20, pois apenas aí é que o botão se encontra visível, como se pode verificar na figura 6.21.



Initiative ID	Initiative Name	DGT Name	DGT Path	Tamr	DGT Status	Include in Consolidated DGTs	DGT's Version
009	Collibra	9_OSGC_AUG_Part_012023/06/19 08:51:57	http://galp-dev.collibra.com/community/e4c5b1b-3ca1-41e6		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 10:19:31	http://galp-dev.collibra.com/community/4390a6d-87e4-4f5c		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 14:57:14	http://galp-dev.collibra.com/community/a5da2751-89a4-48ef		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 15:30:21	http://galp-dev.collibra.com/community/a1420f6-0863-423c		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 16:52:58	http://galp-dev.collibra.com/community/38828768-3b03-402c		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 17:39:25	http://galp-dev.collibra.com/community/5061866c-6a99-4eaa		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 18:06:52	http://galp-dev.collibra.com/community/bf465ad-8fb-4-4d58		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 18:29:29	http://galp-dev.collibra.com/community/97948013-7e0c-4db8		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 22:04:21	http://galp-dev.collibra.com/community/d9e0c5f-899d-4891		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 23:07:46	http://galp-dev.collibra.com/community/11d89fb-306c-488f		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 23:35:26	http://galp-dev.collibra.com/community/424da78e-4aeb-422c		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/19 23:47:37	http://galp-dev.collibra.com/community/6f13502e-9cd3-469f		CV		
009	Collibra	9_OSGC_AUG_Part_012023/06/20 00:30:40	http://galp-dev.collibra.com/community/a38b0caa-8ac3-4d5f		CV		
010	Commercial Analytics	010_ADENIE_RFH_PART_01	http://galp-dev.collibra.com/community/7cab29c-7b74-4d89		WIP		
010	Commercial Analytics	010_ALLGRO_AUG_PART_01	http://galp-dev.collibra.com/community/ie31c9ca-413-472f		CV		
010	Commercial Analytics	010_ARPU_RFH_IND5_PART_01	http://galp-dev.collibra.com/community/95c8b5b-288c-4a6f		WIP		
010	Commercial Analytics	010_DIVA_RFH_PART_01	http://galp-dev.collibra.com/community/3018c0ff-83d2-409f		WIP		
010	Commercial Analytics	010_REFL_RFH_PART_01	http://galp-dev.collibra.com/community/6c19747f-7707-4078		CV		
010	Commercial Analytics	010_REFL_RFH_PART_02	http://galp-dev.collibra.com/community/d6136173-380b-41e		WIP		

Figura 6.20: DGT - Comunidade das DGTs



Figura 6.21: Criação da DGT - Botão

A primeira task, observável na figura 6.22, tem de propósito de obter o nome da DGT, pedindo ao utilizador que preencha os campos respetivos à seleção da Iniciativa, do Sistema, do Data set Layer Type e, opcionalmente, do Tipo de Modelo, aos quais a DGT estará associada. A Iniciativa deve ter sido previamente criada na Plataforma Collibra, utilizando o workflow Criação de Iniciativas, como se pode verificar na subsecção 6.1.6.2.

Caso não exista nenhuma iniciativa no Collibra, em vez de ser mostrada a task da figura 6.22, é mostrada uma notificação para o utilizador que iniciou o processo, a indicar de que não existe nenhuma iniciativa criada e que primeiro tem de ser criada uma iniciativa para englobar a DGT.

Due: 7/28/2023

Start DGT 4.0 Journey

Welcome **David Pinto**! This form triggers the creation of a Data Gathering Tool journey according to the 4.0 template. The fields which are **obligatory** are marked with an asterisk (*).

Related to

07 - Data Gathering Tools

Select the DGT Initiative: *

Select

Select the DGT System: *

Select

Select the DGT Dataset Layer Type: *

Select

Select the Model Type:

Select

Submit

Figura 6.22: Criação da DGT - Preenchimento dos Campos

Na situação em que todos os campos são seleccionados corretamente, e o botão de finalizar é premido, é apresentada então a task da figura 6.23 a informar o utilizador de que a DGT foi criada.

Due: 7/28/2023

DGT Created Successfully

Data Gathering Tool **009_E-REDES_RFN_ONDS_PART_01** has been created successfully.

Related to

07 - Data Gathering Tools

Finish Process

Figura 6.23: Criação da DGT - DGT Criada

6.1.6.2 Criação de Iniciativas

Tal como o caso anterior, para se executar este workflow é necessário que o utilizador seja um elemento do *Data Office* e que se encontre na comunidade das DGTs, esta é observável na figura 6.20, juntamente com o botão para se inicializar o workflow, como se pode verificar na figura 6.24.



Figura 6.24: Criação de Iniciativa - Botão

Semelhante à primeira task do workflow anterior, o utilizador necessita de preencher os campos respetivos à iniciativa, neste caso é o nome e o id da mesma, conforme visível na figura 6.25.

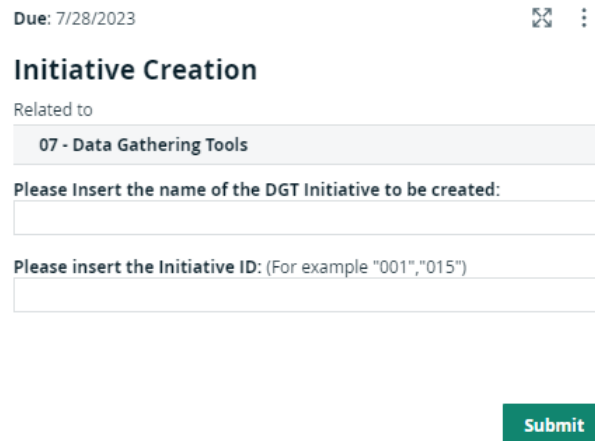


Figura 6.25: Criação de Iniciativa - Preenchimento dos Campos

No entanto, caso a iniciativa que o utilizador esteja a tentar criar já existir, isto é, se já houver uma iniciativa com o nome e o id escolhidos pelo utilizador, então é despoletado uma task para o mesmo, conforme a figura 6.26, a indicar que a iniciativa já existe. Quando o utilizar selecionar o botão "Done", o mesmo é redirecionado novamente para a task de preencher os campos.

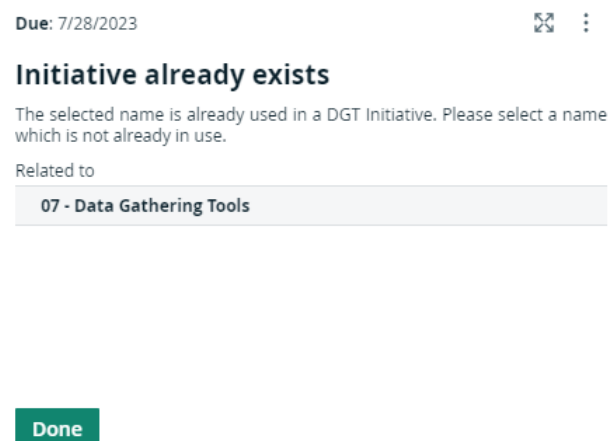


Figura 6.26: Criação de Iniciativa - Iniciativa já Existe

Caso não exista nenhuma iniciativa com os dados fornecidos, o processo é terminado e a mesma é criada e fica disponível na vista das iniciativas, observável na figura 6.27.

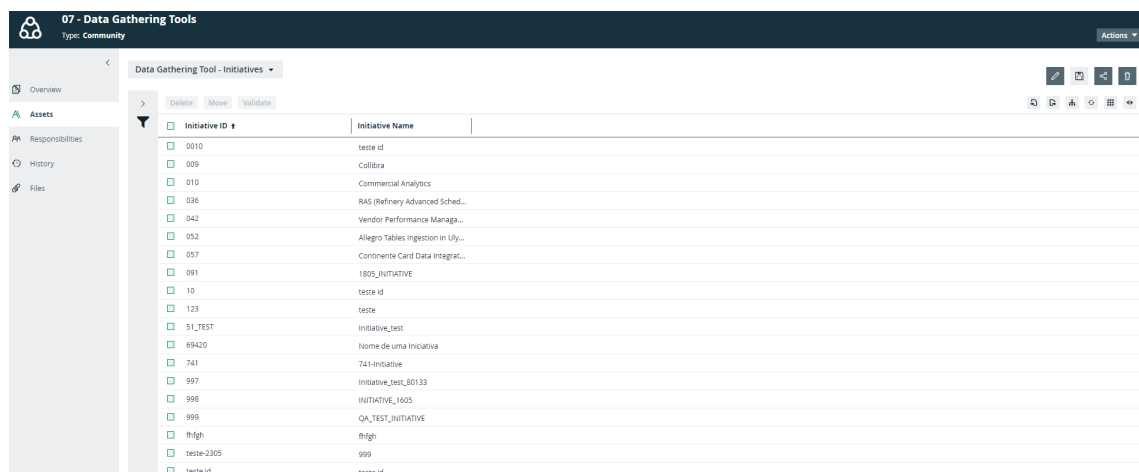


Figura 6.27: Criação de Iniciativa - Vista das Iniciativas

6.1.6.3 Ativação da Equipe

O processo de ativação da equipa permite ao *Data Officer* adicionar elementos à DGT, com o objetivo da mesma ser catalogada pela equipa escolhida. Deste modo, apenas *Data Officers* conseguem visualizar e, conseqüentemente, executar o workflow, no entanto, para tal os mesmos necessitam de se encontram na vista da DGT, mas com o ativo "Register" selecionado, conforme a figura 6.28.

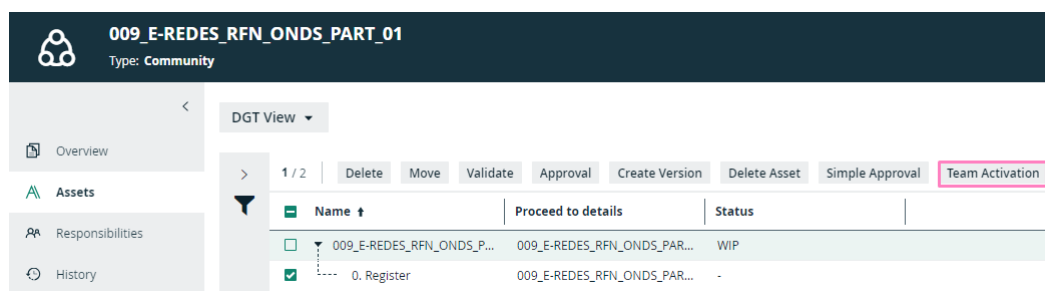
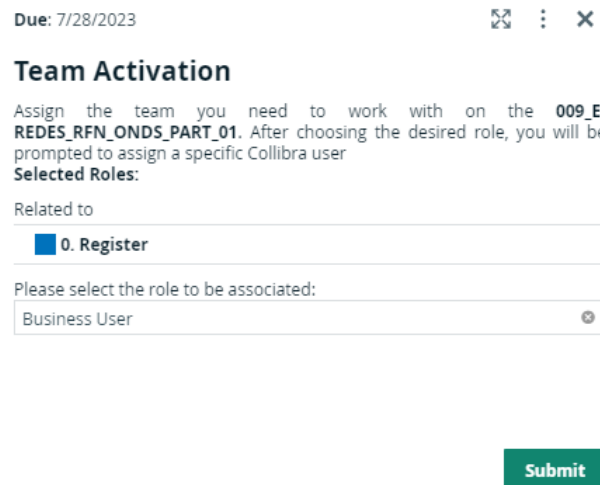


Figura 6.28: Ativação da Equipe - Botão

Após iniciação do workflow, o utilizador que o ativar recebe a task presente na figura 6.29, onde o mesmo tem de selecionar a responsabilidade que pretende adicionar, esta responsabilidade pode ser *Data Steward*, *Data Officer*, *Data Custodian*, *CoE DAAI* ou *Business User*.



