Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-graduação em Engenharia de Software
Mestrado Profissional em Engenharia de Software

OpenEasier: A CKAN Extension to Enhance Open Data Publication and Management

Jonas Jordão de Macêdo

Natal-RN
August 2018

Jonas Jordão de Macêdo

OpenEasier: A CKAN Extension to Enhance Open Data Publication and Management

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Software da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de Mestre em Engenharia de Software.

Universidade Federal do Rio Grande do Norte – UFRN

Instituto Metrópole Digital – IMD

Programa de Pós-Graduação em Engenharia de Software

Supervisor: Dr. Frederico Araújo da Silva Lopes

Co-supervisor: Dr. Nélio Alessandro Azevedo Cacho

Natal / RN

August 2018

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Central Zila Mamede

Macêdo, Jonas Jordão de.
OpenEasier: a CKAN extension to enhance Open Data publication and management / Jonas Jordão de Macêdo. - 2018.
90 f.: il.

Dissertação (mestrado) - Universidade Federal do Rio Grande do Norte, Instituto Metrópole Digital, Programa de Pós-Graduação em Engenharia de Software. Natal, RN, 2018.
Orientador: Prof. Dr. Frederico Araújo da Silva Lopes.
Coorientador: Prof. Dr. Nélio Alessandro Azevedo Cacho.

1. Open Data Publication - Dissertação. 2. Open Data tools - Dissertação. 3. CKAN - Dissertação. 4. Publication by non-IT technicians - Dissertação. I. Lopes, Frederico Araújo da Silva. II. Cacho, Nélio Alessandro Azevedo. III. Título.

RN/UF/BCZM CDU 004.42

Elaborado por Kalline Bezerra da Silva - CRB-15/327

Jonas Jordão de Macêdo

OpenEasier: A CKAN Extension to Enhance Open Data Publication and Management

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Software da Universidade Federal do Rio Grande do Norte como requisito parcial para a obtenção do grau de Mestre em Engenharia de Software.

Trabalho Aprovado. Natal / RN, 7 de Agosto de 2018:

Dr. Frederico Araújo da Silva Lopes
UFRN
Supervisor

Dr. Nélio Alessandro Azevedo Cacho
UFRN
Co-supervisor

Dr. Jair Cavalcanti Leite
UFRN
Examinador

Dra. Bernadette Farias Loscio
UFPE
Examinador

Natal / RN
August 2018

"Good Timber

The tree that never had to fight
For sun and sky and air and light,
But stood out in the open plain
And always got its share of rain,
Never became a forest king

The man who never had to toil
To gain and farm his patch of soil,
Who never had to win his share
Of sun and sky and light and air,
Never became a manly man
But lived and died as he began.
But lived and died a scrubby thing.

Good timber does not grow with ease:
The stronger wind, the stronger trees;
The further sky, the greater length;
The more the storm, the more the strength.
By sun and cold, by rain and snow,
In trees and men good timbers grow.

Where thickest lies the forest growth,
We find the patriarchs of both.
And they hold counsel with the stars
Whose broken branches show the scars
Of many winds and much of strife.
This is the common law of life."

(Douglas Malloch)

Abstract

Open Data is an important concept for our society, and it is being adopted by public and private entities. When embracing Open Data, the companies generate more transparency and collaboration in our society, this enables the enhancement and creation of services, helping to improve many aspects of our lives. Despite the existence of data catalogue platforms to support Open Data, e.g. CKAN, the complexity and costs of achieving the publication of Open Data are still a challenge, hampering the adoption of the activity of publishing Open Data. The existing tools that support Open Data publication demands deep knowledge of IT tools to publish the data, leaving this important task in the hands of few. Hence, the existing tools are not able to properly achieve the main goal which is to make anyone able to publish and maintain Open Data. In this context, this work aims to design and implement a new tool to decrease the complexity and costs, and to make possible non-IT technicians to publish and manage their Open Data. We believe that this strategy will engage the real data producers in the Open Data movement, helping to improve the quality of Open Data.

Keywords: Open Data publication; Open Data tools; CKAN; publication by non-IT technicians.

List of Figures

List of Tables

List of Abbreviations

API Application Programming Interface

CKAN Comprehensive Knowledge Archive Network

CRUD Create, Read, Update and Delete

CSV Comma Separated Values

DQ Data Quality

DSRM Design Science Research Methodology

ETL Extract, Transform and Load

GUI Graphical User Interface

IFRN Instituto Federal do Rio Grande do Norte

INDA Infraestrutura Nacional de Dados Abertos

IT Information Technology

JSON JavaScript Object Notation

OD Open Data

OGD Open Government Data

OKFN Open Knowledge Foundation

ORDBMS Object-relational Database Management System

ORM Object-relational Mapping

PDA Plano de Dados Abertos

SQL Structured Query Language

SUAP Sistema Unificado de Administração Pública

XML Extensible Markup Language

Contents

1 Introduction

Information has become an essential ingredient for the advancement of our society, driving us to considerable improvements in the ways we live our lives. This helps to intelligently execute simple tasks, as buying the best product in the market, or complex ones, as taking an important financial decision in a multinational company.

Progresses in the Computer Science field, such as the Internet, are the cause of a long-lasting relationship between computers and information. By providing methods to collect, store and process data through software approaches, computers make possible to work with high amounts of data, helping us to extract meaningful information. An assignment to costly and almost impracticable to humans in nowadays scale.

In the past few years, the openness movements became popular in our society, leading us to a collaborative mindset in our working approaches. Alongside with the huge amount of important information produced by institutions, and the concept of openness there are the Open Data (OD) initiatives, endorsed by private and public entities, which aims to freely use, reuse, and redistribute data without any type of restriction (MOLLOY, 2011). It is also required to follow a set of rules to make sure data is accessible and easy to be used through software means.

Open Data initiatives leads to many benefits, accomplishing enhancements in political, social, economic, operational and technical aspects of our lives (JANSSEN; CHARALABIDIS; ZUIDERWIJK, 2012). Generating more transparency in the government, the citizens are empowered to push progress through innovations and optimizations of the services and products (LAKOMAA; KALLBERG, 2013) (KUCERA; CHLAPEK, 2014). Creating an environment of collaboration and innovation, Open Data works as a fuel to Smart Cities, being of great relevance when biding Smart Devices to information in an open approach, generating benefits associated to all aspects previously cited (OJO; CURRY; ZELETI, 2015) (DOMINGO et al., 2013).

Aligned with the openness ideals, the Open Knowledge Foundation (OKFN)[1] is a non-profit organization which seeks to create tools and spread insights concerning the Open Data movement, making easier to public and private entities work on their Open Data strategies, not having to focus much of their effort on which ways to publish the produced data. Thus, only following the previous defined standards to adopt the movement.

Due to the results of the OKFN endeavor, the Comprehensive Knowledge Archive Network (CKAN)[2] system has been developed since March 2006, providing ways to

[1] https://okfn.org
[2] https://ckan.org/

publish, share and work with data through a data portal platform (WINN et al., 2013). By abstracting Open Data concepts, the CKAN platform enables others to open their data, or use data already open, in a software-oriented approach. Currently, CKAN is one of the most used platform around the world, officially adopted by countries like USA, Brazil and many other[3]. In fact, many Brazilian public and private institution adopted CKAN to publish their own data.

Open Government Data (OGD) is the concept of Open Data applied in the public domain (MATHEUS; RIBEIRO; VAZ, 2015), where countries create laws and decrees to foster Open Data in the public institutions. Brazil supports OGD through the law nº12.527 (Access to Public Information Act - LAI) (JARDIM, 2013) and the decree nº8.777 (BRASIL, 2016), which seeks for transparency in the Brazilian government. The law nº12.527 and the decree nº8.777 institutes the Open Data Polices to guarantee free access to public information, establishing procedures to be followed by Brazilian public institutions, coordinating and creating a schema in order that the institutions can properly attempt the OGD specifications. To properly establish patterns, create technologies, and procedures related to Open Data in the Brazilian context, there is the Infraestrutura Nacional de Dados Abertos (INDA)[4], which is a national group of institutions with the mission to disseminate and share procedures related to OGD in the Brazilian circumstances. Although CKAN is not the official platform adopted by INDA, it is largely used by the public institutions. A good example is the Portal de Dados Abertos Brasileiro[5], the main portal of Open Data in Brazil.

The Secretaria de Estado da Agricultura, da Pecuária e da Pesca (SAPE)[6] and the sub entities are looking for ways to open the data produced by the institutions, aiming to improve the offered services and to make the citizens aware of the accomplishments, trying to create a collaborative environment, looking to achieve the previous Open Data benefits presented. Attempting the law nº12.527 and the decree nº8.777. For instance, SAPE is a public entity which focuses on promoting improvements in qualitative and quantitative aspects of rural production at the state of Rio Grande do Norte (RN), Brazil. To achieve such improvements, EMATER-RN[7], EMPARN[8], IDIARN[9] and CEASA[10] are sub entities coordinated by SAPE to elaborate and execute the Public Policies, allowing SAPE to attain a broader range of policies. Thereby, to achieve the best of Open Data, the SAPE and sub entities chosen to use CKAN, because it is the platform adopted by the Brazilian government, and contains the main features necessary to attain what is addressed by the

[3] http://ckan.org/instances/
[4] http://wiki.dados.gov.br/
[5] http://dados.gov.br/
[6] http://www.sape.rn.gov.br/
[7] http://www.emater.rn.gov.br/
[8] http://www.emparn.rn.gov.br
[9] http://www.idiarn.rn.gov.br/
[10] http://www.ceasa.rn.gov.br/

transparency law and it is also the platform indicated by INDA.

1.1 Problem Statement

Despite the existence of such platform as CKAN, there are still many challenges in the data publication's process. As a result of the high technical complexity and the ethical concerns involved to open the data, many institutions renounce the activity (KUCERA; CHLAPEK, 2014). The process to release data is usually complex because it relies on specialized Information Technology (IT) technicians to be executed (UBALDI, 2013) (SAYOGO et al., 2014). This factor increases municipality' costs, and hampers the opening data process and fosters data silos.

According to Araújo (2017), there are technical barriers in the process of publishing Open Data. Deficiency in technical knowledge from the IT technicians; lack of human resource; unsatisfactory tools to support the process of publishing Open Data; and difficulties in the maintenance of Open Data. Those are the most relevant problems related to the publication of Open Data exposed by Araújo's research (2017), leaving aside ethical and cultural problems.

For instance, CKAN provides two ways to publish Open Data: (i) by a Web-based interface, where you can create and manage Datasets and their Resources[11]; (ii) and through an Application Programming Interface (API) that provides a set of methods to interact with all the CKAN's core features[12].

Managing the datasets through the CKAN Web-based interface is an error prone task, where the user can easily miss place a resource, or not properly keep the data up to date. Furthermore, when treating a high amount of different datasets this work becomes too costly, having to put specific individuals to execute this operation repeatedly, been only acceptable in situations where there is little quantity of data to be published and managed. Which is not the situation encountered in the SAPE context, where there is not enough work force to execute this task, nor little amount of data to be published and managed.

While the CKAN API is a smarter solution to use when publishing data to CKAN, it is still not an easy path to follow, requiring a certain amount of effort and skills to integrate the institutions' system, or systems, with the CKAN API. Where the complexity can vary depending on the many different software approaches the institutions has, and the level of heterogeneity in the technologies used.

Hence, there is the need for a specialized tool which supports Extract, Transform and Load (ETL) features (KIMBALL; CASERTA, 2011). The solution must extract data

[11] http://docs.ckan.org/en/latest/user-guide.html#features-for-publishers
[12] http://docs.ckan.org/en/latest/api/

from the systems' sources, transform it if necessary, and load it to CKAN through the API. This creates the need of a skilled IT professional to achieve the data publication, demanding a higher cost to create and maintain the solutions.

Although right now there is not much heterogeneity in the systems developed by the SAPE's IT team, there is still the problem of scalability and cost, where every time a new type of resource is introduced to the system, like a new way to register activities, it's going to be needed to build a new bridge to create communication between the CKAN API and the new system's resource. Creating the necessity of more workforce to achieve the task, thus, increasing the cost of opening the data. Being something that the institutions can not afford, preventing the publication of Open Data.

1.2 Motivation

Seeking to provide an user centered approach in the publication of Open Data, this master's thesis aims to contribute with the creation of a tool. The tool aims to enable the business technicians to manage the whole process in the publication of the data produced by themselves, with little intervention of the IT technicians in the process. This will decrease the need of knowledge regarding computer technology, allowing SAPE and SAPE's sub entities to open their data without much struggle and need of great human resources. The business technician is any employee that knows about the business rules and it is also producing or collecting the data of the institution, being aware of the data and the importance of it, making the business technician an important piece in the process of opening the data.

The main benefits of this work is to make simple the process of publishing and managing Open Data, by abstracting the necessary concepts to achieve this task. It is also crucial to provide efficiency and reliability while using the tool to configure the right data to be published. Another import aspect is to make possible to schedule the period of data synchronization, ensuring that the data is always up to date, helping in the process of maintenance of Open Data. By granting more independence to the business technicians, the need of specialized IT technicians is decreased, lowering the costs of the activity.

Another impact that the tools aims to accomplish is related to scalability. The complexity of a system is always growing according to the company's needs, creating more data diversity through this process. With little setting from the IT technicians, a new module is ready to be used by the business technician, allowing to open the intended data, not needing to create a new bridge with CKAN to make the data available to others.

In order to reduce the complexity and costs of publishing Open Data to CKAN, we hope that other organizations will be able to open their data, encouraging the Open Data movement. Looking forward for the great benefits generate to such activity, helping to

improve many aspects of our lives (KUCERA; CHLAPEK, 2014). Other expected result is to foment the engagement from the real producers of the data, motivating them to improve the data gathered. Through a data quality feedback from the tool, showing which portions of the data must be improved, the business technician will be capable to know which portions of the data must be refined. Trying to solve another important problem exposed by Araújo (2017), which is the lack of quality in the data opened.

1.3 Objectives

The main aim of this work is to develop a tool which allows the business technicians to manage and publish data with minimum dependency from the IT technicians.