Due: 7/28/2023

Team Activation

Assign the team you need to work with on the **009_E-REDES_RFN_ONDS_PART_01**. After choosing the desired role, you will be prompted to assign a specific Collibra user

Selected Roles:

Related to

☒ 0. Register

Please select the role to be associated:

Business User

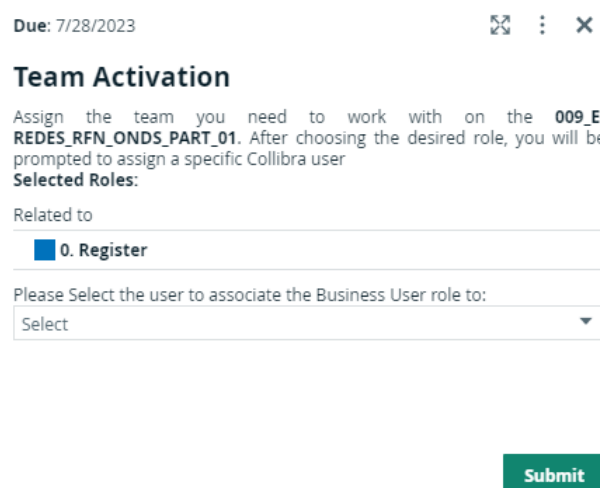
Submit

Figura 6.29: Ativação da Equipa - Seleção da Responsabilidade

Após a escolha da responsabilidade, é necessário seleccionar a quem atribuir a responsabilidade, para tal, conforme observável na figura 6.30, o utilizador necessita, de forma semelhante à task anterior, de seleccionar o elemento que pretende adicionar à responsabilidade.

Estes dois passos, podiam-se encontrar na mesma task, o que faria com que o workflow fosse mais eficiente e intuitivo, no entanto, um dos requisitos para esta etapa é que apenas fosse disponibilizado ao utilizador que a preenche, os utilizadores pertencentes à responsabilidade seleccionada. Por outras palavras, se fosse seleccionada a responsabilidade *Data Officer*, na task de selecção do utilizador, apenas deveriam estar disponíveis os elementos pertencentes ao grupo *Data Office*.

Consequentemente, para se implementar este requisito, foi necessário separar-se em duas etapas, sendo esta a única solução encontrada para se disponibilizar apenas os elementos pertencentes à responsabilidade seleccionada.



Due: 7/28/2023

Team Activation

Assign the team you need to work with on the **009_E-REDES_RFN_ONDS_PART_01**. After choosing the desired role, you will be prompted to assign a specific Collibra user

Selected Roles:

Related to

☒ 0. Register

Please Select the user to associate the Business User role to:

Select

Submit

Figura 6.30: Ativação da Equipa - Escolha do Utilizador

A partir da última task, o processo torna-se num loop, até que o utilizador decida que a sua equipa já se encontra finalizada, conforme observável na figura 6.31, é possível finalizar-se o processo ou continuar o mesmo, selecionando uma nova responsabilidade e um novo elemento. Ademais, como se pode verificar no retângulo cor de rosa, todos os elementos adicionados são apresentados nesta etapa, para que o utilizador que está a executar o workflow, não se esqueça quais elementos é que já foram adicionados e quais faltam adicionar.

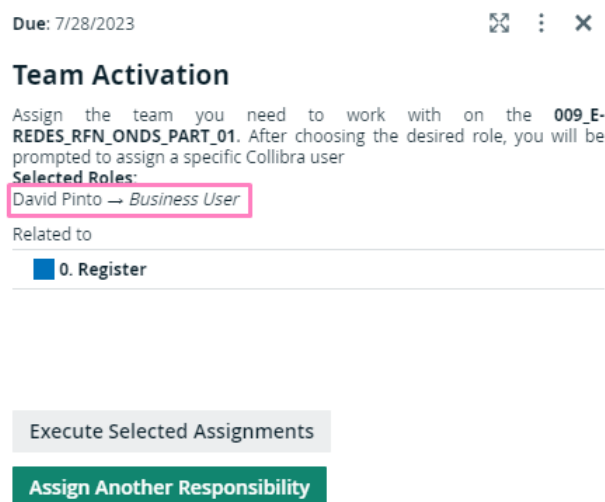


Figura 6.31: Ativação da Equipa - Finalizar ou Continuar

6.1.6.4 Abertura e Fecho de Secções

Como referido na subsecção 5.2.6.4, o objetivo deste requisito é abrir e fechar secções da DGT. A ideia em se abrir secções tem dois propósitos, o primeiro é diminuir a possibilidade de erro, uma vez que, caso todas as secções se encontrassem abertas, os utilizadores que a iam catalogar podiam enganar-se na secção e catalogar a errada. O segundo é que, como secções dentro de fases diferentes podem representar os mesmos ativos, é crucial que a secção anterior se encontre fechada (finalizada) para que a secção seguinte possa ser catalogada tendo em consideração a informação da secção anterior.

Para se inicializar este workflow, é necessário que o utilizador pertença ao grupo *Data Office* e que o mesmo encontre na vista da DGT com a mesma selecionada, conforme observável na figura 6.32, aqui, o mesmo encontrará o botão denominado de *Matrix*. A escolha do nome do botão foi feita pelo cliente.

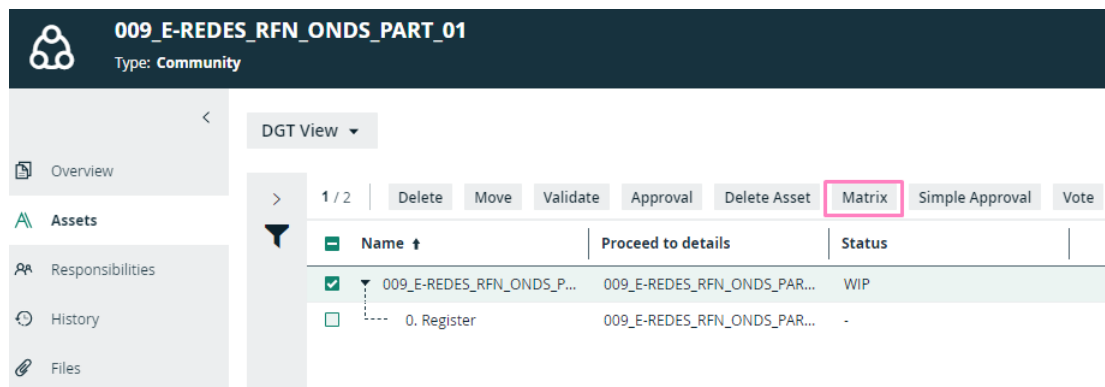


Figura 6.32: Abertura e Fecho de Secções - Botão

A primeira task despoletada para o cliente possui uma lista de todas as secções existentes na DGT, com o valor padrão a "falso", querendo informar de que todas as secções se encontram fechadas. Para abrir uma ou mais secções, o utilizador apenas necessita de alterar o valor para "verdadeiro" e carregar no botão "submit" para finalizar o processo, conforme observável na figura .

Due: 7/31/2023

Matrix

Related to

009_E-REDES_RFN_ONDS_PART_01

Data Management - Business Understanding - Functional (Data Model)
Functional specification (aggregated by possible table)

False

Data Management - Business Understanding - Functional (Data Model Detail)
Functional specification (detailed for all the fields that are needed)

False

Data Management - Business Understanding - Functional (Queries)
If the Business User already has queries, they can share them

False

Data Management - Business Understanding - Data Access
Sheet to indicate the accesses for each dataset (in pre-production)

False

Data Management - Data Understanding - Sources (Data Model)
Technical Specification (Data Model) of the source system.

False

Data Management - Data Understanding - Sources (Data Model Design)
Technical Specification (Visual Data Model) of the source system

False

Data Management - Data Understanding - Sources (Data Model Detail)
Technical Specification (detailed for all fields) of the source system.

False

Data Management - Data Understanding - Sources (Samples)
Technical Specification (sample data for each source table)

False

Data Management - Data Understanding - Sources (Queries)
Technical Specification (extraction query for each source table)

False

Data Management - Data Understanding - Sources (Queries)
Technical Specification (extraction query for each source table)

False

Data Management - Data Understanding - Source Auxiliary Tables
If we need to have subqueries

False

Data Management - Data Modelling - Destination (Data Model)
Technical Specification (Datasets) of the final model.

False

Data Management - Data Modelling - Destination (Data Model Detail)
Technical Specification (detailed for all fields) of the final model.

False

Data Management - Data Modelling - Destination (Data Model Design)
Technical Specification (Visual Data Model) of the final model

False

Data Quality - Data Validation - Quality Guidelines
Standard guidelines to assure before the datasets are send to Data Office

False

Data Quality - Data Validation - Quality Controls (Catalogue)
Definition of quality controls

False

Data Quality - Data Validation - Quality Controls (Outputs)
Results after the execution of the quality controls

False

Data Quality - Data Validation - Acceptance
Acceptance status for each datasets

False

Data Quality - Data Validation - Tamr
If the initiative has Tamr

False

Submit

(a) Início

(b) Fim

Figura 6.33: Abertura e Fecho de Secções - Selecionar Secções

Também já referido na subsecção 5.2.6.4, após a conclusão do processo, uma notificação, igual à observável na figura 6.34, é enviada a todos os elementos responsáveis por a catalogar. Uma vez que no exemplo apresentado foi aberta a secção *Functional (Data Model)*, uma task vou enviada para o *Business User* da DGT em questão.

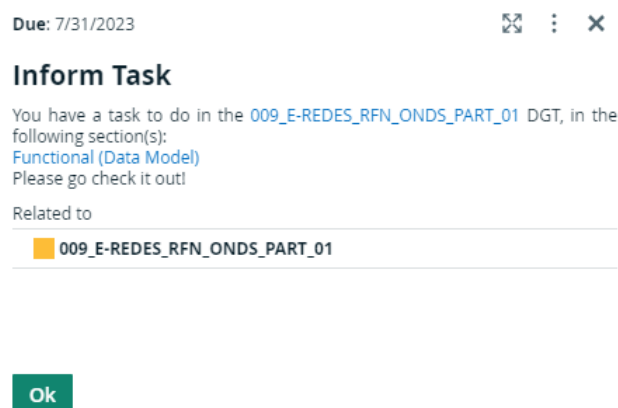


Figura 6.34: Abertura e Fecho de Secções - Notificação

Como é possível observar-se, em contraste com a figura 6.32, a figura 6.35, apresenta uma nova secção aberta.

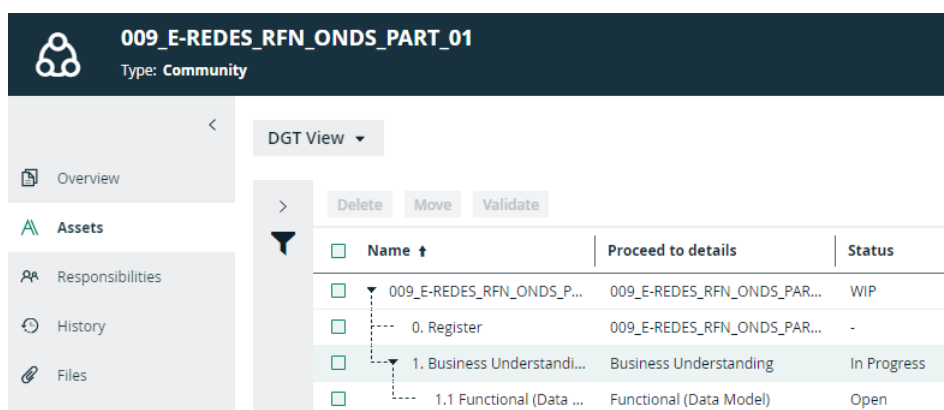


Figura 6.35: Abertura e Fecho de Secções - Secção Aberta

6.1.6.5 Criação de Tabelas

Ao contrário dos workflows anteriores, diversas pessoas conseguem executar este workflows, nomeadamente *Business Users*, *Data Stewards*, *Data Custodians* e *Data Officers*, sendo que o propósito do mesmo seja criar o ativo que será catalogado numa dado secção.

O Processo que será apresentado nesta subsecção é referente à criação de tabelas para a secção *Functional (Data Model)*, no entanto o mesmo processo é aplicado às secções *Functional (Queries)*, *Sources (Samples)*, *Sources Auxiliary Tables*, *Quality Controls (Catalogue)*, *Acceptance* e *Tamr*.

Para seleccionar o botão de execução, o utilizador necessita de se encontrar na vista da DGT e seleccionar a secção *Functional (Data Model)* (para esta situação) ou qualquer uma das outras secções descritas anteriormente para as outras situações. O botão é observável na figura 6.36.

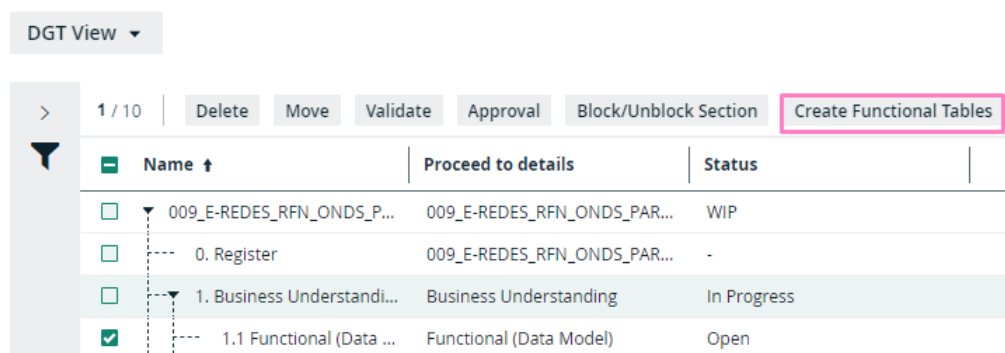


Figura 6.36: Criação de Tabelas - Botão

Após iniciar o workflow, o utilizador necessita de preencher a task apresentada na figura 6.37 com o(s) nome(s) das tabelas que pretende criar, tendo em consideração de que os nomes das mesmas devem estar separados por ";".

O cliente reforçou o facto de que muitas vezes eles possuem os nomes das tabelas num Excel, mas em linhas distintas, e gostaria de ter uma opção eficiente de transformar os nomes no Excel, no formato necessário para ser criado no Collibra.

Com isto em mente, foi criado o ficheiro Excel presente na mesma imagem, onde os utilizadores podem fazer download do mesmo, copiar e colar os nomes das tabelas e automaticamente, o Excel transforma no formato necessário, sendo, novamente, apenas necessário copiar e colar do Excel para a task em Collibra.

Quando tiver escrito o nome de todas as tabelas que pretende criar, o utilizador necessita apenas de carregar no botão "submit" e a(s) tabela(s) são criadas, como é observável na figura 6.38, onde, no exemplo efetuado, foi apenas criada uma tabela denominada "nome-Tabela".

Due: 7/31/2023

Get Tables Name

Please use this [Excel document](#) to facilitate the creation of Tables/Columns.

Related to

☒ 1.1 Functional (Data Model)

Please Insert the names of the Tables separated by semicolons

Submit

Figura 6.37: Criação de Tabelas - Escrita do Nome



Figura 6.38: Criação de Tabelas - Tabela Criada

6.1.6.6 Criação de Colunas

Tal como existe a necessidade de se criar tabelas na secção de *Functional (Data Model)*, na secção de *Functional (Data Model Detail)* existe a necessidade de se criar colunas, que por sua vez, se encontram relacionadas com as tabelas da secção anterior. Como tal, para se criar colunas, primeiro é necessário que existam tabelas, caso contrário, é apresentado ao utilizador uma notificação a informar de que não existem tabelas e que primeiro necessita de as criar.

Semelhante ao workflow anterior, este também é executado por múltiplas responsabilidades, que necessitam de se encontrar na vista da DGT, no entanto para se criar colunas necessitam de seleccionar a secção *Functional (Data Model Detail)*, conforme observável na figura 6.39.

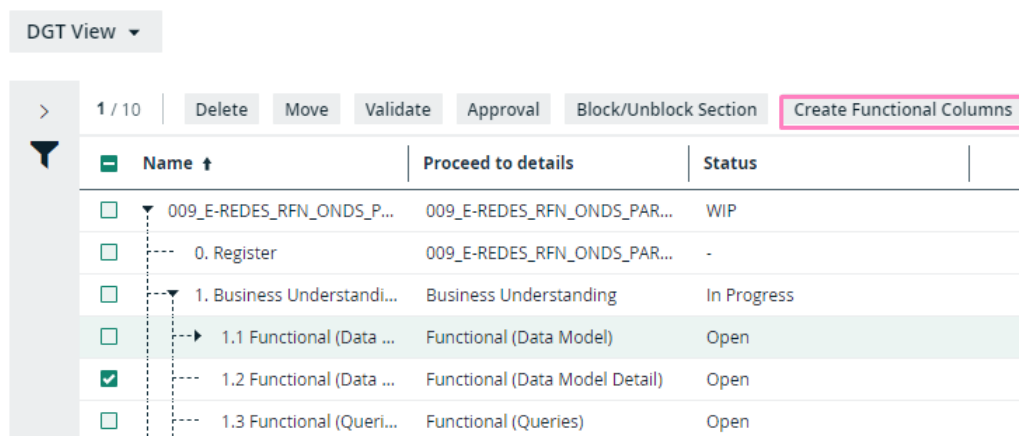


Figura 6.39: Criação de Colunas - Botão

Contrariamente ao processo de criação de tabelas, neste é necessário primeiro seleccionar-se a que tabela se quer adicionar a(s) coluna(s), conforme observável na figura 6.40. Na lista *dropdown* presente na imagem são fornecidas todas as tabelas que tenham sido criadas na secção anterior.

Due: 7/31/2023

Choose Table

Related to

1.2 Functional (Data Model Detail)

Please choose a table to add columns to.

Select

nomeTabela

Submit

Figura 6.40: Criação de Colunas - Seleção da tabela

Após o utilizador seleccionar a tabela a que quer adicionar colunas, o mesmo é apresentado uma task, observável na figura 6.41, que é igual à de criação de tabelas, onde é necessário escrever o(s) nome(s) das colunas separado por ",".

Due: 7/31/2023

Get Column Names

Related to

1.2 Functional (Data Model Detail)

Please Insert the names of the Columns for the Table nomeTabela
separated by comma

Submit

Figura 6.41: Criação de Colunas - Escrita do Nome

Após escrever os nomes das colunas, o utilizador é apresentado com a opção de finalizar o processo ou seleccionar outra tabela para adicionar mais colunas, conforme a figura 6.42.

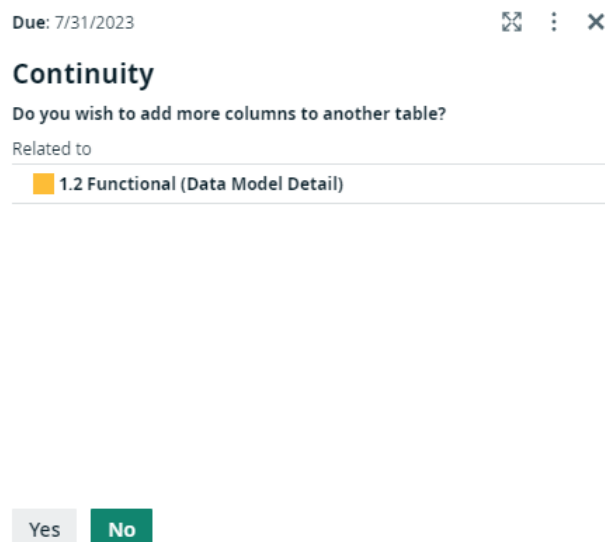


Figura 6.42: Criação de Colunas - Continuidade

No final do processo, as colunas são criadas e observáveis na vista da secção, conforme apresenta a figura 6.43.

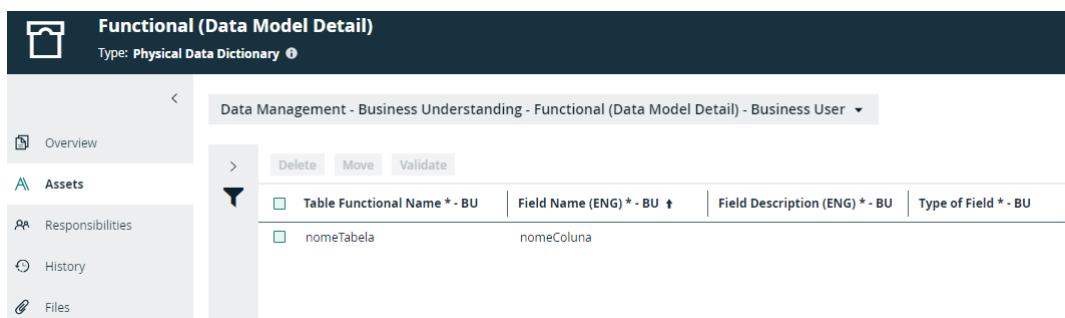


Figura 6.43: Criação de Colunas - Coluna Criada

6.1.6.7 Adição de Utilizadores a Data Access

Este processo é bastante semelhante ao de criação de tabelas, no entanto, em vez de se criar tabelas, cada "linha" apresentada na secção *Data Access* corresponde a um utilizador, conforme observável na figura 6.46.