The specific goals of this work are the following:

- Investigate the existing methods to publish Open Data, focusing on the CKAN platform;
- Specify and design a new tool to improve the activity of publishing and managing Open Data in the CKAN platform;
- Develop a tool which enables business technicians to manage the data publication;
- Define qualitative and quantitative metrics to evaluate and validate the proposed tool;
- Apply an usability experiment to assess the tool based on the previous metrics elaborated.

1.4 Methodology

The proposed research begun with a literature review, aiming to better understand the state of the art regarding to Open Data publication tools. This phase of the research also aimed to find the current problems in Open Data publication in the CKAN platform, likewise, focusing on which were the existing state of the art alternatives to solve the problems. This allowed us to gather insightful information about the key criteria to achieve the objectives. This activity was crucial to define which are the key features to abstract in the process of publishing Open Data.

The search for papers and tools was executed through the following academic search engines: Google Scholar, ACM Digital Library and IEEE Xplore Digital Library engines. We used the combinations of the following key words: *open data*, *publication*, *publishing*, *CKAN*, *tool*, and *automation*. A search using the same keys was executed in GitHub's

repositories, aiming to find works related to Open Data publication. The goal of that search was to find papers and available tools related to Open Data, and the publication process.

To support this initial phase of the research, and improve the understanding in the process of publishing Open Data, an evaluation of the related tools encountered was executed. Providing a measurement of the current works, helping to understand the existing methods and problems related to it. The evaluation is discussed and presented in the chapter 3. It was also important to understand the aspects in the Brazilian's OGD context exposed by Matheus *et al.* (2015), identifying patterns adopted in the publication of Open Data.

The main research method chosen to achieve the expected results is the Design Science Research Methodology (DSRM), which aims to build and evaluate the proposed artefact (MARCH; SMITH, 1995). The artefact proposed by this research was designed aiming to solve the problems, or to improve the actual process of Open Data publication. Such method is proved to be a good choice when related to engineering research activities, following a well-defined methodology to make research contributions (PEFFERS et al., 2007).

The DSRM process model has five steps to achieve the research's results, which are: (i) identification of the problem and the motivation to solve it; (ii) definition and design of the artefact; (iii) development of the artefact; (iv) validation and evaluation of the artefact; and (v) conclusion and communication of the results (PEFFERS et al., 2007) (VAISHNAVI; KUECHLER, 2015). Figure 1 shows a overview of the DSRM process model.

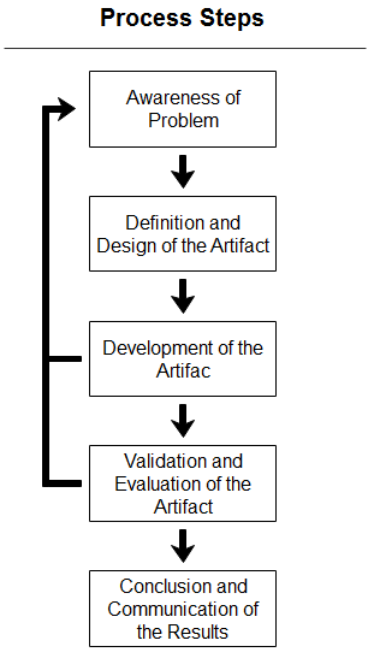
The first step was achieved after we understood the difficulties related to Open Data publication. The literature review and evaluation of the tools in chapter 3, provided a full understanding of the current situation. Thus, this step was accomplished by that effort, being essential for the definition of the goals of this dissertation.

Defining and designing the tool based on the previous knowledge acquired was necessary to achieve the second step. Thus, creating a software tool to simplify the process of publishing Open Data. The tool created enabled the business technician to execute the task of publishing and managing Open Data.

After defining and designing the tool, the activity of creating the tool started, following an agile development methodology. Trying to achieve a first stable version, with the minimum of features for the users' satisfaction, in order that the study can proceed to the next stage of evaluation.

Elaborating the metrics is one of the finals steps in the DSRM, which are going to be used to evaluate the artefact (the tool) created in the process of researching. Providing

Figure 1 – DSR Process Steps



Font: Adapted from Vaishnavi *et al.* (2015)

quantitative and qualitative feedback from the resulted artefact. To assist the measurements of the elaborated metrics, an usability experiment was executed to better understand the impacts of the new tool in the process of Open Data publication. Therefore, helping to validate the proposed artefact, extracting the final results of the research. Providing knowledge to identify weaknesses and areas of improvements (VENABLE; PRIES-HEJE; BASKERVILLE, 2012), if there are any. As the final activity, results will be presented and discussed, as a mean to achieve the last step of communication necessary in the DSRM. Thus attempting to all steps of the DSRM.

1.5 Chapters Overview

This section presents a short description about each specific chapter of this dissertation.

Theoretical Background Overview. Describes the concepts and technologies applied along the research development, and the reasons of adoption of the technologies and methods chosen.

Related Works Overview. Describes the related work conceived as important to the research, approaching the pros and cons of each one.

Design Overview. This chapter presents a description of the process of designing the tool.

Development Overview. Describes and documents the development process, showing the approach to achieve the creation of the tool to support, the technologies, and the tool itself.

Evaluation Approach Overview. Presents the metrics chosen to validate the tool, why those metrics were chosen, and which methods were applied to test those metrics. Lastly, it also presents how the experiment of evaluation was applied, along with the results.

Final Considerations. Presents a review of the conclusions, limitations and the future works of the study.

2 Theoretical Background

This chapter introduces the main concepts related to the research goals, with the intention to provide context to the readers. We start the chapter with section 2.1, presenting aspects about Open Data and Open Government Data in the Brazilian context. The section 2.2 is an introduction to CKAN, exposing general and technical aspects considered as important to the research, related to the platform. In section 2.3 we explain the process of Extract, Transform, and Load (ETL). Such process is pertinent to the activity of publishing Open Data. The section 2.4 gives a brief introduction about Object-relational Mapping (ORM) and the relevance of it to the study. In the section 2.5 we present the concept of Data Quality (DQ) and the importance of applying it to the study, so it can improve the data being published.

2.1 Open Data

The general definition of Open Data (OD) is "data that can be freely used, re-used and redistributed by anyone" (DIETRICH et al., 2009). But to better understand the specificities and technical aspects of the subject here discussed, the Open Definition[1] gives a more detailed description. The main requirements of an Open Data Work defined by the Open Definition are:

- The data must have an Open License or have the absence of copyrights and similar restrictions;
- The data must be provided via the Internet without charge, and must be complete, not restricting part of it to a specific public;
- The data must be provided in a machine readable form, following patterns to make access easier;
- The data must be in an open format, where it can be read by an open-source software tool.

Following the requirements presented at the previous list, a public or private entity can make data open and available to all, creating value in many aspects of our lives. Janssen *et al.* (2012) highlights the many benefits that the adoption of Open Data can provide. The main categories those benefits are clustered are: (i) political and social; (ii) economic; and (iii) operational and technical aspects.

[1] http://opendefinition.org/od/2.1/en/

It is possible to provide improvements and innovation in services and processes through Open Data, those improvements are categorized as political and social benefits to our society, creating a more democratic and transparent government. In the economic view, there is growth and stimulation of the competitiveness, with the development of new products and services from the information available. This allows the creation of new economic sectors, thus adding value to the services provided to the society in general. The main benefits related to operational and technical aspects are the capability to reuse data, the creation of new data based on combination, and the external view of others to apply problem solving thinking with insights from the data. Thereby, with the explanation provided, it is assured the importance of Open Data to us as a society, demonstrating the importance of the topic to the research.

2.1.1 Open Government Data

As a subcategory of OD, Open Government Data (OGD) aims to benefit the society creating a more transparent government, allowing the citizens to be more aware of the services provided and activities executed by the government (UBALDI, 2013). Since the public entities produce a large amount of data about the surroundings that involves the society, because of the public services provided, opening the data (as OGD prescribes) the government and the citizens could both benefit themselves with all the improvements that Open Data can provide.

2.1.2 Infraestrutura Nacional de Dados Abertos

The Infraestrutura Nacional de Dados Abertos (INDA)[2] [3] is a set of patterns, technologies, procedures and mechanism of control to promote the sharing and use of Open Data in the Brazilian government. The main contribution of INDA is the Portal Brasileiro de Dados Abertos[4], which is the official Open Data portal of Brazil, gathering data from many public institutions[5].

The Plano de Dados Abertos (PDA) [6] is a document created by INDA, so the public institutions can use this document as a guide to create a plan describing how and which data will be open. In the document, the institution must describe which data will be published, and must also follow the patterns established by INDA.

[2] http://wiki.dados.gov.br/
[3] INDA can be directly translated to National Infrastructure of Open Data
[4] http://dados.gov.br/
[5] http://dados.gov.br/organization
[6] PDA can be directly translated to Open Data Plan

2.1.3 Data Dictionary

One important aspect to facilitate the use of Open Data is the Data Dictionary. The Data Dictionary provides information about the data contained in a resource, making possible to understand general and technical aspects about the data being provided. We have not found any pattern provided by INDA on how to create a Data Dictionary, but we have analyzed the Data Dictionaries provided by the institutions at the Portal Brasileiro de Dados Abertos. Most of the dictionaries have at least the following metadata:

- A name informing the resource which the dictionary is created to;
- The version of the dictionary;
- The author or font of the dictionary, which can be a person or a company;
- The last date the dictionary was updated;
- and a general description about the purpose of the dictionary.

The dictionary also provides technical knowledge about the resource. Those technical knowledge are all about the data itself, it provides the name, type and size for each column. There is also a description of the column, possible values contained at the column, and if the column accepts null values. Most of the technical information provided by the dictionary, can be, and should be generated by a software, making more precise and easier to describe the data. Thus, the tool must provide ways to generate the Data Dictionary, this shows the importance of this topic to the study.

2.2 Comprehensive Knowledge Archive Network

The Comprehensive Knowledge Archive Network (CKAN) is a Web-based data portal platform, developed with the goal to provide ways to others publish, find and use Open Data in a software approach. CKAN is open-source, thus, it has a large community of developers improving the platform and creating new extensions, enriching the CKAN environment. The platform is a good choice to use as an Open Data provider, being the official choice by INDA to be the main tool in the Brazilian Open Data context.

2.2.1 Dataset

CKAN organizes data in units called datasets. The datasets contains information about a group of resources. It provides a title, date of creation and last update, formats of the data, owners of the dataset, and a description about the dataset[7]. For example, a

[7] ahttp://docs.ckan.org/en/latest/user-guide.html

dataset can be a group of resources that has the information of each year's spendings of a public institution.

2.2.2 Resource

A resource is any file contained in a dataset, it is a more specific subset of the dataset. It can be the data by itself, a data dictionary (metadata), a link and etc. It can be of different formats, as JSON, CSV, PDF and etc. CKAN is able to internally store the resource, or just point to a link where the resource can be found.

2.2.3 CKAN Application Program Interface

The CKAN Application Program Interface (API)[8] provides a complete set of features so other applications can interact with CKAN. It is possible to manage datasets, resources, groups, organizations and users, with the necessary authentication. Applications from external sources can also interact with the CKAN API without the need of authentication, but having limited functionalities, as executing simple queries to retrieve datasets information, or complex queries to extract a specific set of data from a resource. The DataStore [9], for example, is an important official extension for CKAN, which can be used through the CKAN API. When the extension is installed in the CKAN instance, an authenticated user can send a spreadsheet, as CSV or Excel, and the extension will read it, create a table in the internal database using the columns of the spreadsheet and store the data. By doing that, CKAN provides the data through the CKAN API so unauthenticated users can query it in a software approach, without the need of downloading the file to read the data. This is one of the main features that the CKAN API provides, when storing the data from a spreadsheet in the CKAN's database it makes possible to other query those data.

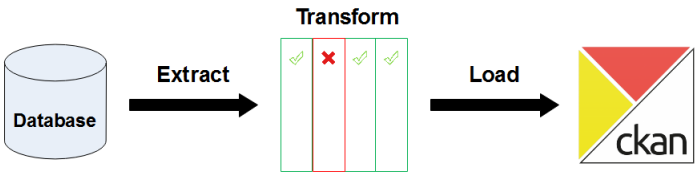
2.3 Extract, Transform and Load

The Extract, Transform and Load (ETL) processes can nowadays be defined as any software developed with the following basic responsibilities: (i) extract the data from one or more sources; (ii) execute the necessary transformation and cleansing; and (iii) load the data to the appropriated destination (VASSILIADIS, 2009). This is an activity executed in most of the cases to extract data from different systems, and properly load it to a data warehouse; when publishing Open Data in an automatic approach using software it is used a similar process, as mentioned in the section 1.1. An IT-technician must create an ETL script to achieve the publication of Open Data to the CKAN instance. Figure 2 presents

[8] http://docs.ckan.org/en/latest/api/index.html
[9] http://docs.ckan.org/en/latest/maintaining/datastore.html

an overview of the interoperability of ETL features and CKAN. Since the main subject of the research is Open Data, we present each aspect of the ETL process relating to the OD context. Further more, we consider that the 8 Principles of Open Government Data[10] as a fundamental part of this section.

Figure 2 – ETL and CKAN interoperability overview



Font: Authors

The first step is the extraction, which can be performed from many sources (e.g., a database-management system, tabular files and others). In the OD context, it is more suitable to extract the data from a live source, thereby, a database-management system is the best option, where it will ensure that the data being extracted is the latests version. This will assure the *Timely principle*, which holds the definition that the data must be available as quickly as possible, preserving the importance of the data. The second step is the data transformation, but as the *Primary principle* defines, there should be no aggregation or modification of the data. Thus, the only tasks to be executed in the transformation step is related to the remotion of portions of the data (columns of the table) that must be no public, or transforming the data to many formats, e.g., CSV, JSON and etc. The final step is the loading, which is the simplest of all three steps presented. In this step, the data must be inserted or updated into the CKAN instance to complete the ETL process, publishing the data at the end.

2.4 Object-relational Mapping

The technique of ORM will be used at this work as a way to do a live mapping of the tables and columns of a database. This will allow the system to present the data as objects to the non-IT technician, providing a simple way to choose the tables and columns wanted to have the data published.