Os utilizadores que conseguem executar este workflow são os mesmos que nos últimos dois processos, e para o executarem, necessitam de se encontrar na vista da DGT com a secção seleccionada, conforme é visível na figura 6.44.

DGT View ▾

> 1 / 10 | Delete | Move | Validate | **Add Data Access Elements** | Approval

☐	Name ↑	Proceed to details	Status
☐	009_E-REDES_RFN_ONDS_P...	009_E-REDES_RFN_ONDS_PAR...	WIP
☐	0. Register	009_E-REDES_RFN_ONDS_PAR...	-
☐	1. Business Understandi...	Business Understanding	In Progress
☐	1.1 Functional (Data ...	Functional (Data Model)	Open
☐	1.2 Functional (Data ...	Functional (Data Model Detail)	Open
☐	1.3 Functional (Queri...	Functional (Queries)	Open
☒	1.4 Data Access	Data Access	Open

Figura 6.44: Adição de Utilizadores a Data Access - Botão

Na única task despoletada para o utilizador, o mesmo simplesmente necessita de seleccionar um utilizador, de todos os existentes na plataforma Collibra, conforme a figura 6.45. Após o envio da seleção, o utilizador é criado, já com alguns campos preenchidos a partir das informações disponíveis sobre a pessoa seleccionada, tal como é o caso do id, observável na figura 6.46.

Due: 7/31/2023 🔍 ⋮ ✕

Please Insert Users to Add

Related to

1.4 Data Access

Please select the users to add Data Access

David Pinto ⋮

Submit

Figura 6.45: Adição de Utilizadores a Data Access - Escrita do Nome

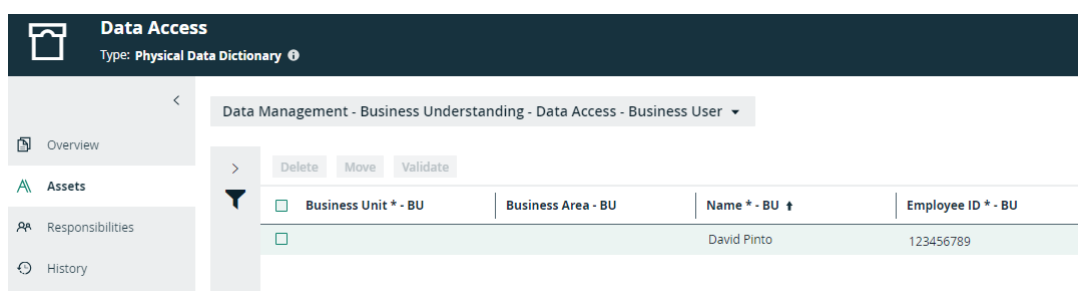


Figura 6.46: Adição de Utilizadores a Data Access - Utilizador Criado

6.1.6.8 Importação de Business Understanding e Data Understanding

Tal como discutido anteriormente, existem secções dentro de fases diferentes que são referentes aos mesmo ativos, mas catalogadas por elementos diferentes, como tal faz sentido importar-se todos os ativos catalogados na secção anterior de forma a haver uma coesão e coerência entre secções, também para o elemento que for catalogar a seguir, ter uma ideia do que está a catalogar, precisando então da catalogação efetuada pelo elemento anterior.

Com isto em mente, é necessário um workflow que efetue essa importação corretamente, e havendo duas fases distintas onde seja necessário efetuar-se essa importação na fase de *Data Understanding* importar os ativos da fase *Business Understanding* e na fase *Data Modeling* importar os ativos da fase *Data Understanding*.

Para se executar, o utilizador basta encontrar-se na vista da DGT, conforme observável na figura 3.2 da subsecção 5.2.6 e selecionar a secção para onde quer importar e pressionar o botão de importação, conforme a figura 6.47.

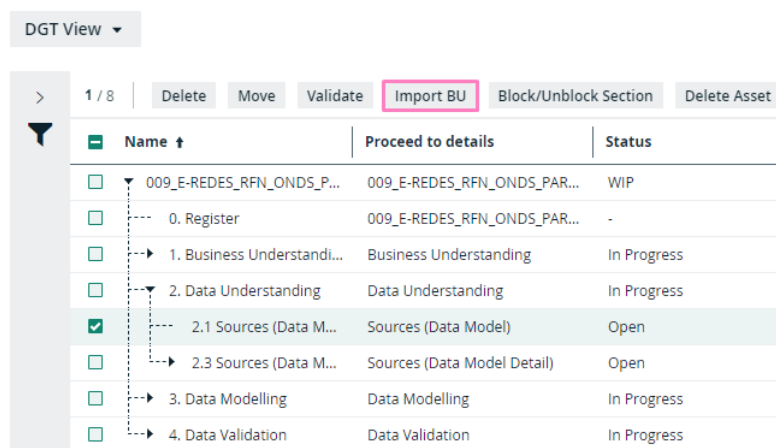


Figura 6.47: DGT - Botão de Importação de Business Understanding e Data understanding

Nas figuras 6.48 e 6.49 pode-se observar os ativos presentes na secção *Functional (Data Model)* e que após a importação, os ativos na secção *Sources (Data Model)* são os mesmos que os da anterior.

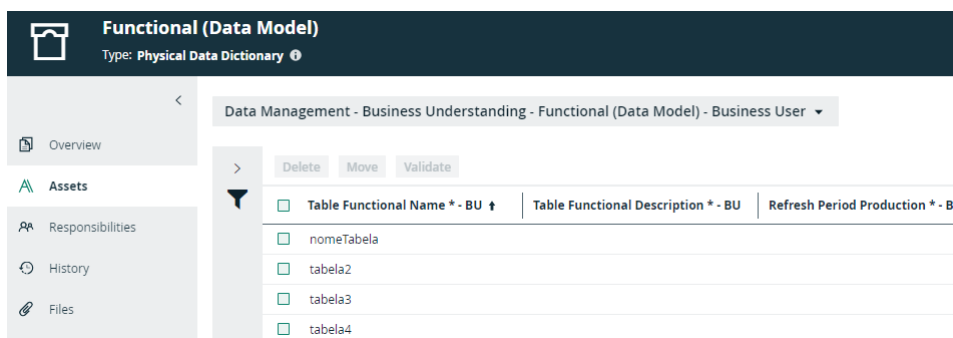


Figura 6.48: DGT - Importação de *Business Understanding* e *Data Understanding* - *Business Understanding*

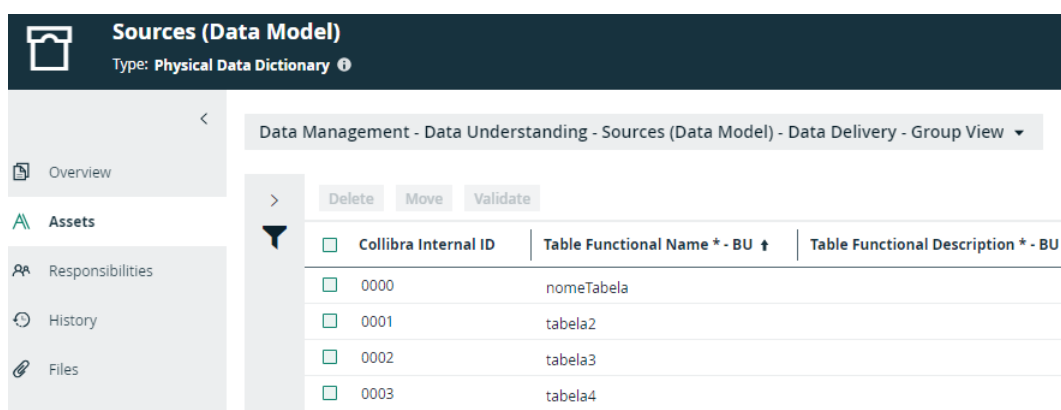


Figura 6.49: DGT - Importação de *Business Understanding* e *Data Understanding* - *Data Understanding*

6.1.6.9 Importação dos ativos de Controlo de Qualidade

Esta secção, *Quality Controls (Outputs)*, é a subsequente à *Quality Controls (Catalogue)* e os ativos desta são os mesmos da secção anterior, deste modo, é necessário que seja possível importar-se os ativos da secção anterior para esta, de modo a informação se encontrar coerente e coesa, facilitando assim, a catalogação do utilizador.

Semelhante ao processo anterior, é necessário que o utilizador se encontre na vista da DGT, conforme observável na figura 3.2 da subsecção 5.2.6 e seleccionar a secção para onde quer importar, que nesta situação é a secção *Quality Controls (Outputs)* e pressionar o botão de importação, conforme a figura 6.50.

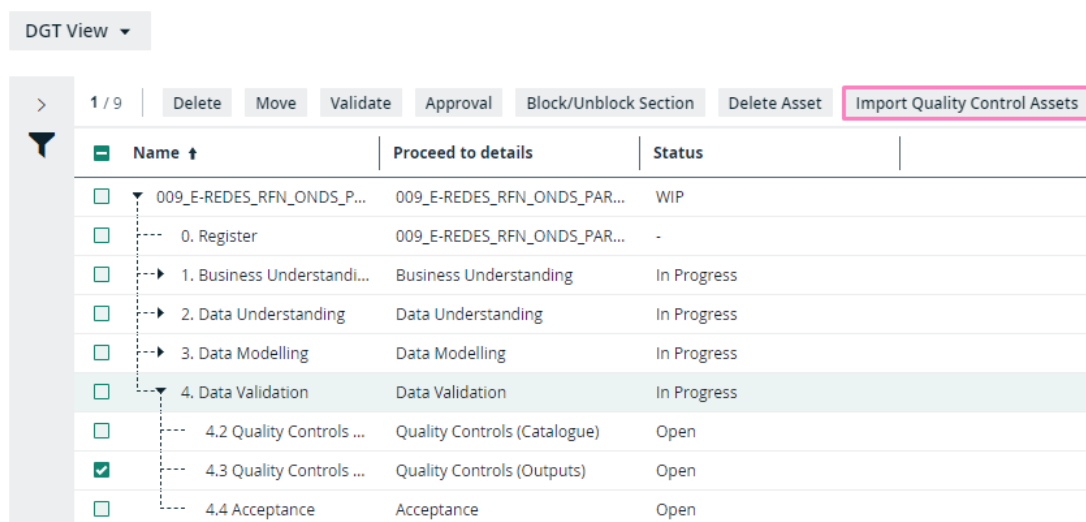


Figura 6.50: DGT - Botão de Importação dos Ativos de Controlo de Qualidade

Nas figuras 6.51 e 6.52 pode-se observar os ativos presentes na secção *Quality Controls (Catalogue)* e que após a importação, os ativos na secção *Quality Controls (Outputs)* são os mesmos que os da anterior, inclusive os atributos referentes ao nome do data set e ao valor do atributo "Control Status" onde o mesmo é catalogado com o valor padrão de "fail", devido ao facto de, se retratarem de problemas e após o mesmo ser resolvido, o valor é passado a "pass".