2.5 Data Quality

The most well known definition of Data Quality (DQ) is: data that *fits for the use*. But the assessment of the quality of the data can be really complex, there are multi-

[10] https://opengovdata.org/

dimensional aspects to handle while checking the quality (PIPINO; LEE; WANG, 2002). When assessing the level of quality of the data, it is necessary to choose dimensions as metrics to guide the process, and those dimensions are often domain-specific. Thus, it is a complex and critical activity (BATINI et al., 2009).

The task of creating the domain-specific metrics is arduous. It is necessary to analyze the whole context of a company to define those metrics. The domain where the data will be used must be well defined, assuring that the data fits for the use in this domain. And there is also the need of technical knowledge to execute it. Thereby, since the purpose of the research is to produce a tool for non-IT technicians to publish and manage Open Data, this topic was worked carefully during the design of the tool.

3 Related Works

With the intention of presenting the works directed related to the tool proposed in this master's thesis, this chapter describes the objective of each work, and also provide an overview of the supported features. The works presented here were conceived as important to the development of the tool, trying to achieve a higher level of what is already done, providing a more abstract way to publish Open Data, focusing in the CKAN.

Few solutions were found, some not even achieving the minimum criteria to be considered as related work since they only presented implementations where the goals were to make easier to communicate with the CKAN API (being considered simple wrappers of the API) [1], not even having ETL capabilities.

The works here presented and discussed were not defined by the authors as tools with the main user being the business technician. By studying each tool, it was clear that the tool's features were only restricted to abstractions to the ELT process, relating with interoperability with the CKAN platform, and simplifying the task of the IT technician in the Open Data publication process, but still needing to code in most of the cases.

Although the analysis of the FME-CKAN Connector [2] is not presented in this work, an experimentation was performed, but without success, not being able to properly use the tool, because it was specific for the context of the Australian Government's CKAN instance. Thereby, this tool will not be discussed in this chapter.

3.1 WPRDC-ETL

The Western Pennsylvania Regional Data Center (WPRDC) [3], from University of Pittsburgh[4], has an Open Data Portal that provides technological and legal infrastructure for data sharing, allowing the Allegheny County and the City of Pittsburgh to publish their Open Data[5].

The portal uses CKAN as the platform to abstract the Open Data services needed, and while the amount of data being published was growing, the IT team started to realize that the ETL processes used needed to become automated, to attend the needs of the portal[6].

1 https://github.com/ckan/ckan/wiki/CKAN-API-Clients
2 https://github.com/OpenCouncilData/FME-CKAN
3 http://www.wprdc.org/
4 http://www.pitt.edu/
5 http://www.wprdc.org/about/
6 https://www.wprdc.org/news/automated_etl/

WPRDC-ETL is an open-source Python library that abstracts the process of ETL to the CKAN platform. This solution was used in the WPRDC Open Data Portal, allowing the IT team to easily design pipelines to automate the process of creating and maintaining datasets, by defining the data to be published and scheduling the update frequency of the Dataset, also helping to monitor the process through a simple log.

As the own developers of the tool described[7], the tool is designed for the use of the IT team. I.e., this tool is not suitable for the data producers (business technician). However, this project showed importance to the research, abstracting some of the features to introduce, or improve, in the final work. Further information will be presented in the detailed analysis topic of the tools.

3.2 CKAN Automator

The CKAN Automator is a non-official extension for the CKAN platform, developed by the Open North[8] a nonprofit organization which seeks to develop Open Data solutions to achieve better and more open democracies.

The extension wrappers features of the CKAN API into a Python library, simplifying the management of groups, packages and resources. It also provides a way to upload CSV files contained in a local directory, by using a metadata file in JSON format, the programmer can specify the information about the resources to be uploaded, setting things such as which dataset the resource makes part in the CKAN instance[9,10].

The extension is use-only by IT team since there is the need to code. However, it is a important to study in our research process.

3.3 Open Data Node

Open Data Node (ODN)[11] is an open-source Web platform with ETL features which includes tools to help publish, manage and exchange Open Data, using CKAN in the background to provide some of this processes. The tool is supported by the Methodology for Open Data publication (KUCERA et al., 2015), providing guidelines to achieve the best in Open Data, being developed by the COMSODE project[12].

With a more broad and integrated approach, the Open Data Node uses CKAN as the internal data catalog, making possible to open the data through it. The main integrated

[7] http://wprdc-etl.readthedocs.io/en/latest/
[8] http://www.opennorth.ca
[9] http://extensions.ckan.org/extension/ckan-automator/
[10] https://github.com/opennorth/ckan-automator
[11] https://www.opendatanode.org
[12] www.comsode.eu/

tool to the ODN is the UnifiedViews[13], which allow users to create ETL processes, using an interface, to publish the data in the internal catalog, also having support to linked data (KNAP et al., 2015), proving to be a more mature solution.

Although the creators of the platform explain that the main user is the business technician[14], the process of publishing Open Data proved to be, in some level, very hard. It still requires technical skills to achieve the task, in which a good level of knowledge is necessary to understand the tool and to design SQL queries to extract the data to be published.

3.4 Pentaho CKAN

Pentaho Data Integration (PDI) [15], also known as Kettle, is a powerful desktop ETL tool, where it is possible to create jobs through a Graphical User Interface (GUI), automating the processes of Extracting, Transforming and Loading data. Working in the Loading layer of ETL, a custom plugin allows interoperability between PDI and the CKAN DataStore API, enabling to use PDI to Extract, Transform and Load tabular data to the CKAN's data catalog[16].

3.5 A taxonomy for Open Data publishing tools

To evaluate the Open Data publishing tools presented in chapter 3, a taxonomy was created based on features and quality attributes identified as important to achieve a good level of abstraction for municipality non-IT staff. The taxonomy comprises two groups: (i) the features group, which contains every functionality discerned as needed to manage the process of publishing and maintaining Open Data; and (ii) the quality attributes group, including all the attributes defined to achieve a higher level of abstraction in the Open Data publication process, to accomplish the goal where the main user is the municipality business technician. The attributes created for the taxonomy were based on the experience acquired on the literature review and also based on the hypothesis of the study.

3.5.1 Features Group

Understanding the process of publishing and managing Open Data, we have identified the following features as necessary to achieve this tasks: configure, extract, transform, load, pipeline, schedule, log and notify. In the next sub-subsections, we describe each cited feature.

[13] https://unifiedviews.eu/
[14] https://www.youtube.com/watch?v=FulhbmXFWqU
[15] http://www.pentaho.com/product/data-integration
[16] https://github.com/OpenGov-OpenData/CKAN-DataStore-Writer-for-Pentaho-Data-Integration

3.5.1.1 Configure

In the configuration feature, the main user is the IT technician. It is necessary to provide ways to easily configure different CKAN instances to have the data uploaded, if needed, and to manage the users (business technician) which will participate in the process of publishing Open Data. Is also part of the IT technician to configure each source of data that each business technician has access, i.e., the database connection information and accessible schema.

3.5.1.2 Extract

To extract the data, it is necessary the ability to connect to a database, and execute a query to fetch the data; by retrieving it from the main source it assures that it is the most recent version of the data produced, keeping the resource in consistency with the data that is in the production environment. The extracting process can be complex in some situations, this is why a quality attribute as the presented in the subsubsection 3.5.2.2 is so important. Extracting data from multiple sources, e.g. tables, is needed when publishing high quality open data, enriching the data.

3.5.1.3 Transform

The transformation process is necessary to have the ability to ignore some columns with restricted data, or transform the data into many formats, e.g., CSV; JSON; XML. Moreover, it is also important to create a data dictionary to describe the data type in each column. Furthermore, in cases where there is little variation, the data presented in the records can be listed in the data dictionary, as in a column to inform the sex of a person, where is always going to be only male or female. The data dictionary is necessary to help others to understand and use the resource, being a crucial part in the Open Data process.

3.5.1.4 Load

Loading is the final step in the ETL process, gathering the results from the previous steps and sending the data to a specific location, in our case, a specific CKAN instance. It is also important to emphasize that the load process should be able to execute an upset operation in the CKAN instance, inserting new data or updating the existing data when necessary.

3.5.1.5 Pipeline

The pipeline acts as a binder to all of the ETL processes, creating a single element which controls the execution of everything contained in it. Thus, the pipeline has the role to assure that each step is being executed in the correct order.

3.5.1.6 Schedule

Obsolete data is considered of little or no used in most of the situations. This is why scheduling the pipeline to execute periodically is so important, making sure the data came from a fresh source, and is good to be used, keeping the data up to date is an essential activity in the Open Data.

3.5.1.7 Log

The log feature must provide a simple way to check the results of each execution of the pipelines. And also some information like the latest changes in the pipeline, including the users who did the changes. It is also necessary to store every success or error occurred during the process of ETL. This feature is important for the user of the IT technician, to make possible to check if things are running properly.

3.5.1.8 Notify

The notification through e-mail of each execution of the pipeline is essential to keep track of the Open Data publication process. It allows the knowledge about what is being updated, or what went wrong and fix it. Therefore, it helps to maintain a resource in a good state.

3.5.1.9 Data Dictionary

The Data Dictionary is an important element in the publication of Open Data, being used in the Brazilian's Open Data Portals. The goal is to describe the structure of a resource in PDF format, being considered a metadata of the resource, this is of great importance to the users of the data. The Data Dictionary must have technical information, such as the columns' names, types and sizes, and also an explanatory description of the importance of the column to the business. Thus, the tool must be able to automatically generate the technical information mentioned, and allow the business technician describe each column of the resource.

3.5.2 Quality Attributes Group

The quality attributes group has the following criteria: Focus on the Business Technician; an Object-Relational Mapping capability; and a Data Quality Evaluation mechanism. Thus, it contains important attributes to provide enough abstractions, so the business technician can achieve the process of publishing and maintaining Open Data, having little dependency of specialized IT technician.

3.5.2.1 Focus on the Business Technician

The requirements to be a solution focused in the business technician is to not need to code, and not need technical knowledge in the computer field while publishing Open Data. This involves creating abstractions of the Open Data publication process. By dividing the process in small steps, and simplifying it, the non-IT technician should be able to select the data which he wants to be published. The main focus of this attribute is at the GUI, providing simply ways to execute the features presented in the subsection 3.5.1.

3.5.2.2 Object-Relational Mapping

To extract the data in a proper manner, it is necessary to have Object-Relational Mapping (ORM) quality attribute, to make possible to abstract the connection to any database, and present the data to the user. Showing the tables, tables' columns and the tables' data, it enables the business technician to choose the correct ones to be published.

3.5.2.3 Data Quality Evaluation

Consistency in data is an important matter in the Open Data movement. Therefore, it is crucial to have a tool which provides a mechanism of evaluation of the quality of the data being published. By helping the business technician to improve the data, focusing in aspects as completeness, reliability and accuracy, it is going to have a return in benefits to who uses the data. Such evaluation has the potential to keep the business technician engaged in the Open Data movement.

3.6 Tools Evaluation

This section uses the taxonomy defined in the section 3.5 to evaluate, through experimentation, each of the tools, creating the Table 1 as result. Weights were assigned to enable the measurement of each aspect contained in the tools, and also to calculate a complete dimension through the sum of each score, each tool only being able to achieve a max score of 22. The weights 0, +1 and +2, are respectively: does not fulfill, partially fulfill, and fulfill. The experimentation of the tools was by using the GUI provided, or coding when necessary, seeking to find the features, or attributes presented in the section 3.5, assigning the proper weights depending on the results of the tests. It is provided in the following subsections a detailed analysis of each category classified by the taxonomy.

Table 1 – Evaluation of Open Data Publication Tools for CKAN

Taxonomy Aspects		WPRDC-ETL	CKAN-Automator	Open Data Node	Pentaho CKAN
Features	*Configure*	+1	+1	+1	+2
	Extract	+1	+1	+2	+2
	Transform	+1	0	+2	+2
	Load	+1	0	+2	+1
	Pipeline	+1	+1	+2	+2
	Schedule	0	0	+2	+2
	Log	+1	0	+2	+2
	Notify	0	0	+2	+2
	Data Dictionary	0	0	0	0
Quality Attributes	*Focus on the Business Techinician*	0	0	+1	+1
	ORM	0	0	0	0
	Data Quality Evaluation	0	0	0	0
Final Score		**6**	**3**	**16**	**16**

Font: Authors

3.6.1 Features Group

3.6.1.1 Configure

Both the WPRDC-ETL and CKAN-AUTOMATOR tools achieved a low level of capabilities in the configuration criteria, thus, scoring 0 in this feature. The ways to specify the CKAN instance and the user API KEY to publish the data were through a simple configuration text file, with static variables to receive the CKAN instance's URL, and the user API KEY. This makes it difficult to extend and use a flexible solution, moreover, not allowing for multiple users participating in the publication process, nor it is able to publish data to many CKAN instance.

Although the Open Data Node easily supports many users, there is a problem in the ways things work. Because of the integrated solution, with the internal catalog, only users registered in the internal CKAN instance are capable of accessing the UnifiedViews tool to publish the data. Not being a good solution for those who already have a running CKAN instance, making completely adoption of the ODN the only simple way to properly use the features provided. The data will only be published at the internal catalog, not having direct ways to publish data to different CKAN instances, thus, receiving the +1 weight.

In Pentaho CKAN, while creating the ETL's pipeline, you can configure the CKAN instance which the data will be loaded, and also the CKAN user which is uploading the data. This has to be achieve through a manual approach, where each time the pipeline is created, those information must be provided. However, the Pentaho CKAN was the tool that achieved a closer solution of the *configuration* feature, therefore, scoring +2.

3.6.1.2 Extract

Because of the limits in the WPRDC-ETL and the CKAN-AUTOMATOR, the solutions only fulfilled partially the criteria, scoring +1. In both tools it is only possible to extract information from CSV files, not having the means to extract and combine from

different resources. There is the need of creating a script to do this task, extracting and preparing the CSV file to be used, proving to be a costly approach.

In a more flexible approach, the extract feature of ODN fulfilled the criteria, scoring +2. It is possible to collect and work with data from many sources, e.g., remote or local database, a remote file, or a local CKAN file from the internal catalog. But all of those options need technical knowledge in the computer science field, as knowing what is a relational database, or what is a HTTP request, being something difficult for the business technician to properly use.