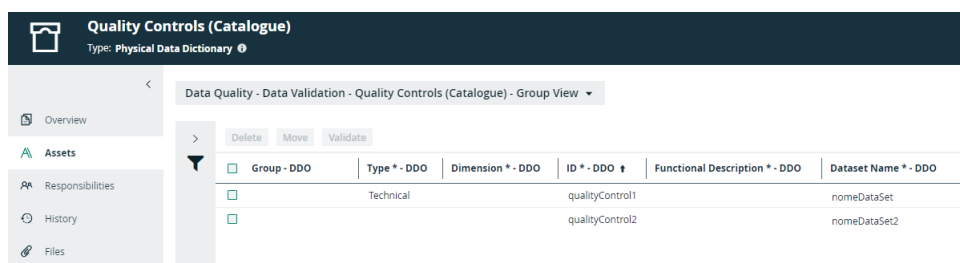


Figura 6.51: DGT - Importação dos Ativos de Controlo de Qualidade (Secção 1)



Figura 6.52: DGT - Importação dos Ativos de Controlo de Qualidade (Secção 2 Após Importação)

6.1.6.10 Importação de Secções de outras DGTs

Muitas vezes, existem secções bastante semelhantes, se não iguais, entre DGTs. Para tal, é crucial que exista a possibilidade de se importar todos os ativos dessa secção para a que se encontra em desenvolvimento.

Com esta situação em mente, foi desenvolvido um workflow que se encontra disponível apenas para *Data Officers* e apenas podem ser importadas secções de DGTs já finalizadas, uma vez que não faria sentido importar-se algo que poderia ainda sofrer efeito ou possuir defeitos.

Da mesma maneira que os workflows anteriores, o *Data Officer* necessidade de se encontrar na vista da DGT, selecionar a secção para onde quer importar dados e selecionar o botão presente na figura 6.53.

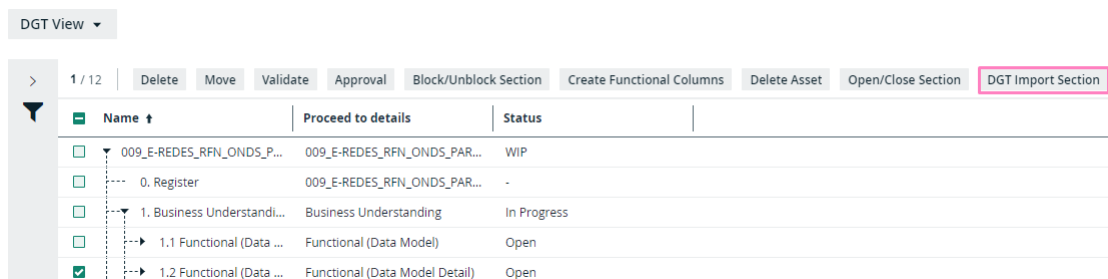


Figura 6.53: DGT - Botão de Importação de Secções

Na figura 6.54 é possível observar-se a secção *Functional (Data Model)* de uma DGT e na figura 6.55 a mesma secção, mas noutra DGT com os mesmo ativos da primeira mais os que ela já possuía (que neste exemplo foram as tabelas criadas para o requisito de criação de tabelas).

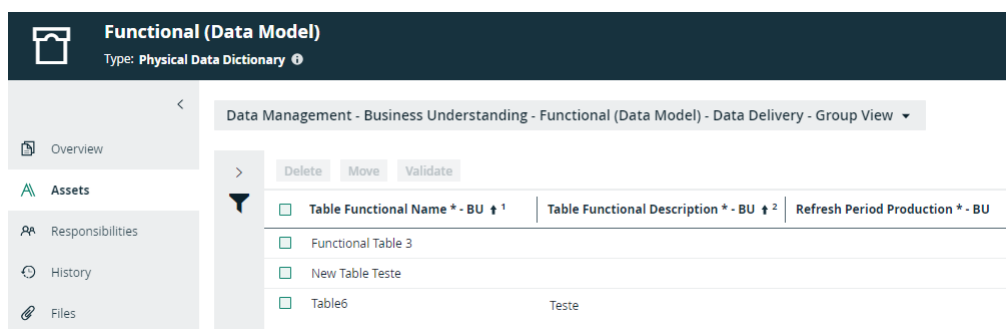


Figura 6.54: DGT - Importação de Secções (Secção 1)

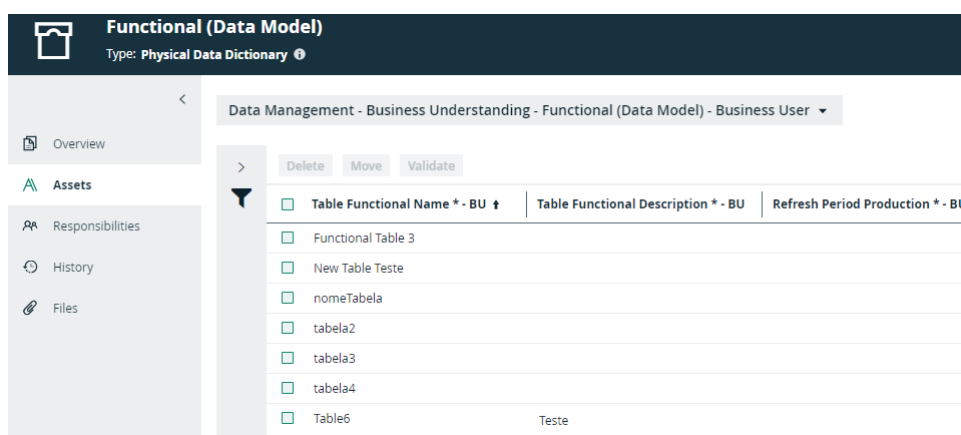


Figura 6.55: DGT - Importação de Secções (Secção 2 Após Importação)

6.1.6.11 End DGT

Uma vez que o objetivo de uma DGT é criar novos ativos, que por sua vez representam novos dados, é necessário efetuar-se essa criação após a catalogação do DGT se encontrar finalizada.

Este workflow não possui nenhum processo visual para o utilizador, o botão é pressionado e todos os ativos, relações a atributos são criados.

Este processo apenas pode ser executado por *Data Officers*, em DGTs concluídas, por outras palavras, que tenha o status a "CV", e para se ter uma DGT finalizada, todas as secções necessitam de se encontrar fechadas, conforme observável na figura 6.56.

Deste workflow são então criadas, tal como descrito na subsecção 5.2.6, tabelas, colunas e data concepts.

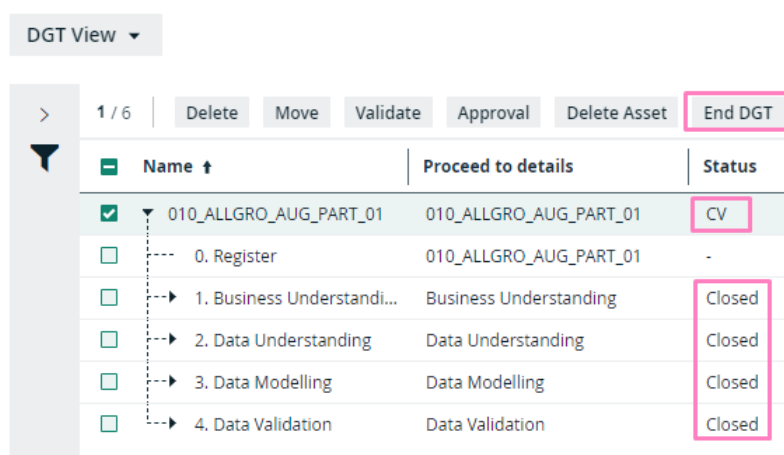


Figura 6.56: DGT - Botão End DGT

6.2 Customização da Landing Page

A *Landing Page* tem como objetivo guiar os utilizadores dentro do Collibra, possuindo links para as comunidades, domínios ou vistas mais utilizadas pelos mesmos, tal como auxiliar os

novos utilizadores em conhecer e entender melhor a nova plataforma que é o Collibra.

Deste modo a *Landing Page* foi implementada utilizando HTML, CSS e JavaScript, sendo que para se apresentar a mesma dentro da plataforma Collibra, é necessário a utilização de widgets, que a mesma dispõe para casos como este. Para tal foi necessário importar os ficheiros do código da *Landing Page* para o Collibra e adicionar-se ao *widget* o url local dos ficheiros, como é observável na figura 6.57.

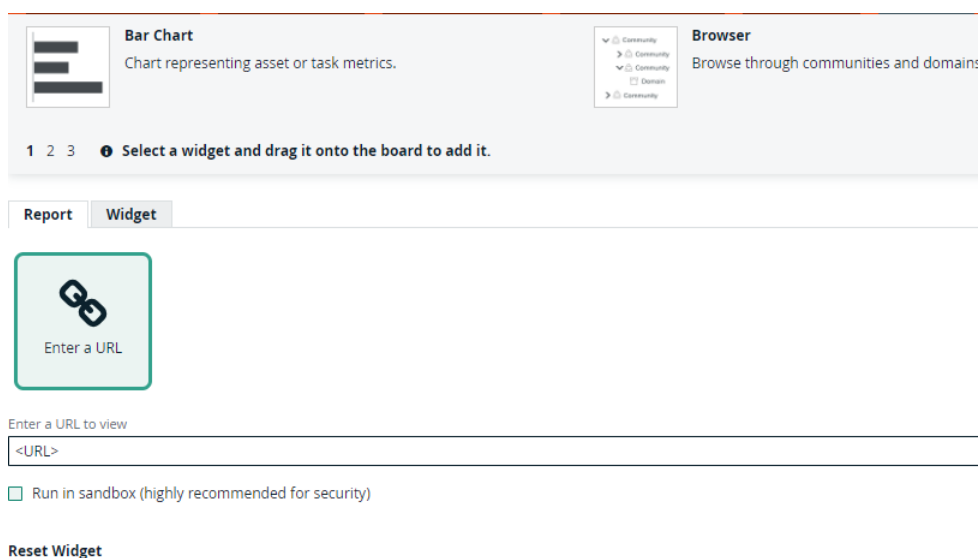


Figura 6.57: Landing Page Widget

Após a implementação, a landing page implementada é observável conforme as figuras 6.58 e 6.59. Sendo que todos os requisitos foram implementados com a exceção da parte onde eram mostradas as tasks do utilizador. Esta não implementação deveu-se ao facto de que informação sobre as tasks do utilizador já é apresentada na barra de navegação da própria plataforma, como se pode verificar na figura 6.60, logo expor esta informação duplicada seria redundante, pelo que se optou por não a implementar.