It is possible to extract data from different sources with Pentaho CKAN, as database tables, XML files and others. But is necessary IT technical knowledge to achieve this task, e.g., using SQL queries to extract data from tables. BUt Pentaho CKAN also properly achieved the specifications in the *extract* criteria, thereby, scoring +2.

3.6.1.3 Transform

The CKAN-AUTOMATOR was not able to perform any type of transformation, only extracting and loading the data into CKAN, as a result, not fulfilling the criteria, scoring 0. The WPRDC-ETL tool had some capabilities of transforming data, e.g. formating dates to a previous pattern defined by the programmer. But those features did not fulfill completely the criteria, only providing simple methods to achieve simple transformations. Not properly attempting to the requirements in the *transform* criteria, thus, scoring only +1.

Using the UnifiedViews in ODN it is possible to perform 21 different types of transformation, with options as merging, converting, validating and treating the data. It is possible to merge data from different tabular or CSV resources, or also to validate if the XML is correctly formatted, and other options. Proving to properly attempt the criteria, scoring +2, but all of those tasks need a good level of computer technology to use those methods.

Pentaho CKAN's GUI gives many possibilities to transform data, being possible to select only specific values, split rows, sort and replace data. Properly attempting the *transform* criteria, scoring +2.

3.6.1.4 Load

The WPRDC-ETL achieved the *load* criteria, upserting the data correctly when specifying the key fields to be used as primary keys, being able to update the correct record. But in the CKAN-AUTOMATOR tool there was no method to upsert the resource, it only overrides the data completely, removing the existence resource in the dataset and inserting the new ones. Thus, to execute an update in a resource, the programmer must

create a complete new CSV file with the data contained in the resource and with the new record inserted or updated. Both the solutions are not able to load PDF, which is necessary when providing a data dictionary of the resource published. As a result, WPRDC-ETL and CKAN-AUTOMATOR scored respectively +1 and 0.

UnifiedViews has 10 methods to load data, being possible to store a file, RDF or relational data in to the CKAN internal instance, working properly with the provided features, achieving the last criteria of loading the data, scoring +2.

Data can be loaded to many places in Pentaho CKAN, but the method evaluated was the CKAN DataStore step provided by the plugin, where you can inform the CKAN instance and the user which is loading the data. Informing the package ID, resource title, description, resource ID and resource primary key, the data will be upserted. The only problem is that you can only load CSV files, because of the link with the CKAN DataStore API, therefore, you cannot load a PDF or XML file, only partially achieving the criteria, scoring +1.

3.6.1.5 Pipeline

The tools achieved a good level of abstraction when managing a pipeline, being able to encapsulate all the ETL process discussed before and execute it in proper order to achieve the final result of loading data into CKAN. The main difference is that in the WPRDC-ETL and CKAN-AUTOMATOR it can only be achieved by coding, therefore, both scored +1. In the ODN and Pentaho CKAN, you can use a GUI to create and configure a pipeline or job, thus, scoring +2.

3.6.1.6 Schedule

The WPRDC-ETL and the CKAN-AUTOMATOR are Python scripts with no features to schedule the execution, therefore, the only way is to use a job scheduler included in the operational system to run the scripts periodically, being not a scalable approach when handling a considerable amount of scripts, both receiving 0.

UnifiedViews provides an easy-to-use interface with options to schedule the pipeline to run daily, weekly or monthly in a specific hour, or periodically in a defined interval, e.g. every two hours. There is also an option to run the pipeline after another pipeline have finished the execution. Providing many methods to schedule the pipeline, achieving the *schedule* criteria, scoring +2.

As in ODN, you can use Pentaho CKAN to schedule the job in different intervals, choosing the periodicity to execute the job, e.g., day; week; month; or every specific interval in minutes or seconds. Making a scoring +2 for the *schedule* feature.

3.6.1.7 Log

Using a SQLite database, the WPRDC-ETL tool automatically stores information about the last execution of the pipeline, like title, last execution time, status and few more information. Analyzing the SQLite table which the log is stored, it is possible to determine that no history will be saved, only the last execution information, because the primary key of the record is the pipeline name. Not achieving a proper method of registering logs, needing improvements as informing who created the pipeline, and the complete history of the pipeline. Thus, scoring only +1.

Storing each step of the pipeline's execution in a text file, the CKAN-AUTOMATOR did not achieved the criteria of the log feature, being difficult to track the last running process of each pipeline, and almost impossible when executing many pipelines, because everything is stored in a single text file, scoring 0 in the *log* feature provided.

Through the UnifiedViews' GUI it is possible to see every step of the pipeline execution, containing information about events that happened while the pipeline was running, showing a time stamp and a short message for each event, and also a more detailed log with technical information, as the methods executed and the exceptions that happened. Thereby, scoring +2 in the *log* feature.

Pentaho CKAN produces a log during the execution of the pipeline, the logs can be stored in a specific file, or in a database table. The logs can be customized, with options as a basic log level, only errors, or a very detailed log, with technical information. Therefore, scoring +2 in the *log* feature.

3.6.1.8 Notify

There is no possibility to notify through e-mail in the WPRDC-ETL tool, therefore, not fulfilling the criteria, scoring 0. In the CKAN-AUTOMATOR it is possible to inform one e-mail, in the configuration file, to receive notifications about the last pipeline execution, but it is not possible to send it to many users, thus, not properly achieving the *notify* criteria, scoring 0 too.

The UnifiedView provides an interface to configure notifications for each pipeline, being possible to custom the notification, e.g. choose if a notification is going to be send every time a pipeline starts, or if should be an instant notification or a daily bulk report of the successful executions. And it is also possible to register many e-mails to receive the notifications, attempting to the criteria of the *notify* feature, scoring +2.

When creating a job in Pentaho CKAN, you can chose to send e-mails with the information of the last execution of the job, where you can customize a message, and even chose to send the log of the execution, but is necessary to configure the SMTP server to send the e-mail. Thus, achieving the criteria described in the taxonomy, also scoring +2.

3.6.1.9 Data Dictionary

None of the tools analyzed were able to help creating or to create a Data Dictionary of the Resource, therefore, the score 0 was attributed to each tool.

3.6.2 Quality Attributes Group

3.6.2.1 Focus on the Business Technician

By the previous analysis it is clear that the WPRDC-ETL and the CKAN-AUTOMATOR are not tools focused in the business technician. The necessity to code, and a high level of technical knowledge is necessary to one be able to publish Open Data in to the CKAN platform, not reaching the criteria, thus, scoring 0.

Although the UnifiedView provides an interface, as said before, it is focused in the business technician. The analysis made possible to arrive at the conclusion that the tool is not completely to use of the business technician. The nomenclatures used all over the tool is, in most of the times, necessary of technical knowledge to understand what is being done.

When creating a pipeline in UnifiedView, the labels used to express the functionalities of each option in extracting, transforming and loading data are all terminologies of the computer field of study, e.g. meta data; HTTP request; SQL; SPARQL; RDF, being of little understanding of the ordinary business technician.

To extract the data in UnifiedView it is also necessary knowledge in the database science, where the user has to create a query, using the SQL language, to extract information from a remote database, for example. Thus, not achieving the criteria in a complete sense. Thereby, scoring only +1.

Pentaho CKAN is known to be a tool of use of ETL engineers, abstraction features in a GUI, but still needing the knowledge in many aspects related to the data and database science, e.g., creating SQL queries to extract the data. Thus, scoring +1.

3.6.2.2 Object-Relational Mapping

The tools compared in this study does not have any ORM functionalities. In the WPRDC-ETL and the CKAN-AUTOMATOR the data is extract directly from a CSV file. In the UnifiedView and Pentaho CKAN, the options are the ones cited in the paragraph of the extraction feature. Therefore, none of the tools achieved the *Object-Relational Mapping* criteria, all of them scored 0.

3.6.2.3 Data Quality Evaluation

None of the tools have any type of feature to provide evaluation about the data published, not achieving the criteria of the *Data Quality Evaluation*, scoring 0.

3.7 Conclusion

All of the aforementioned tools did not enable non-IT technicians to publish the data produced by themselves. The tools presented had features to make the process easier, but not enough abstractions to allow the non-IT technicians to use it. Both the WPRDC-ETL and CKAN-AUTOMATOR required higher need of IT knowledge, being necessary to code to automate the process of publishing a resource. Despite the higher score from ODN and Pentaho CKAN, and the GUI provided, both of the tools were not able to allow the non-IT technicians to publish Open Data. Moreover, both tools requires IT knowledge in the process of extracting the data, being necessary to provide the database connection information, and also to create a SQL query to extract the wanted data. It was also noticeable the vast use of IT technical words at the GUI, making harder to non-IT technician use the tools. Thus, the IT knowledge required to properly use the tools to publish the data hampers the Open Data movement.

4 OpenEasier Design Decisions

This chapter details the tool proposed in this dissertation, which was named OpenEasier [1]. We present the main design decisions undertake to define the tool's architecture and implementation. We used the taxonomy created and the patterns encountered by the literature review as a baseline of the OpenEasier requirements. Understanding the Brazilian's Open Data context was also a fundamental part to elaborate the system's requirements. In the Design step we used principles of Software Engineering to conceive and describe the proposed artefact. At this chapter we leverage Sommerville (2010) and Ambers (2005) works as foundations to the process of elicitation, design and documentation of the requirements of the proposed solution.

Although the smarter path to choose would be to use the best open-source tool analyzed in chapter 3 – which in this case would be the Open Data Node – there were a few complications that did not make it possible. First, we were not able to compile the project from the source code. Second, the project has been developed for more than four years by a team of more than 10 developers. This by itself, and the lack of a good documentation, made it difficult to understand the project structure in the time we had to implement all the new features necessary, and make the changes, without compromising the software tool native features. Thereby, due to those reasons, we choose to create a new tool, which will be proposed in the next sections.

In this chapter, we will present in the section 4.1 an overview of OpenEasier, discussing some technical aspects of the architecture in a high level, and presenting the benefits of the organization of how the system works. In section 4.2 we have a more detailed examination of the architecture of OpenEasier, providing a better understand on how the internal components are organized and how the components communicate with each other. The section 4.3 will provide a brief discussion explaining the requirements, and also an use case digram to illustrate the interactions the users can have with the tool.

4.1 OpenEasier overview

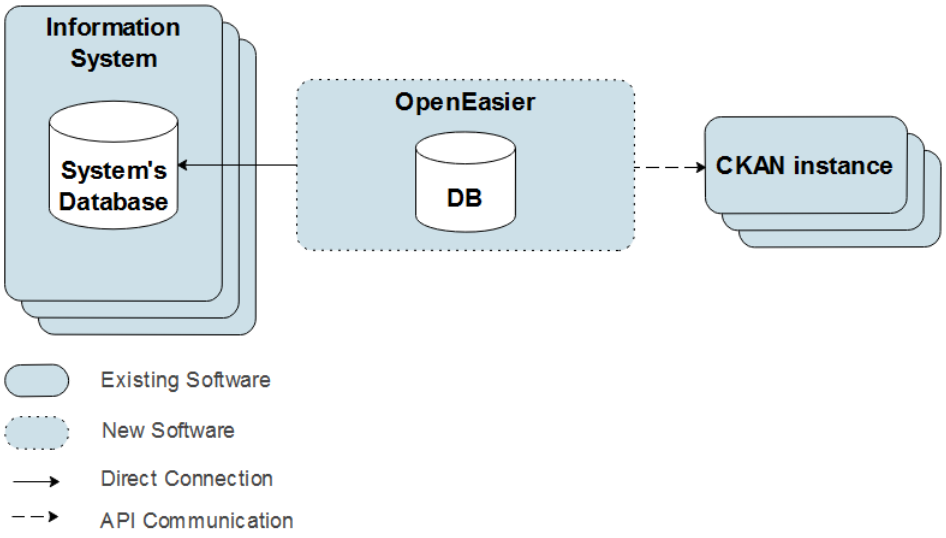
OpenEasier is an open-source Web-based software tool, which has features to enable non-IT technicians to easily publish data in an open pattern, according with the Brazilian's OGD context. The process of publishing and maintaining Open Data is broken in simple steps, those steps are abstractions of the IT technical process to publish Open Data. This will enable non-IT technicians to manage the process of Open Data, with little dependency

[1] https://github.com/Jonas452/openeasier

on IT technicians.

The Figure 3 presents a high level overview of how the solution is organized. The Information System and the CKAN instance displayed in the Figure 3 must be previously configured by the IT Technician, by informing the connections configuration in the OpenEasier. This will enable the non-IT Technician to choose the data to be opened from the configured system, load it to the CKAN instance configured, and keep track of the resources' updates executed automatically by OpenEasier.

Figure 3 – OpenEasier integration overview



Font: Authors

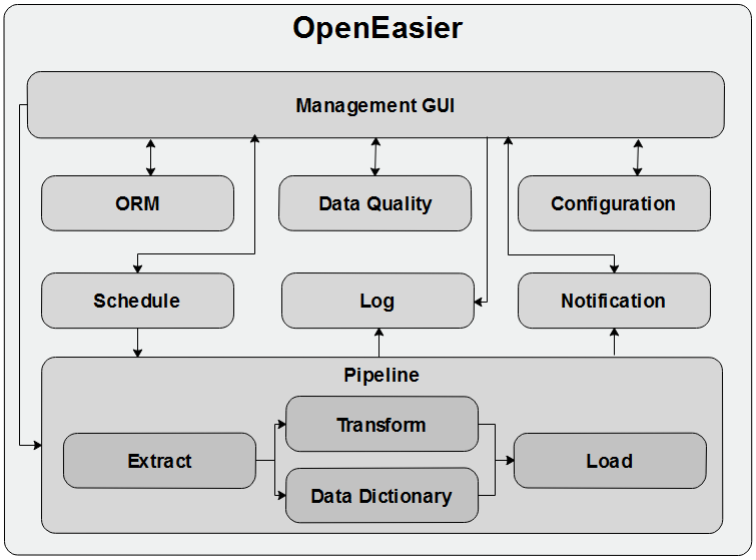
By designing OpenEasier to function as a separated software from the data source (Information System) and from the CKAN instance, we can have three main benefits with this approach presented in Figure 3. The first is performance since OpenEasier natively is not coupled to none of those other softwares can improve performance. By not affecting the other system and working as a separated system, OpenEasier is free to execute the necessary tasks in a dedicated server, if required, not overloading any of the other systems. Another important factor to be considered is scalability, since OpenEasier is connected to the other system only by configuration, it is easy to have connections between many Information Systems, or even different CKAN instances, if desired. The last benefit we can have with this approach is security. Since OpenEasier natively is not coupled to the CKAN instance — as it is in the ODN approach — we do not need to have big concerns with security. Since CKAN must be in a public domain, accessible to external sources, this can lead to security problems. By decoupling OpenEasier from CKAN, those problems are less difficult to solve and can be outcome from a Network Firewall security, for example. Thus, having the possibility to let OpenEasier decoupled from any of this softwares is a

good choice, gaining in the aspects discussed in this paragraph.

4.2 Architecture

As a mean to provide a broader understand of how the system works and how it is organized, we will discuss in this section the architecture established to proper achieve the requirements, explaining in the next paragraphs each component contained in the OpenEasier tool, and also the interactions between those components. The Figure 4 shows the Architecture Diagram in discussion. It is important to highlight that those components will all be deployed in a Web Server, working in a client-server model. The *Pipeline* must have a interface to interact with the *Management GUI* component. However, the *Pipeline* must run as an independent application, working in the server-side, being assign the task to execute the ETL processes. Besides that, all the other components will work as a client-server application, reacting to the clients requests in the back end.

Figure 4 – Architecture Diagram



Font: Authors

The *Management GUI* is a component which the main intention is to provide a Graphical User Interface (GUI), allowing the user to interact with the existing OpenEasier's features. As said before, the GUI is an important component of the tool, it must have abstractions of the ETL functionalities, being of easy manipulation. The *Management GUI* component interacts with all the existing components, only to present the results, status or interactions of each feature.

Another main component is the *Pipeline*, containing the *Extract*, *Transform*, *Data Dictionary* and the *Load* components. The *Pipeline* will connect and control the other

components, allowing the user and other components to interact with it and receive the results in summarized way. First, the *Pipeline* will start by calling the *Extract* component to retrieve the data from a database, than the *Extract* component will send the data retrieved to the *Transform* and *Data Dictionary*. The transformation will be executed in the *Transform* component, excluding and formating the necessary data, and by the end generating the final CSV file. Running concurrent with the transformation, the *Data Dictionary* component will read basic informations as columns name, type, size and if the column can be null, and with that information, it will create a PDF file containing those metadata. As the final step, the *Load* component will send the CSV file resulted from the transformation to the DataStore CKAN API, and the PDF file outputted from the *Data Dictionary* component to the FileStore CKAN API.

The *Schedule*, *Log* and the *Notification* components are all components that have interactions with the *Pipeline*. The *Schedule* component will enable the user to schedule the pipeline execution in determined periods by using the GUI. The *Pipeline* will use the *Log* component to register the results of each step of the pipeline, making a history of the execution, being of great importance to track errors in the process. And the final component to interact with the *Pipeline* is the *Notification*, which will be used to send the results from the pipeline's execution to the users that were previously registered to receive it.

The *Object-Relational Mapping (ORM)* component has the task to provide an abstraction of the database's tables to the user, so the user can decide which table and columns will be published. Through the GUI, the Business Technician will select the table to publish the data, and than select each column to be published. This will allow the Business Technician to proper choose the data being published. As a mean to improve the quality of Open Data, the *Data Quality* component will provide a special feedback to the user about the data being published, specifying disturbances in the data.

The last to be discussed is the *Configuration* component. As the name indicates, the configuration of the project, and the permissions will all be contained in this component. This is where, for example, the configuration of the users, and each the permission to access the schemes in the database will be configured. The IT Technician is the only user of this component, allowing to proper prepare the tool so the Business Technician can be able to use it. By that, we are making the job of the Business Technician easier.

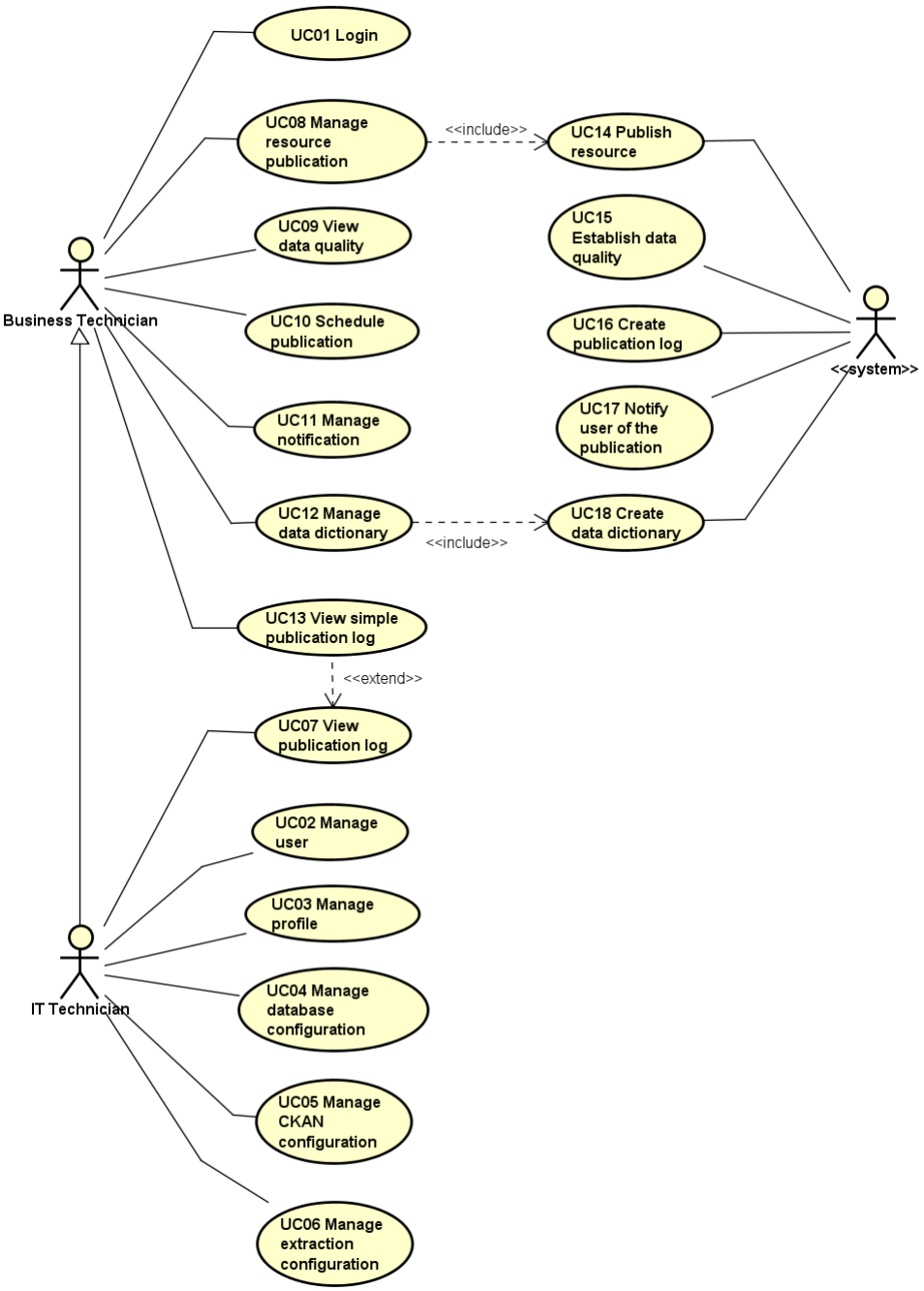
4.3 User requirements

This section will present the description of the user requirements and the diagrams outputted from the literature review and analyses performed at the chapter 3, it is important to highlight that the review and analyses executed was of great importance in

the process of elicitation of those requirements.

Since the main user of OpenEasier is the Business Technician, an usability quality attribute as *Focus on the Business Technician*, presented in the section 3.5, is highly important to the success of the solution. Thus, we had in mind this important attribute while elaborating those requirements, being careful to not add to much complexity to the process. Figure 5 presents an Use Case diagram, displaying the actors and their interaction with the system. In the essence, each Use Case that the Business Technician can interact must look as simple create, read, update and delete (CRUD) actions divided in straightforward steps. When summing each of those steps, the Business Technician will be able to publish and to manage the data produced by themselves at the end of the process.

Figure 5 – Use Case Diagram



Font: Authors

5 OpenEasier Implementation

In this chapter we will discuss about the development process of OpenEasier, explaining the technical decisions and also the achievements, and why the technologies, such as programming language and framework, were chosen. The main focus is to provide an understand of the reasons and decisions of the technical aspects in the tool implementation. The section 5.1 describes each main technology chosen and the reasons of such choices. And to demonstrate the implementation results, the section 5.2 present the OpenEasier tool in detail.

5.1 Technologies

Here we discuss all the main technologies used to develop OpenEasier, giving a brief description of each technology, and presenting the reasons of the choices.

5.1.1 Python

Python is a high-level programming language with multi-paradigms, and it is a language of general purpose, with the focus in simplicity and high performance. By being a programming language of general purpose we can use it in the Web and also to run in the server side as a script language. As explained before, components of OpenEasier must run in the Web environment, and other components must run only in the server side. Thus, this is one of the things we can benefit from the language. This characteristics made Python a good choice for the development of the tool.

5.1.2 Django

Django[1] is a Web Framework written in Python with the main purpose to be an easy to use framework, having many encapsulated features for the Web, such as user authentication and security concerns. Django is a mature open-source framework, it has being developed since 2005, and also has a big community supporting it. With all those facilities, and following the best patterns for the development of Websites, Django was a good choice to use it to develop the solution for the Web environment.

[1] https://www.djangoproject.com/

5.1.3 PostgreSQL

PostgreSQL[2] is a free and open-source Object-relational Database Management System (ORDBMS), it is a powerful and mature solution to handle database operations. The choice of PostgreSQL was because of the authors expertises, and also because it has the necessary features that OpenEasier will need from a DBMS also with a good community and documentation, along with no costs to use it.

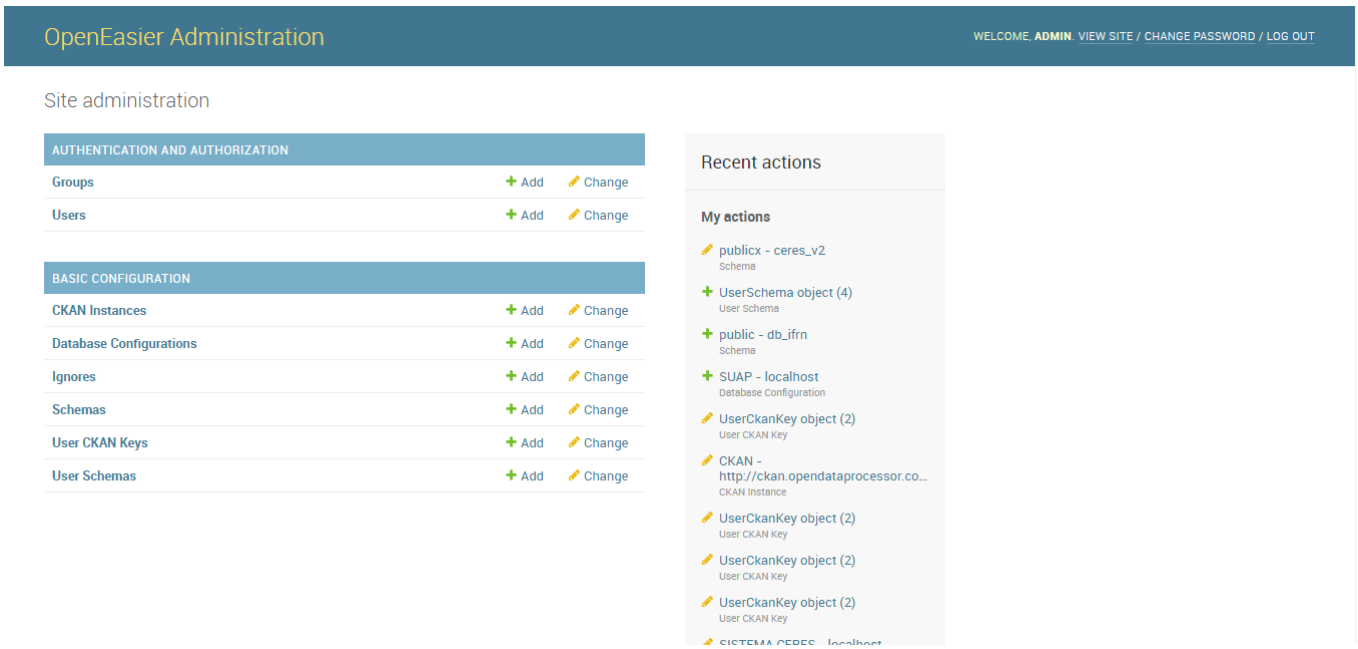
5.2 The tool

In this section we present OpenEasier, showing the steps necessary to publish and manage Open Data through the tool. We explain each step and present images of the screens.

5.2.1 Administration Area

The Administration Area has essential features to make OpenEasier work. This area is intended to be used only by the IT Technicians, where they can configure the tool so that the Business Technicians can use OpenEasier. At this area, the IT Technicians can configure the CKAN Instance to have the data published, the databases to have the data extract, and other things such as the users and CKAN credentials for the users. The Figure 6 shows the Administration Area discussed in this subsection.