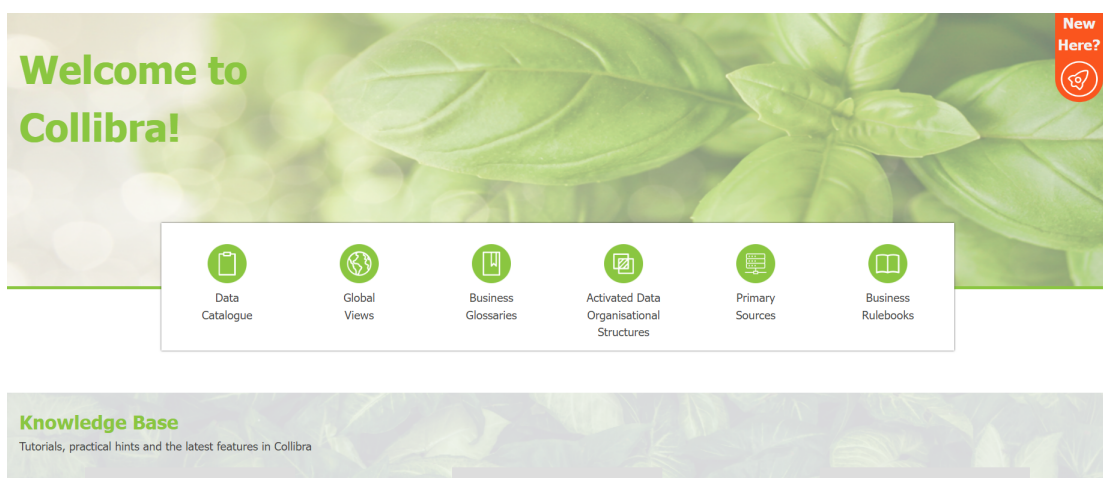


Figura 6.58: Landing Page Image - Parte Cima

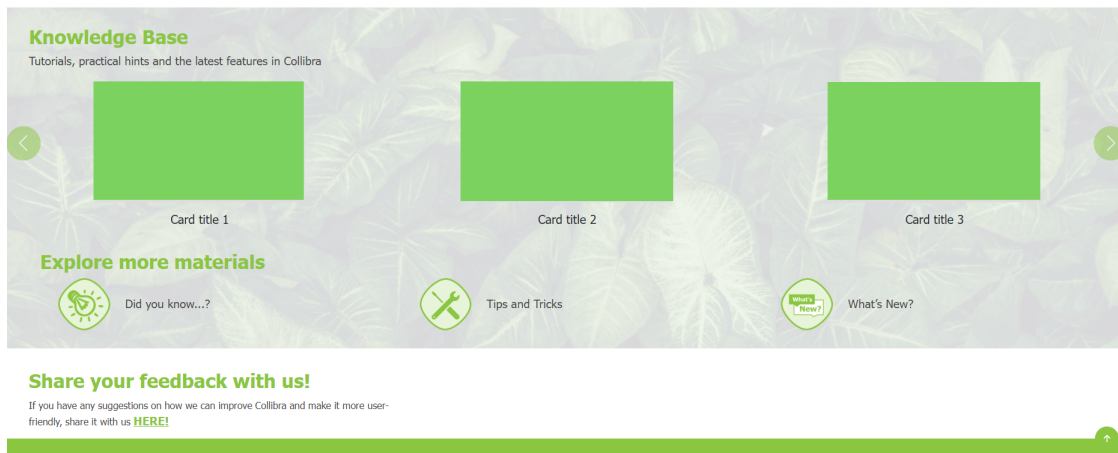


Figura 6.59: Landing Page Image - Parte Baixo

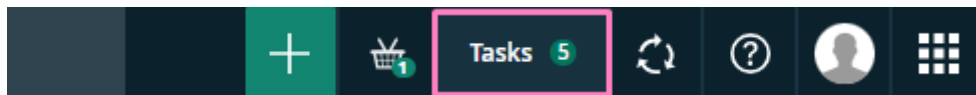


Figura 6.60: Landing Page Barra Navegação

Capítulo 7

Avaliação da Solução

Em engenharia de software a qualidade é um aspeto vital que deve ser considerado ao desenvolver qualquer tipo de trabalho. Um bom sistema é considerado aquele que para além de atender às necessidades do cliente, também o faz de forma eficaz.

Neste capítulo são descritas as hipóteses de investigação ligadas ao problema em causa, referidos os indicadores e fontes de informação para avaliar a qualidade da solução proposta com este trabalho em relação aos objetivos e requisitos estabelecidos. Por fim, especificada a metodologia de avaliação empregue.

7.1 Hipóteses de Investigação

Esta secção apresenta as hipóteses que guiam a avaliação do sistema deste projeto e permitem dizer se tem ou não qualidade.

Uma hipótese é uma possível resposta a uma pergunta de pesquisa. Trata-se de um palpite ou presunção com base no qual um estudo deve ser conduzido. Esta hipótese é testada para uma potencial aprovação ou rejeição. No caso de esta ser aceite, demonstra-se que a presunção estava correta (ReadingCraze 2015).

De acordo com os objetivos e requisitos funcionais e não funcionais descritos nos capítulos 1 e 3, este projeto é considerado viável se cumprir algumas dimensões de qualidade que precisam de ser medidas.

As dimensões que serão utilizadas para avaliar a qualidade do sistema são:

- A solução proposta é capaz de suportar todos os processos de Data Governance (DG);
- Os requisitos funcionais e não funcionais referentes à solução proposta são cumpridos na íntegra e apresentam um alto grau de usabilidade;

Considerando uma percentagem predefinida de 70% para o resultado final da avaliação da qualidade do sistema, pode-se definir uma hipótese nula (H0) e uma alternativa (H1):

- H0 - O sistema não é viável pois está avaliado abaixo da percentagem predefinida;
- H1 - O sistema é viável desde que seja avaliado com uma percentagem de valor igual ou superior ao do predefinido.

7.2 Indicadores e Fontes de Informação

Esta secção apresenta os indicadores e fontes de informação que precisam de ser recolhidas para avaliação posterior da qualidade e validade da solução.

Tendo em consideração as dimensões da secção anterior, podem-se apurar duas características passivas de serem avaliadas. Uma delas é a usabilidade do produto final e o seu processo de avaliação consiste em efetuar testes unitários e UAT's (Teste de Aceitação do Utilizador) que são testes executados pelo utilizador final ou o cliente para verificar/aceitar a solução.

A outra característica para avaliar corresponde ao grau de qualidade da solução final, que será avaliada com recurso ao modelo de avaliação Quantative Evalution Framework (QEF).

7.3 Metodologia de Avaliação

Esta secção visa obter a percentagem da qualidade final deste projeto usando a ferramenta QEF, que é descrita na primeira subsecção. No final, a aplicação do QEF a este projeto é feita e a medida de qualidade é obtida.

7.3.1 Quantitative Evaluation Framework

QEF é um modelo desenvolvido por Escudeiro e Bidarra (2008) que avalia uma solução real (sistema implementado) comparada com uma solução ideal (sistema que atinge todos os parâmetros de avaliações considerados).

Esta comparação consiste em colocar ambas as soluções num espaço composto por três dimensões, como é possível observar-se na figura 7.1 a fim de se calcular a distância euclidiana entre elas.

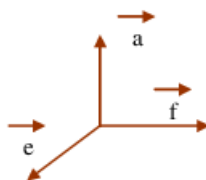


Figura 7.1: Espaço Tridimensional

O modelo em causa assinala que a qualidade do sistema é inversamente proporcional à distância entre o sistema ideal e o sistema real (Escudeiro e Bidarra 2008), isto significa que quanto maior for a distância calculada mais longe a solução real se situa da solução ideal.

QEF envolve três conceitos (Escudeiro e Bidarra 2008):

- **Dimensão** - Calculado a partir dos valores de importância e conformidade dos fatores que o compõem;
- **Fator** - Um fator é composto por dois elementos:

- *Importância* - Diz respeito à divisão da soma dos graus de relevância dos requisitos a si associados pela soma de graus de relevância de todos os requisitos da dimensão onde o fator se encontra inserido;
- *Cumprimento* - Diz respeito à divisão da soma da avaliação de cumprimento dos requisitos pelo valor total de cumprimento de requisitos que fazem parte do fator.
- **Requisito** - Um requisito é composto por dois elementos:
 - *Avaliação* - Grau de relevância do requisito e pode ser classificado numa escala de 0, 2, 4, 6, 8 e 10, sendo 10 a classificação que corresponde ao maior nível de relevância;
 - *Relevância* - Grau de avaliação que corresponde à taxa de cumprimento de um requisito e pode ser classificada numa escala percentual de 0, 25, 50, 75 e 100, sendo 100 a taxa que corresponde ao maior nível de cumprimento.

A qualidade do sistema é então medida, em percentagem, quanto ao cumprimento dos critérios definidos para a solução ideal. Para isso, calcula-se a distância entre o real e a solução ideal. Quanto maior essa distância, mais longe a solução real está de ser comparável à solução ideal, o que significa que a qualidade é baixa.

7.3.2 Testes de Aceitação do Utilizador (UATs)

O teste de aceitação do utilizador (UAT), também chamado de teste de aplicação ou teste de utilizador final, é uma fase de desenvolvimento de software na qual o software é testado no mundo real pelo seu público-alvo. O UAT geralmente é a última fase do processo de teste de software e é realizado antes do software testado ser lançado para produção. O objetivo do UAT é garantir que o software possa lidar com tarefas do mundo real e executar de acordo com as especificações de desenvolvimento (Software-Testing-Help 2023).

Nos UATs, os utilizadores têm a oportunidade de interagir com o software antes do seu lançamento oficial para ver se algum requisito foi esquecido ou se contém algum bug. Os resultados dos primeiros testes são encaminhados aos programadores, que fazem as alterações finais antes de lançar o produto para produção (Software-Testing-Help 2023).

7.3.2.1 Testes de Integração

O significado de teste de integração refere-se ao processo de teste das interfaces entre dois componentes ou módulos de software para avaliar a forma como os dados são transferidos entre eles. As estratégias de teste de integração permitem às equipas de desenvolvimento detetar defeitos que podem ser introduzidos quando dois ou mais módulos de software são integrados, bem como avaliar o ajuste global e a função dos elementos de software combinados (Zaptest 2023).

7.3.2.2 Testes funcionais

Os testes funcionais permitem que os testes ocorram de uma forma mais eficiente e rápida, possibilitando encontrar as não conformidades do software em relação aos requisitos do sistema. O principal objetivo dos testes de software é a localização de erros, falhas, defeitos e a verificação das funcionalidades do software em desenvolvimento ou finalizado (DevMedia 2023).

7.3.2.3 Testes de ponta-a-ponta

Teste de ponta a ponta replica o comportamento de um utilizador com o software num ambiente de utilização completo. O mesmo verifica se vários fluxos de utilizador funcionam como esperado e podem ser tão simples quanto carregar uma página da web ou efetuar login ou cenários muito mais complexos verificando notificações por e-mail, pagamentos on-line, entre outros (Pittet 2023).

7.3.2.4 Testes de aceitação

Teste de aceitação é o processo de comparação do programa com os seus requisitos iniciais e as necessidades atuais dos seus utilizadores finais. Semelhante a outros tipos de teste, a melhor maneira de efetuar este é criar casos de teste que tentam mostrar que o programa não cumpre com os requisitos; se estes casos de teste são mal sucedidos, o programa é aceite (Myers 2004).

7.3.2.5 UAT efetuado

Os UATs à solução implementada foram efetuadas pelas partes interessadas do cliente. Ao longo do processo de testes, foram efetuados testes de integração, testes funcionais, testes de ponta-a-ponta e testes de aceitação. Deste processo, originou o dashboard presente na figura 7.2.