Figure 6 – Administration Page



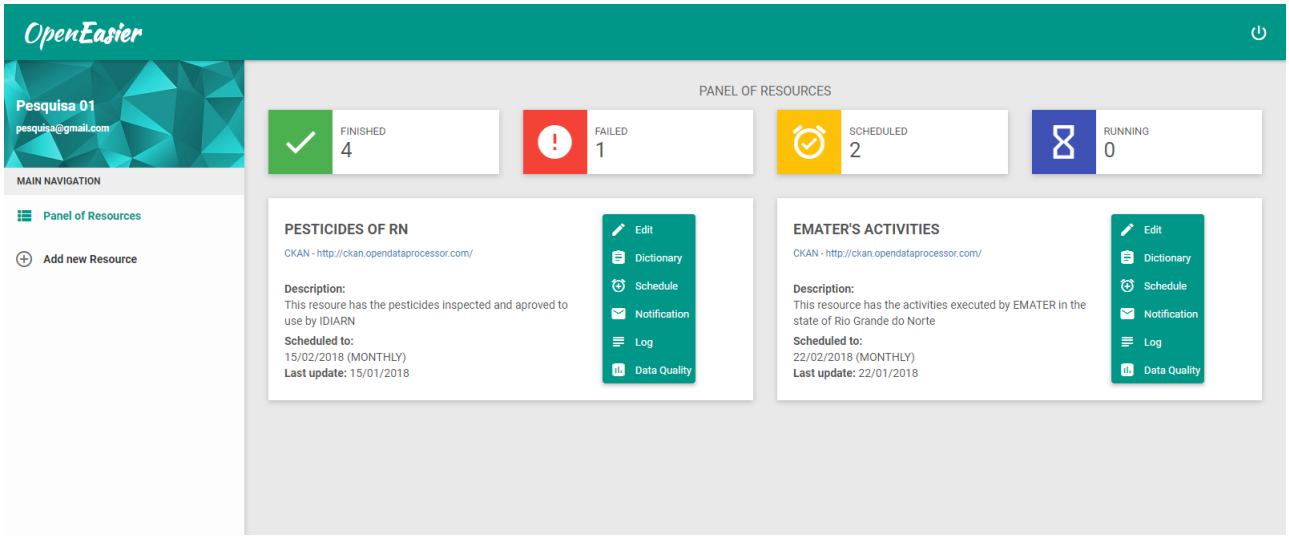
Font: Authors

[2] https://www.postgresql.org

5.2.2 Resources Panel

As a way to make possible the Business Technicians to manage the Open Data being published, and to follow the latests results from the resource publication, the Resource Panel presented in Figure 7 shows all the results of the publication, and also a detail of each resource created. The panel is the main page of the tool, where the user can interact with each resource. At the right corner of the resources' box, from up to down, the user can use the buttons to do the following actions: (i) edit the resource; (ii) create or edit the resource's data dictionary; (iii) schedule the resource to be published; (iv) register the e-mails that will receive notifications about the resource execution status; (v) view the log from the last execution; (vi) view the data quality feedback of the resource; and (vii) open the resource link of the CKAN instance. Each of those actions will be explained in the next subsections.

Figure 7 – Resource Panel Page

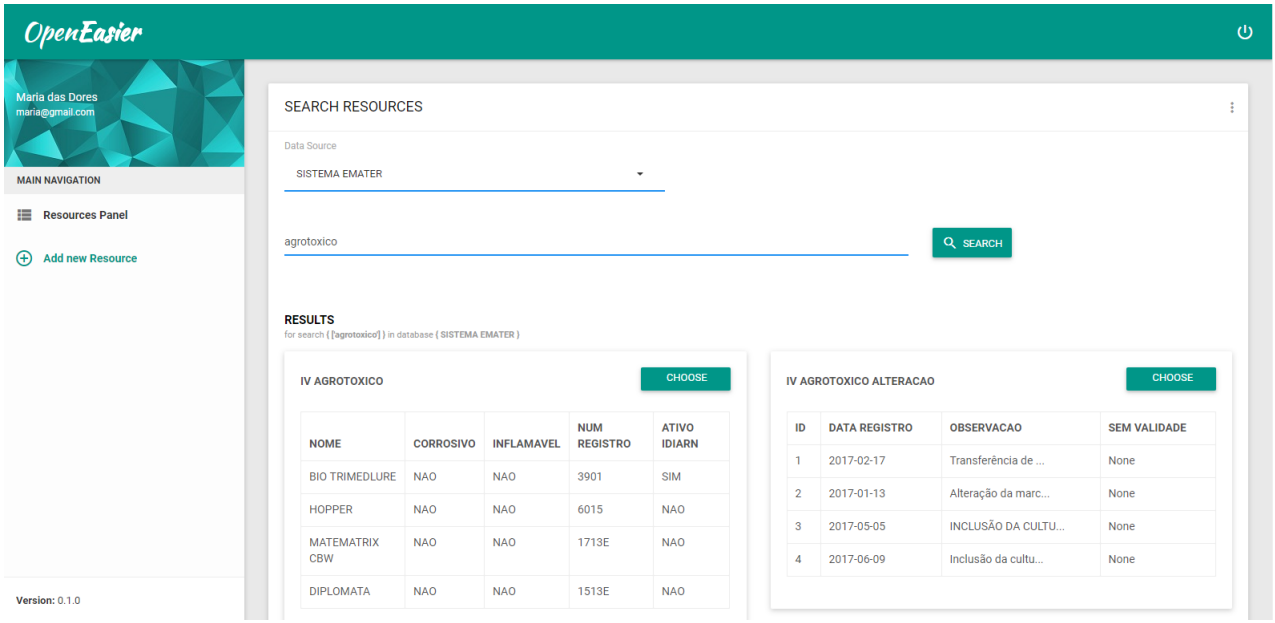


Font: Authors

5.2.3 Publishing Resource

The publication of a Resource is divided in four simple steps: (i) search and choose the data (a database table); (ii) select the wanted columns from the main table to be published; (iii) select the columns from the secondary tables (foreign keys); and last, (iv) fill the basic information of the resource to be presented in the CKAN instance. The Figure 8 presents the Web page of the first step, where the user will be able to search and choose the wanted table. To be able to do it, the user must select a data source (a database) to search for the data, this data source was previously attributed to the user by the IT Technician. Then, the user must type the keys words in the search box, where a result with the tables and a sample of the data in it will be displayed in boxes, and by clicking in the *choose button*, the user will be redirect to the page displayed in Figure 9.

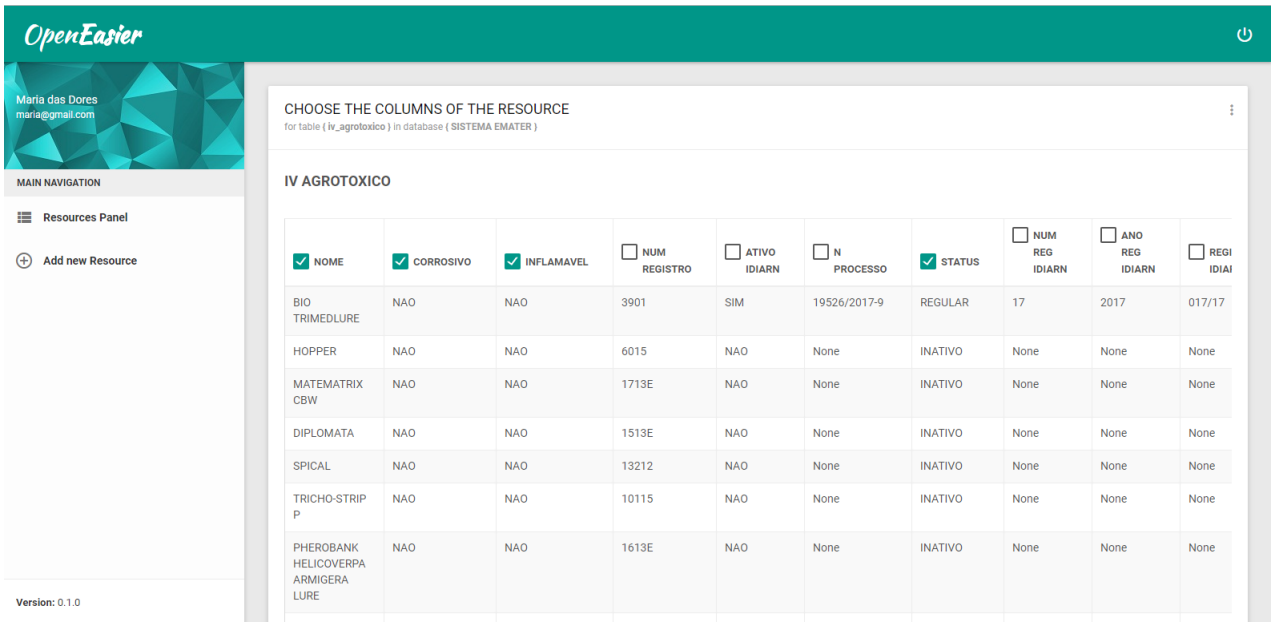
Figure 8 – Search for data source Page



Font: Authors

The Figure 9 shows the page with the second step, where the full information of the table will be displayed. The user must verify each column, and the data contained in it, to check (by recognition) if it is the wanted data to be published. Then, the user will select the wanted columns to be published and click at the continue button. Since in some situations not all the columns of the table will be published, this is an important step in the process of publishing Open Data.

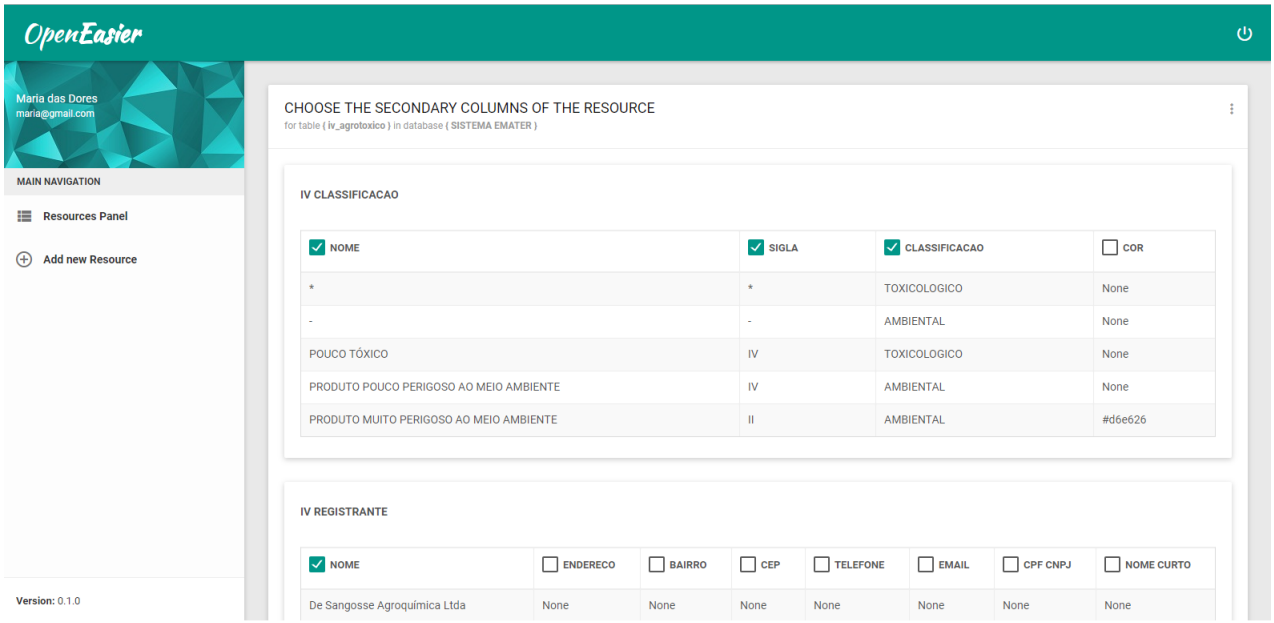
Figure 9 – Select the columns Page



Font: Authors

In the Figure 10 it is showed the third step. In this step the user is presented with the tables that are foreign keys of the table selected in the previous step. This allows the user to select columns to enrich the data which will be published, representing the reality of the data that is stored at the database.

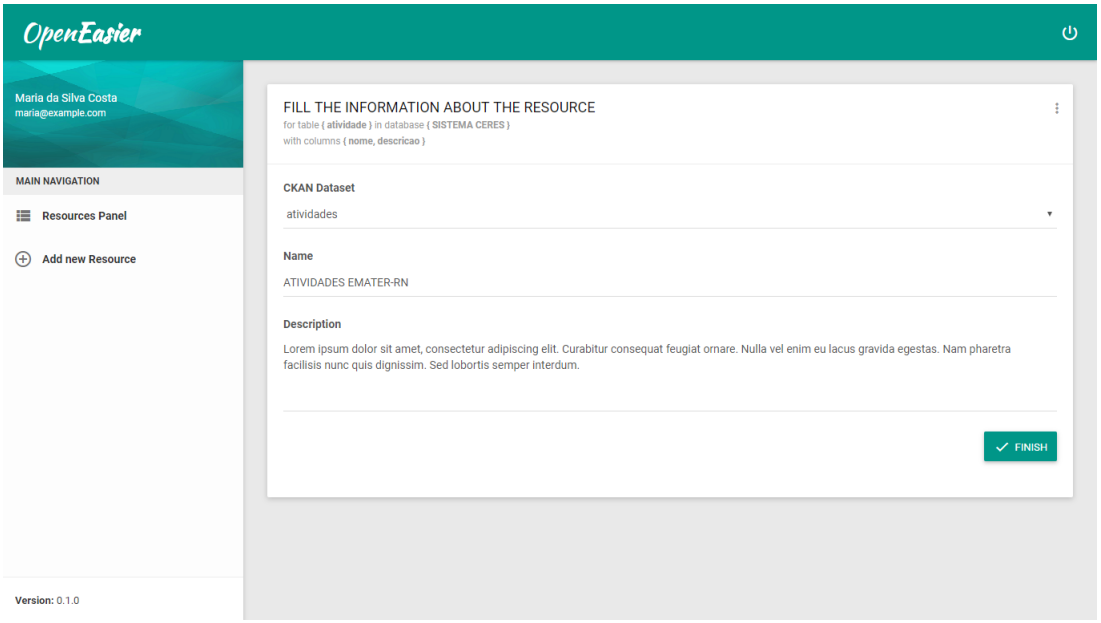
Figure 10 – Select the secondary columns Page



Font: Authors

The Figure 11 displays the last step to achieve the creation of the resource. First, the user must select the existing Dataset in the CKAN instance. The CKAN instance must be previously configured and set to the user profile, this must be done by the IT Technician. Then, the user must provide a descriptive name of the resource, and a description. At the end, by click in the *finish button*, the user will end the process of creating the resource, and must continue to the steps of scheduling it (subsection 5.2.4), and also to create the Data Dictionary (subsection 5.2.5).

Figure 11 – Page to describe the resource

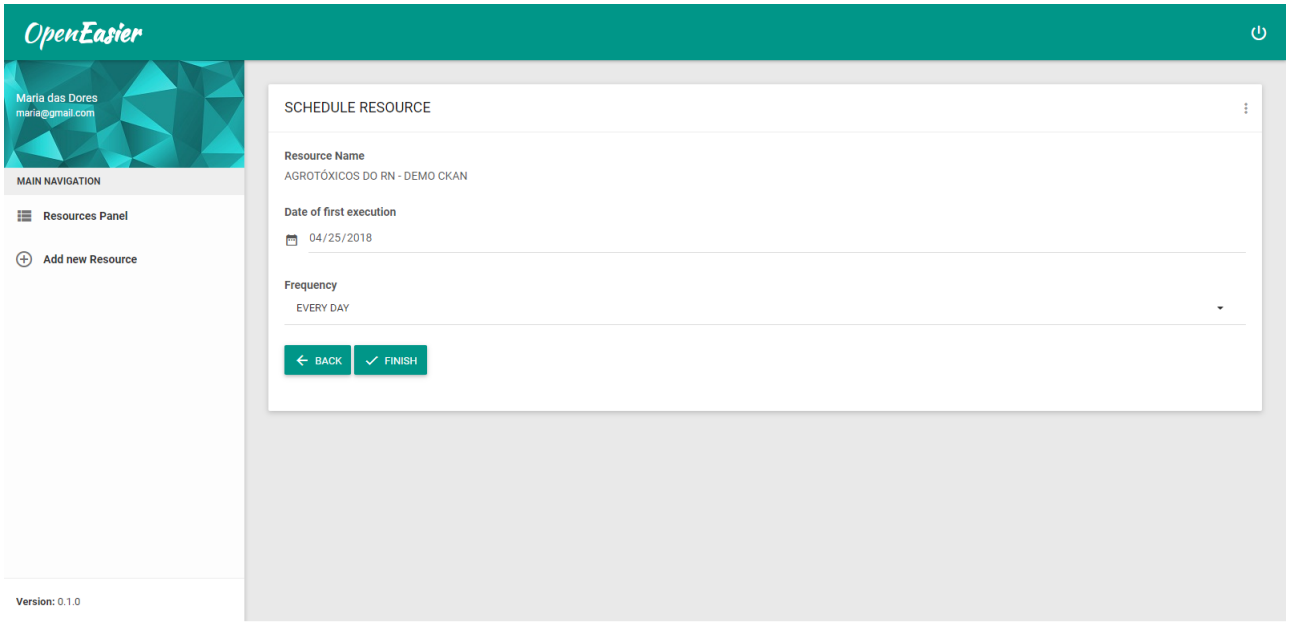


Font: Authors

5.2.4 Resource Scheduling

The Resource scheduling is a simple operation, which is presented in the Figure 12, the user must select the resource wanted to be schedule at the resources panel presented previously. The user must choose the first date to start the resource publication, and also the frequency of the resource's updates, where it is possible to choose 4 options: (i) every day; (ii) every week; (iii) every month; and last, (iv) every year. This simple operation will guarantee that the resource's data will be up to date, where it will be extracted directly from the source, with the periodicity desired.

Figure 12 – Page to schedule the resource

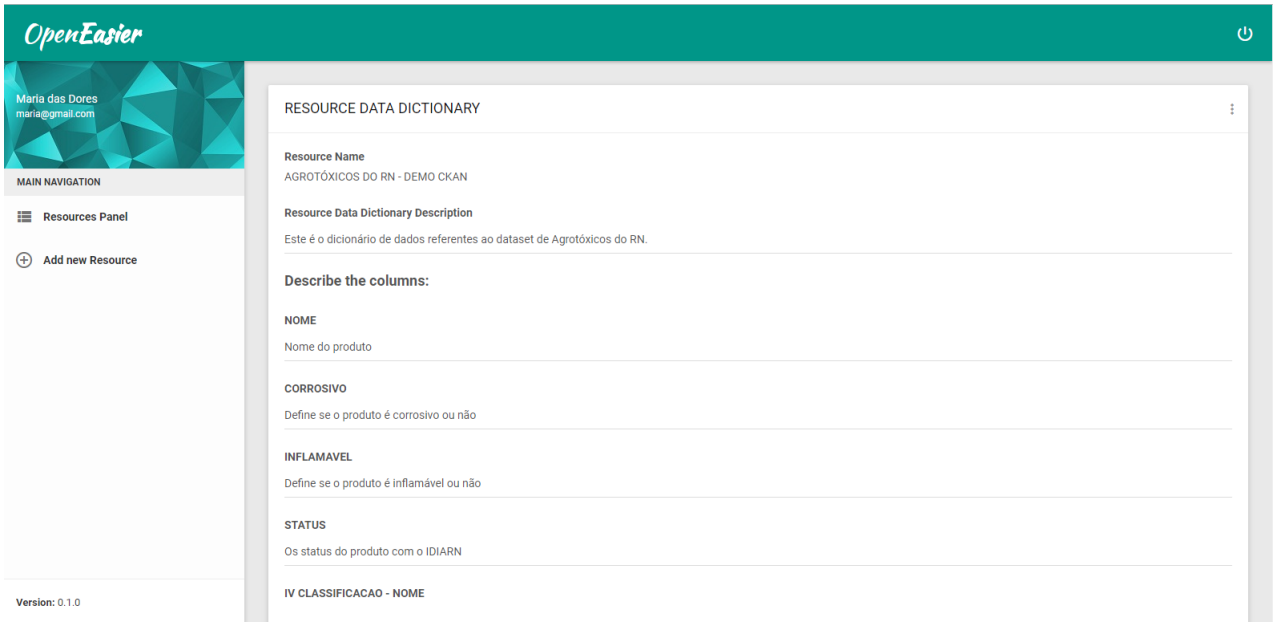


Font: Authors

5.2.5 Data Dictionary

The Figure 13 presents the page where the user will be able to create the data dictionary. For now, it is only necessary to describe each column of the resource, providing information to who will use the data. This will enable the process of automatic creation of the data dictionary, not being a concern of the Business Technician to keep it in a more manual way.

Figure 13 – Page to describe the Resource's Data Dictionary



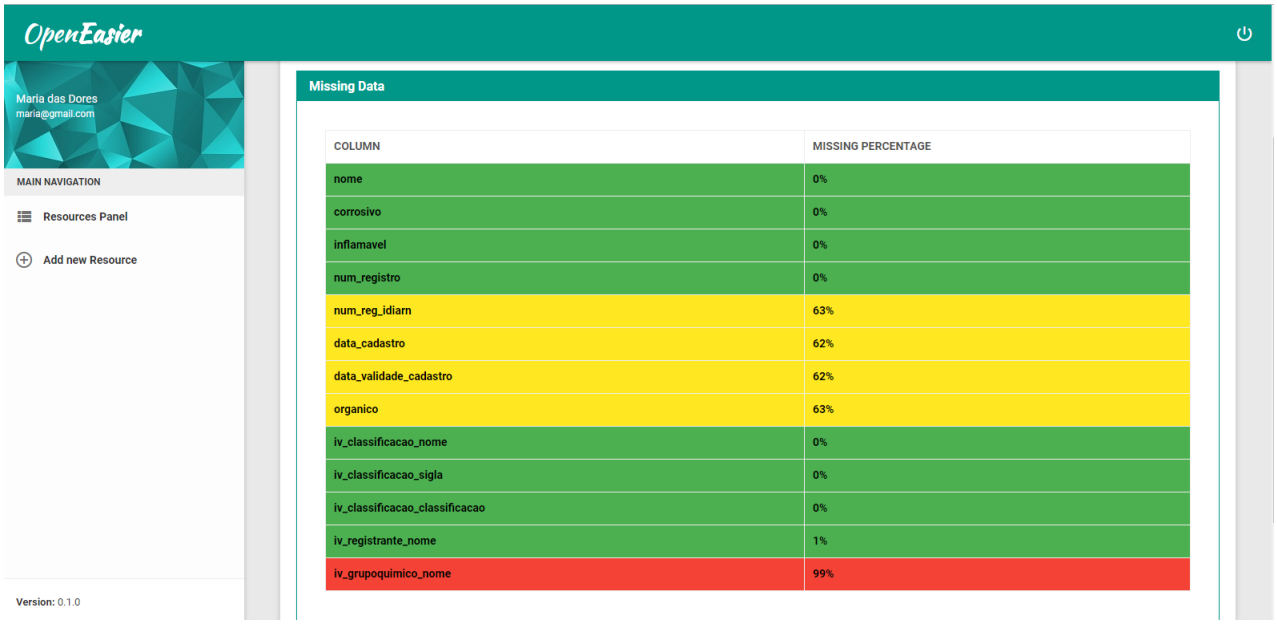
Font: Authors

5.2.6 Data Quality Evaluation

Figure 14 shows the Data Quality Evaluation feature. In this page the user can check some aspects of the data published. The user has two options: the (i) Missing Data evaluation, where it is shown the percentage of missing data for each column; and the (ii) Format Consistency evaluation, which shows the data types percentage of each column, providing feedback to the user if a column has different percentage of data types. Those data types vary between text, numbers and dates. This is still a simple evaluation of the data, and must be enhanced. This was developed with the use of Pandas [3], an open source Python library for data analysis.

[3] https://pandas.pydata.org/

Figure 14 – Data Quality Evaluation



Font: Authors

6 OpenEasier Evaluation

As a mean to achieve the validation and evaluation of the artefact, following the steps of the DSRM, we executed an evaluation of OpenEasier. This evaluation had the focus on usability, since it is a crucial factor for the success of the proposal of this work. Therefore, in this chapter we will detail the approach we adopted to evaluate the tool, and also present the results we obtained.

6.1 Institutions

In this section we will present each institution that participated of the evaluation. The selected institutions can be divided in two categories of domain: (i) agriculture, being the domain in which SAPE, EMATER-RN and IDIARN focus on; and (ii) education, being covered by the efforts of the IFRN.

6.1.1 SAPE

SAPE is a public entity which focuses on promoting improvements in qualitative and quantitative aspects of rural production at the state of Rio Grande do Norte (RN), Brazil. Fomenting the internal and external market. To achieve such improvements, there are sub entities coordinated by SAPE to elaborate and execute the Public Policies, allowing SAPE to attain a broader range of policies. The sub entities are the following:

- EMATER-RN;
- EMPARN;
- IDIARN;
- CEASA.

In this study we opted to choose to work only with EMATER-RN and IDIARN, because the other institutions do not have information system. Thus not being able to apply the evaluation, since is necessary to have an information system to extract the data from.

6.1.2 EMATER-RN

The EMATER-RN is a public institution that promotes the agribusiness in the state of Rio Grande do Norte, Brazil. The main focus of the institution is in the family

farm business. By applying public policies the EMATER-RN seeks to create a sustainable environment. Nowadays, 167 municipalities are assisted by EMATER-RN, reaching a total of 2,140 rural communities, aiding 58,742 rural families that live in those communities. It is possible to access the EMATER-RN works at the Ceres Cidadão [1], which is a Website that provides transparency to the society.

6.1.3 IDIARN

The IDIARN is a public institution that has the mission to defend and inspect the agricultural sector in the state of Rio Grande do Norte, Brazil. This work is executed with the intention to certify the quality of agricultural products, secure the public health and the environment, and increase competitive in the market in the RN. Thus, the institution can fine those who does not follow the standards, or perform activities that can cause harm to public health or to the environment.

6.1.4 IFRN

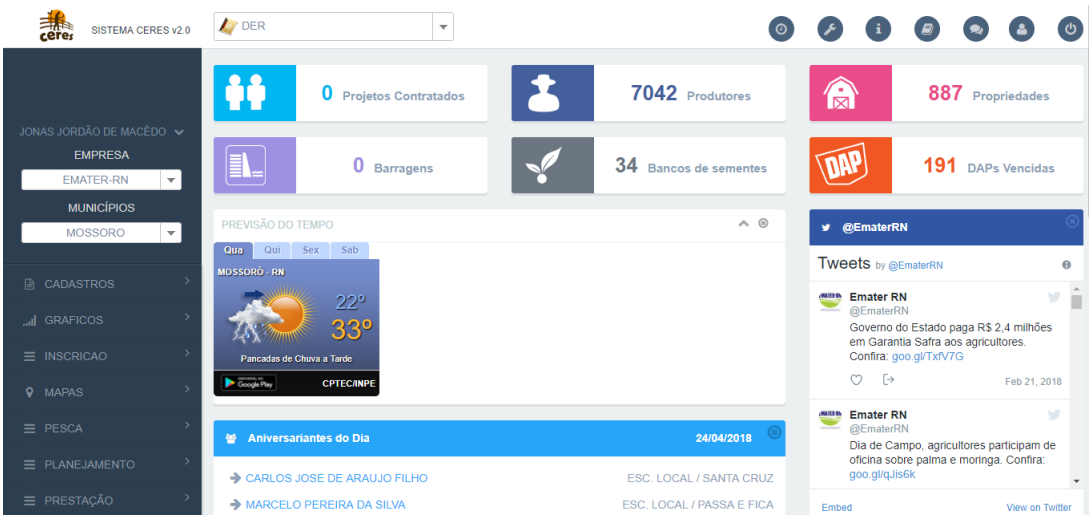
The Instituto Federal do Rio Grande do Norte (IFRN)[2] is a federal public educational institution with professional and technological courses. The institution has over 28 thousand students and 21 campus distributed in the RN, offering undergraduate degrees and postgraduate degrees. Nowadays the institution has 109 courses in many areas of expertises.