Este dashboard encontra-se dividido em 5 grupos. O primeiro a amarelo informa que todos os requisitos principais encontram-se testados a 100% e quantos requisitos, dentro de cada grupo, foram testados, sendo possível observar-se que dentro do grupo de *DOS Activation*, foram testados 4 requisitos, nomeadamente ativação de *Data Owner*, *Data Steward*, *Data Custodian* e em massa.

No grupo seguinte, a azul, retrata informação semelhante ao grupo anterior, mas com o tópico das DGTs mais detalhado, sendo que o mesmo se encontra dividido em 7 subgrupos.

A vermelho no centro, temos informações sobre os defeitos e melhorias encontradas, neste caso, foram encontrados seis defeitos, no entanto, todos foram corrigidos devidamente e já não existem defeitos. Foram também encontradas nove melhoras que foram adicionadas ao desenvolvimento futuro.

No canto superior direito, a verde, encontra-se, em forma de tópicos, um pequeno resumo do dashboard, nomeadamente que todos os requisitos dos três grupos principais foram totalmente testados, que foram encontrados seis defeitos, mas que já foram todos resolvidos e que os defeitos que se encontravam em observação do status anterior foram descartados a adicionado para melhorias futuras, uma vez que não faziam parte do âmbito do projeto.

Por último a roxo, temos uma representação do número de testes efetuados (102) e se os mesmos seguiram o script criado para os testes (Test-plan) ou se foram Ad-hoc, por outras palavras, se a pessoa que testou, o fez seguindo os passos que assim entendeu como necessários para efetuar o teste da melhor maneira.

UAT Overview

Last update: 23th June

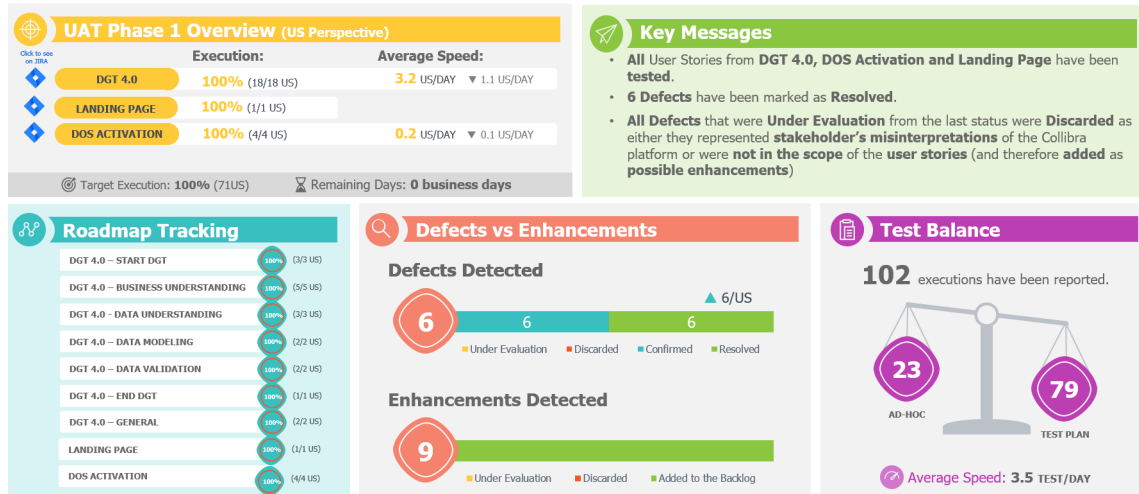


Figura 7.2: Resumo das UATs

Capítulo 8

Conclusão

O projeto incidiu na passagem do processo de Data Governance (DG) do cliente para o ambiente Collibra, existindo um elevado ênfase no processo Data Gathering Tool (DGT) e nas várias possibilidades de ativação de responsabilidades.

Considerando os objetivos para este projeto mencionados na secção 1.3, é possível dizer que estes foram alcançados. E tendo em consideração o capítulo 7, onde são apresentados os testes efetuados e os seus resultados, todos os requisitos que se mantiveram em âmbito para o trabalho, passaram nos múltiplos testes efetuados, tendo sido descobertos seis defeitos, mas rapidamente corrigidos.

No início do projeto, os requisitos referentes à DGT eram esperados serem algo simples e rápidos de se implementar, no entanto rapidamente se tornou em algo complexo e demorado. Deste modo, no âmbito do projeto, houveram requisitos que foram removidos, por parte do cliente, pois foi atribuída uma prioridade mais elevada aos da DGT.

Embora não muito relevante, houve uma falta de especificação referente aos requisitos do projeto, uma vez que, foram necessárias múltiplas reuniões com o cliente para se conseguir obter uma ideia correta e final do que seria esperado, o que originou acréscimo de tempo dado aos objetivos da DGT e a remoção de requisitos de âmbito.

A passagem do processo de *Data Governance* para a plataforma Collibra, trouxe múltiplos pontos positivos, mais pormenorizadamente, as DGTs vão passar a estar no mesmo ambiente centralizado onde vão coexistir com os metadados, sistemas e data roles críticos para o governo do ecossistema de dados do cliente, podendo beneficiar diretamente da informação existente no catálogo de dados corporativos; A comunicação e *stewardship* aumenta a sua eficiência, na medida em que é possível a atribuição de roles, responsabilidades, menções, atribuição de tasks, notificações e recolha de informação de utilizadores; Monitorização mais próxima do histórico de alterações de catalogação através do histórico nativo; Enriquecimento instantâneo do catálogo de dados corporativos após a conclusão de cada DGT.

Como trabalho futuro sugere-se a realização de inquéritos de qualidade às partes interessadas, como método de avaliação da solução, uma vez que foi um ponto menos tocado ao longo desta dissertação. Outro trabalho futuro a considerar, seria os requisitos que não foram cobertos no âmbito do trabalho, os mesmos não apresentam uma complexidade exorbitante em termos de implementação, sendo, portanto, uma oportunidade promissora a ser considerada no futuro, uma vez que são características que melhorariam a plataforma Collibra e facilitariam a interação do cliente com a mesma.

Por último, as melhorias, brevemente discutidas, no capítulo 7, provenientes das UATs, são, tal como as características que saíram de âmbito, aspetos importantes para a solução, e consequentemente, importantes para serem considerados no futuro.

Bibliografia

- (1997). Em: url: <https://www.bpmn.org/>.
- Abdel-Basset, Mohamed et al (2019). «A Hybrid Plithogenic Decision-Making Approach with Quality Function Deployment for Selecting Supply Chain Sustainability Metrics». Em: *Symmetry* 11.7. url: <https://www.mdpi.com/2073-8994/11/7/903>.
- Agbo, Innocent, Alexander Kupriyanov e Igor Rytsarev (2020). «Big Data Architecture: Designing the Right Solution for Social Network Analysis». Em: *8th International Symposium on Digital Forensics and Security (ISDFS)*, pp. 1–5.
- Batini, Carlo e Monica Scannapieco (2006). «Data Quality: Concepts, Methodologies and Techniques». Em: *Springer* 1, pp. 1–18.
- Billigence (2023). «Collibra Make Data Meaningful». Em: url: <https://billigence.com/collibra/>.
- Bulao, Jacquelyn (jul. de 2023). «How Much Data Is Created Every Day in 2023?» Em: *techjury*. url: <https://techjury.net/blog/how-much-data-is-created-every-day/>.
- Carlos Vazquez, Guilherme Simões (2016). «Engenharia de Requisitos: Software Orientado ao Negócio». Em: url: <https://www.fattocs.com/engenharia-de-requisitos/livro-de-engenharia-de-requisitos/>.
- Catro, Alfonso et al. (ago. de 2021). «An Ontological-Based Model to Data Governance for Big Data». Em: 9, pp. 7–10.
- Chan, Lai-Kow e Ming-Lu Wu (2002). «Quality function deployment: A literature review». Em: *European Journal of Operational Research* 143.3, pp. 463–497. url: <https://www.sciencedirect.com/science/article/abs/pii/S0377221702001789>.
- Chung, Lawrence (2012). «Non-functional requirements in software engineering». Em: *Springer*.
- Coepd (ago. de 2014). «What is FURPS+?» Em: url: <https://businessanalysttraining%20hyderabad.wordpress.com/2014/08/05/what-is-furps/>.
- DAMA (abr. de 2017). «DAMA-DMBOK - Data Management Body of Knowledge». Em: *Technics Publications* 2, pp. 18–19. url: <https://www.datameer.com/what-is-alation/>.
- Datameer (2023). «What is Alation?» Em: url: <https://www.datameer.com/what-is-alation/>.
- Dehghani, Zhamak (mar. de 2022). «Data Mesh: Delivery Data-Driven Value at Scale». Em: *O'Reilly Media, Inc* 1, p. 443. url: <https://www.oreilly.com/library/view/data-mesh/9781492092384/>.
- DevMedia (2023). «Testes funcionais de software». Em: url: <https://www.devmedia.com.br/testes-funcionais-de-software/23565>.
- Escudeiro, Paula e José Bidarra (2008). «Quantitative Evaluation Framework (QEF)». Em: *Conselho Editorial/Consejo Editorial* 16.
- Glavic, B. (2014). «Big data provenance: Challenges and implications for benchmarking». Em: *Revised Selected Papers of the First Workshop on Specifying Big Data Benchmarks* 8163, pp. 72–80.

- Governance, Ieee Big Data e Metadata Management Industry Connections Activity (dez. de 2020). «BIG DATA GOVERNANCE AND METADATA MANAGEMENT: STANDARDS ROADMAP TRADEMARKS AND DISCLAIMERS». Em: *Review of Scientific Instruments*, pp. 5–6. url: <http://standards.ieee.org/about/sasb/iccom/>.
- Graf, Albert e Peter Maas (2008). «Customer value from a customer perspective: a comprehensive review». Em: *Journal für Betriebswirtschaft*.
- Harwanto, Ikhsan M. e Achmad N. Hidayanto (dez. de 2022). «Data Governance Maturity Assessment: A Case Study Directorate General of Corrections». Em: *International Conference on ICT for Smart Society (ICSS)*, p. 2.
- Hyderabad, Business Analyst Training in (ago. de 2004). «What is FURPS+?» Em: url: <https://businessanalysttraininghyderabad.wordpress.com/2014/08/05/what-is-furps/>.
- IBM (2023). «What is data modeling?» Em: url: <https://www.ibm.com/topics/data-modeling>.
- Institute, Data Governance (jan. de 2020). «The DGI Data Governance DGI Data Governance Framework Framework». Em: *Review of Scientific Instruments*. url: https://datagovernance.com/wp-content/uploads/2020/07/dgi_data_governance_framework.pdf.
- Kaewkamol, Porntida (dez. de 2022). «Data Governance Framework as Initiative for Higher Educational Organisation». Em: *International Conference on Digital Arts, Media and Technology (DAMT)*, p. 1.
- Karlsson, Joachim (2004). «Managing software requirements using quality function deployment». Em: *Software Quality Journal* 6, pp. 311–326.
- Kermode, G. R., S. Sivaloganathan e T. M. M. Shahin (2000). «Value analysis—the technique: State of the art and future directions». Em: *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 214, pp. 301–312. url: <https://doi.org/10.1243/0954405001517694>.
- Lamsweerde, Axel van (2000). «Requirements Engineering in the Year 00: A Research Perspective». Em: *Proceedings of the 22nd International Conference on Software Engineering*, pp. 5–19. url: <https://doi.org/10.1145/337180.337184>.
- LexisNexis (2023). «Why data stewardship is a winning business investment». Em: url: <https://risk.lexisnexis.com/>.
- Liu, Zhidong (2020). «Practice research on University Data Center Construction from the Perspective of Data Governance». Em: *2nd International Conference on Information Technology and Computer Application (ITCA)*.
- Loshin, David (2001). «Enterprise Knowledge Management: The Data Quality Approach». Em: *Morgan Kaufmann*.
- Malan, Ruth e Dana Bredemeyer et al (2001). «Functional requirements and use cases». Em: *Bredemeyer Consulting*.
- Maulina, Joanita e Yova Ruldeviyani (ago. de 2019). «Data Governance and Data Architecture for the Ministry of Foreign Affairs of the Republic of Indonesia». Em: *2019 International Conference on Information Management and Technology (ICIMTech)*, pp. 409–414. url: <http://link.aip.org/link/?RSI/62/1/1>.
- Myers, Glenford J. (2004). «The Art of Software Testing». Em: *John Wiley & Sons, Inc* 2, pp. 143–145. url: <http://library.lol/main/24A555D1011FA1E62F0374963B072E04>.
- Neap, Halil Shevket e Tahir Celik (1999). «Value of a Product: A Definition». Em: *International Journal of Value-Based Management* 12, pp. 181–191.
- Newman, D. e D. Logan (mai. de 2006). «Governance Is an Essential Building Block for Enterprise Information System». Em: *Gartner Research*.

- O'Leary, D (2014). «Embedding AI and Crowdsourcing in the Big Data Lake». Em: *University of Southern California*.
- Oliveira, Paulo (out. de 2021). «Armazenamento e Processamento Analítico de Dados». Em: *Departamento de Engenharia Informática (DEI/ISEP)*. url: https://moodle.isep.ipp.pt/pluginfile.php/154583/mod_resource/content/3/DW_Introduction.pdf.
- Osterwalder, Alexander e Yves Pigneur (2003). «Modeling Value Propositions in E-Business». Em: *Proceedings of the 5th International Conference on Electronic Commerce. ICEC*, pp. 429–436. url: <https://dl.acm.org/doi/10.1145/948005.948061>.
- Osterwalder, Alexander, Yves Pigneur e et al. (2015). «Value proposition design: How to create products and services customers want.». Em: *John Wiley and Sons 2*.
- Paschalidi, Charikleia (2015). «Data Governance: A conceptual framework in order to prevent your Data Lake from becoming a Data Swamp». Em: pp. 16–20.
- Patrizio, Andy (dez. de 2018). «IDC: Expect 175 zettabytes of data worldwide by 2025». Em: *Networkworld*. url: <https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html>.
- Peter, A. e Koen (2002). «Fuzzy Front End: Effective Methods, Tools, and Techniques». Em: *Marketing management*.
- Pittet, Sten (2023). «Types of software testing». Em: url: <https://www.atlassian.com/br/continuous-delivery/software-testing/types-of-software-testing>.
- Practices, B. e Y. C. Data (2023). «Introduction To Data Governance». Em: *Data Governance Study*.
- Quality-One (dez. de 2017). «QFD: Quality Function Deployment». Em: 62. url: <https://quality-one.com/qfd/>.
- Ravald, Annika e Christian Grönroos (fev. de 1996). «The value concept and relationship». Em: *European Journal of Marketing* 30, pp. 19–30. url: https://www.researchgate.net/profile/Christian-Groenroos/publication/235270562_The_value_concept_and%20_relationship/links/0deec5224f05e75b8f000000/The-value-concept-and-relationship.pdf.
- Ray, Dewey E. (2012). «The IT professional's merger and acquisition handbook». Em: *Cognitive Diligence*.
- ReadingCraze (2015). «Hypothesis formulation in research». Em: url: <http://readingcraze.com/index.php/hypothesis-formulation-research/>.
- Rich, Nick e Matthias Holweg (2000). «Value analysis». Em: *Value engineering*.
- Saaty, R.W. (1987). «The analytic hierarchy process—what it is and how it is used». Em: *Mathematical Modelling* 9.3, pp. 161–176. url: <https://www.sciencedirect.com/>.
- Saaty, Thomas L. (1988). «What is the Analytic Hierarchy Process?» Em: *Mathematical Models for Decision Support*, pp. 109–121.
- SAP (2023). «SAP Master Data Governance». Em: url: <https://www.sap.com/products/technology-platform/master-data-governance.html>.
- Sarsby, Alan (2016). «SWOT ANALYSIS: A GUIDE TO SWOT for business studies students». Em: *Spectaris Ltd*,
- Smallwood, R. (ago. de 2014). «Defining the Differences Between Information Governance, IT Governance, and Data Governance.». Em: *AiIM Community*. url: <http://community.aiim.org/blogs/robert-smallwood/2014/08/18/defining-the-differences-between-information-governance-it-governance--data-governance>.
- Software-Testing-Help (abr. de 2023). «What Is User Acceptance Testing (UAT): A Complete Guide». Em: url: https://www.softwaretestinghelp.com/what-is-user-acceptance-testing-uat/#What_Is_User_Acceptance_Testing.

- Tang, Mingjie et al. (2019). «SAC: A System for Big Data Lineage Tracking». Em: *IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 1964–1967.
- Tejada, Z. (2018). «Big data architectures». Em: url: <https://learn.microsoft.com/en-us/azure/architecture/data-guide/big-data/>.
- Teoli, Dac, Terrence Sanvictores e Jason An (2021). «SWOT Analysis». Em: *StatPearls Publishing, Treasure Island (FL)*. url: <http://europepmc.org/books/NBK537302>.
- Teperi, Anna-Maria, Nadezhda Gotcheva e Kirsi Aaltonen (2021). «16 - Design thinking perspective for developing safety management practices in nuclear industry». Em: *Human Factors in the Nuclear Industry*, pp. 309–326. url: <https://www.sciencedirect.com/science/article/pii/B9780081028452000168>.
- Thakur, Dinesh (nov. de 2013). «What is software requirement? Types of requirements». Em: url: <https://ecomputernotes.com/software-engineering/softwarerequirement>.
- Thomas, G. (2006). «Alpha Males and Data Disaster.» Em: *Brass Cannon Press*.
- Uлага, Wolfgang e Samir Chacour (jan. de 2001). «Measuring Customer-Perceived Value in Business Markets: A Prerequisite for Marketing Strategy Development and Implementation». Em: *Industrial Marketing Management* 30.6.1, pp. 525–540. url: <https://www.sciencedirect.com/science/article/abs/pii/S0019850199001224>.
- Wamars, Harco e Richard Randriatoamanana (nov. de 2016). «Datawarehouse: A Data Warehouse artist who have ability to understand data warehouse schema pictures». Em: *IEEE Region 10 COnference*, pp. 2205–2208.
- WATT, ADRIENNE (2014). «DATABASE DESIGN». Em: *BC Campus 2*. url: <https://opentextbc.ca/dbdesign01/chapter/chapter-5-data-modelling/>.
- Wong, Doris Hooi-Ten, Nurazean Maarop e Ganthan Narayana Samy (2020). «Data Governance and Data Stewardship: A Success Procedure». Em: *8th International Conference on Information Technology and Multimedia (ICIMU)*, pp. 54–61.
- Woodall, Tony (jan. de 2003). «Conceptualising 'Value for the Customer': An Attributional, Structural and Dispositional Analysis». Em: *Academy of Marketing Science Review*.
- Zaptest (2023). «O que são testes de integração? Mergulho profundo nos tipos, processo e implementação». Em: url: <https://www.zaptest.com/pt-br/o-que-sao-testes-de-integracao-mergulho-profundo-nos-tipos-processo-e-implementacao>.

Apêndice A

Tipos de Ativos e as suas Relações

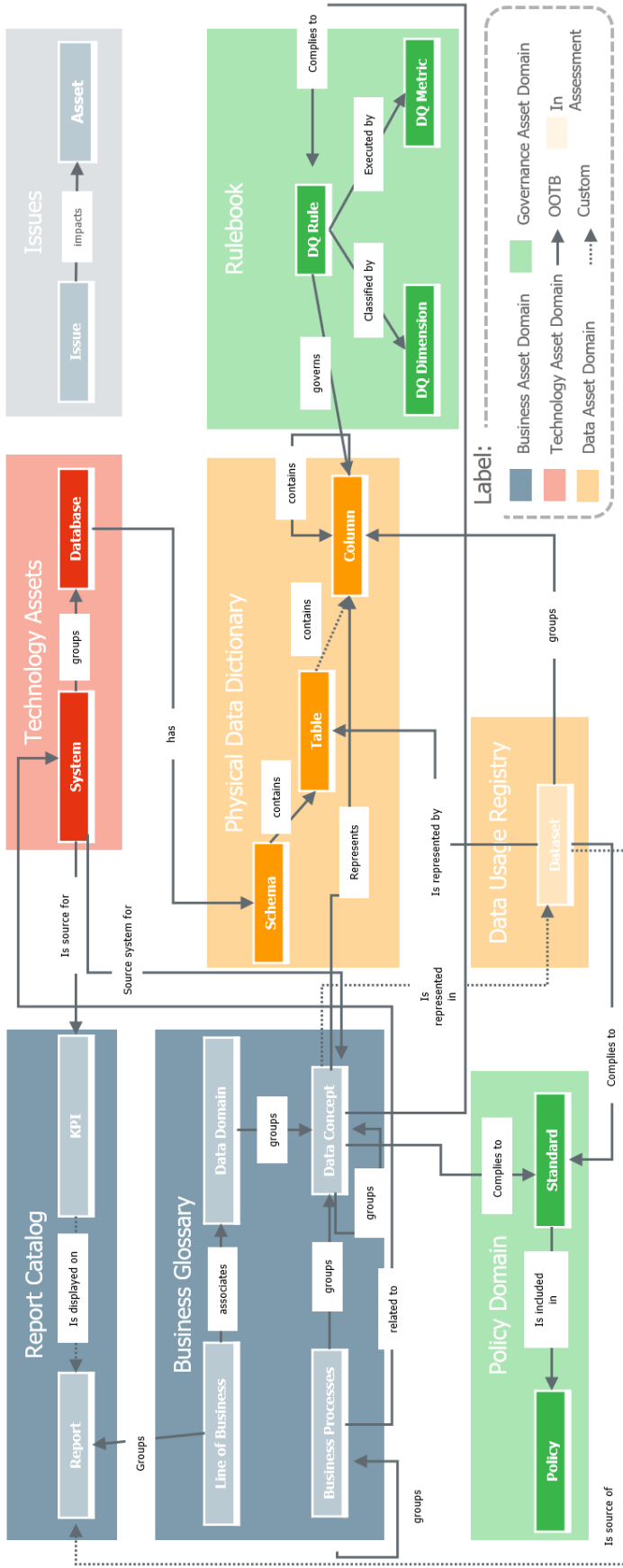


Figura A.1: Tipos de Ativos e as suas Relações