6.2 CERES

CERES is an information system adopted by SAPE, EMATER-RN, IDIARN and other public institutions, which provides features to monitor and to evaluate the public policies that each entity apply. The system is equipped with a great amount of reports and graphs to help with the activities of monitoring and evaluating the institution work, those reports and graphs are all built by data registered at the system by the users. The system has 19 modules nowadays. There are modules to manage the human resources of the institutions, the use of vehicle to transport the employees, the assets, and other modules specifics for each institution. Being used by a vast number of users with different profiles. The Figure 15 presents the home page of CERES. Therefore, this is one of the system which we chosen as the source of data to use in our evaluation of the OpenEasier tool, being a good case because of the variety of users and data.

[1] http://cerescidadao.rn.gov.br/
[2] http://portal.ifrn.edu.br/

Figure 15 – CERES 2.0

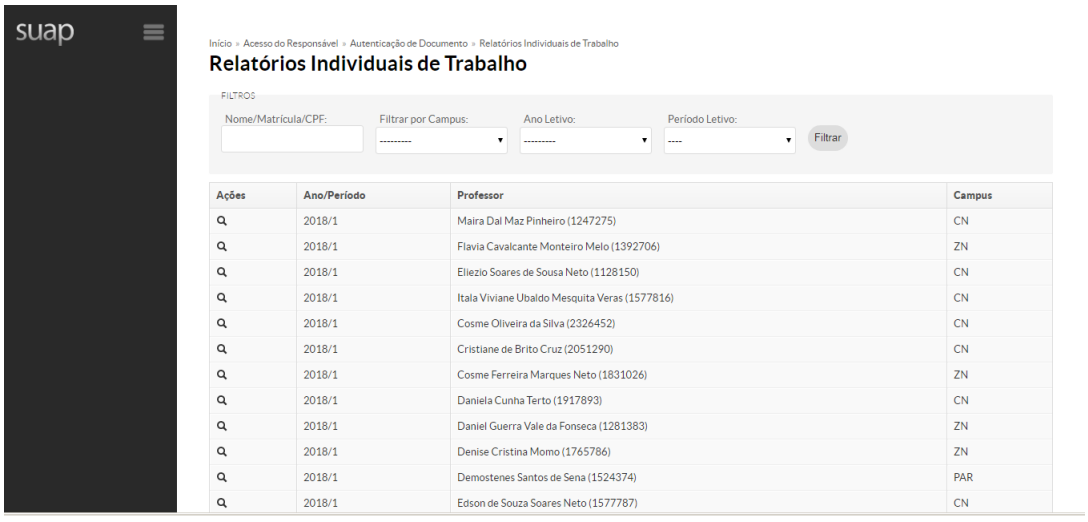


Font: Authors

6.3 SUAP

Sistema Unificado de Administração Pública (SUAP) is the system developed and used by the IFRN to manage the administrative processes of the institution. The users of the system are the institution's work force and also the students. The system has many modules, such as the modules to Manage Assets, Projects, Contracts and others. The professors and students also use the system for their academic activities. The SUAP was adopted by other federal educational institutions from Brazil, having a great acceptance.

Figure 16 – SUAP



Font: Authors

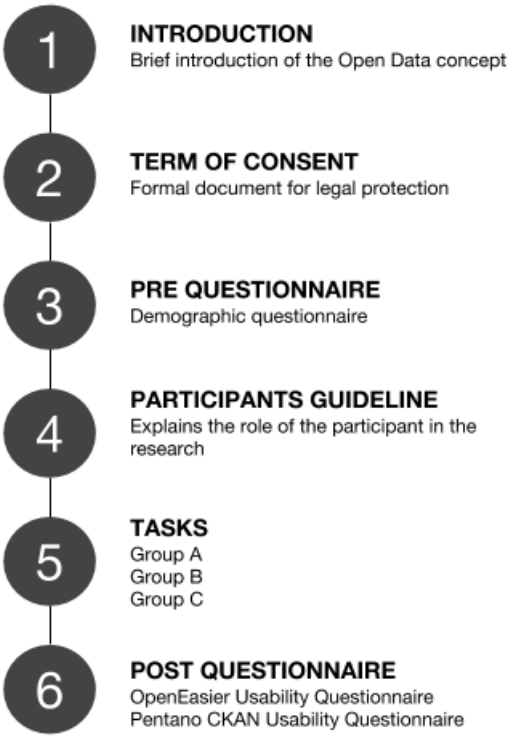
6.4 Evaluation Approach

The main aim of the evaluation was to assess the *usability* quality attribute. Even though software usability has slightly distinct definitions in the academia (ABRAN et al., 2003), it can be broadly defined by the ISO 9241-11:2018 Standard[3] as “the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". In an even more precise definition, focused in the field of Software Engineering, the ISO 25010:2011 [4] defines usability as the "degree to which a product or system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use". Therefore, with that said, this evaluation focus on assessing if the tool OpenEasier properly fulfill four aspects of usability: (i) effectiveness, which is a measure of the success rate of accomplishing the tasks (SEFFAH et al., 2006); (ii) efficiency, being related to time and human effort to execute the tasks; (iii) satisfaction, as the feelings that the user has while using the software (e.g. discomfort, positiveness, capable) (BEVAN; CARTER; HARKER, 2015); and (iv) "learnability", which can be measured in how easy is to the user identify and learn how to use the software (NIELSEN, 2003).

The evaluation approach chosen was a comparison test, where the goal is to compare two or more designs to establish which design is easier to use and learn (RUBIN; CHISNELL, 2008). Thus, we developed questionnaires and a series of tasks to perform this experiment, comparing OpenEasier with Pentaho CKAN (presented in the section 3.4). We opted to use Pentaho CKAN as the competitor for this experiment because it scored higher in our evaluation presented in section 3.6, and it is also a tool broadly adopted in the industry. The evaluation was planned and divided in six steps, which will be presented in the next paragraphs. The Figure 17 shows the steps executed in the experiment.

[3] https://www.iso.org/standard/63500.html
[4] https://www.iso.org/standard/35733.html

Figure 17 – Experimental procedure



Font: Authors

Introduction. A brief introduction to explain to the participants important concepts related to OD, such as the definition of OD and what is a dataset. This was relevant so the participants could be able to understand the tasks that they must execute. The intention is to put the participants in a real situation, where they have to publish the data of their institutions. Thus, it is important to explain such concepts.

Term of Consent. The Term of Consent (Appendix A) is a formal document describing the intentions of the research, clarifying the goals and assignments for the participants. And also explaining in which situations the data collected (e.g. video of computer screen and participant face, and audio) will be used, making clear that in under no circumstances the participants identity will be revealed in the final report of the evaluation. Thus, this document was created to legally protect the participants and researchers, being signed by both and given a signed copy to each one (participant and researcher).

Participants Guideline. This document (Appendix B) must be given to every participant so they can read and understand what is their role in the research. And to also make clear that the tools are being evaluated, not them. The document provides instructions on how to act during the research, underlining things such as speaking out loud during the execution of the tasks, and not asking questions to the researchers.

Tasks. The tasks which the participants executed were created in accordance with their context, and there were a total of 5 tasks, described in the Table 2. Each of the tasks has a narrative, serving has an introduction to the participant. The participants must execute each task 3 times. The first time the participant has no training at all regarding using the tools, only the task sheet to read and follow. The second time the participant will receive training in each tool and must execute the tasks using the task sheet. The third, and last time, the participant will have no task sheet to follow, they must execute the previous tasks with only what they remembered. This was elaborated with the intention to test if the tools were easy to use, and also to measure how easy it is to learn to use them. We also alternated the sequence of the use of the tools, where one participant started using OpenEasier and Pentaho CKAN, the next one used Pentaho CKAN and OpenEasier, and so on.

Table 2 – Evaluation tasks

Label	Tool	Description
T01-PC (Appendix C)	Pentaho CKAN	the participant must create a transformation to select and publish the data
T02-PC (Appendix D)	Pentaho CKAN	the participant must schedule the execution of the transformation
T01-OE (Appendix E)	OpenEasier	the participant must select and create the resource that must be published
T02-OE (Appendix F)	OpenEasier	the participant must describe each column selected (this creates the data dictionary)
T03-OE (Appendix G)	OpenEasier	the participant must schedule the publication of the data

Font: Authors

It is important to make clear that the task T01-PC and T01-OE are equivalent (select the data to be published), and both the task T02-PC and T03-OE are also equivalent (schedule the data publication) in the tools. The features to achieve the task T02-OE only existed in the OpenEasier, but we kept it in the experiment because it was a crucial step in the publication of the Open Data (the creation of the of the data of dictionary). Thus, the task T02-OE was not used to compare Pentaho CKAN and OpenEasier, but to validate the feature existent in OpenEasier.

Questionnaires. This paragraph details the third and sixth steps of the experiment. We developed three questionnaires for this evaluation, which are: (i) a demographic questionnaire (Appendix H), with questions about the profile of the participants, such as age, experience with computer tools, and understanding of Open Data; a (ii) questionnaire based on the 5-point Likert Scale with usability questions to evaluate the OpenEasier tool (Appendix I); and (iii) a questionnaire also based on the 5-point Likert Scale, with the same usability questions, but focused on the Pentaho CKAN tool (Appendix J). The

demographic questionnaire was answered at the beginning of the experiment, and the others two questionnaires was answered at the end of the experiment, after the participants execute each task 3 times.

As for the experiment environment, we used a conference room. And the participants used the researchers' laptop to execute the tasks and answer the questionnaires. The laptop camera was used to record the participants reactions (face), and the microphone was used to record the participants voice. We also recorded the desktop screen of the laptop during the execution of the experiments. Every participant had the same configured environment for each tool, e.g. database configuration and CKAN instance configuration.

The evaluation was executed seeking to answer four usability research questions, which were elaborated based on the four aspects of usability presented early in this section. The questions are:

- **RUQ-01** - Does the effort to use OpenEasier is greater or lesser than to use Pentaho CKAN?
- **RUQ-02** - Is using OpenEasier faster or slower than using Pentaho CKAN to select and schedule data to be published?
- **RUQ-03** - How easy is to use OpenEasier compared to Pentaho CKAN without training?
- **RUQ-04** - Does the user feel more confident when using OpenEasier rather than Pentaho CKAN?

Each research usability question was developed to measure and compare one aspect of the tools' usability. The **RUQ-01** was created with the intention to measure effectiveness. The **RUQ-02** is related to efficiency. **RUQ-03** was planned to measure "learnability". And last, the **RUQ-04** was developed to measure the aspect of satisfaction.

6.5 Participants Profile

Since the present study is focused on the idea that the business technician is capable of publishing OD — if enough abstraction is provided in a tool —, the target audience (participants) of the evaluation were selected according with their context of work. By context we mean the participants working area, e.g. Human Resource (HR), Accounting, Finance and other departments. We analyzed the data from each institution to elaborate the tasks for the participants, then we selected the participants for each context created.

The Table 3 presents the demographic profile of the participants. A total of 5 participants participated of the experiment, being a number considered ideal to exposing

80 percent of the usability problems (RUBIN; CHISNELL, 2008), and in our situation of limited time and resources that was an acceptable amount of participants.

Table 3 – Participants Profile

CHARACTERISTICS	NUMBER OF PARTICIPANTS
Age	
21-30	2
31-40	3
41-50	0
51+	0
Gender	
female	3
male	2
Educational Level	
high school	0
bachelor's degree	3
master's degree	2
doctoral degree	0
Use of Computer Tools	
does not use	0
uses occasionally	0
uses frequently	5
Institution	
EMATER-RN	2
IDIARN	1
IFRN	2
Knew about Open Data	
yes	0
no	5
Used or Published Open Data	
yes	0
no	5
Total of Participants	**5**

Font: Authors

6.6 Results

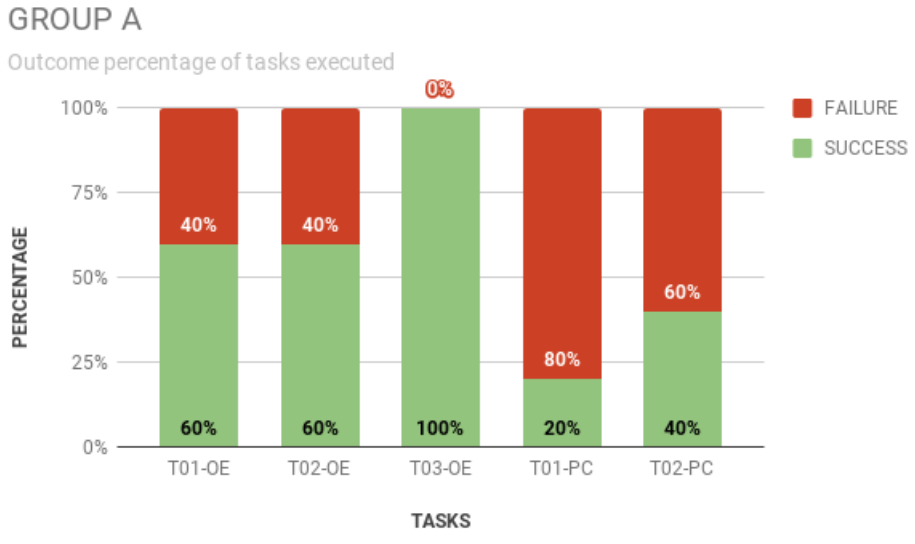
The execution of the experiment provided data from video and responses from the questionnaires. The raw data collected was summarized and analyzed, enabling us to extract the results of the experiment. This activity provided us the base to determinate the outcome of the experiment, answering the questions presented in the section 6.4. In this section we present the results from the experiment.

For the tasks executed by the participants, we analyzed the videos and extracted the

time to complete each task, and also classified each execution with two possible outcomes: (i) success, which means that the participant concluded the task and also did everything correctly; and (ii) failure, meaning that the participant was not able to complete the task, and/or did something wrong. In addition, we collected data from the comments and reactions of the participants during the execution of the tasks. The tasks mentioned in this paragraph are the ones presented in the Table 2.

As explained before, the participants executed each task 3 times, thereby, for this analyzes we separated the execution of the tasks in three groups: (i) Group A in which the participant first executed the tasks only with the task sheets, without training; (ii) Group B, where the participant was trained and also had the task sheets to follow; and last, (iii) Group C, with training and without the task sheets to follow, having to relay in what they learned. We have created outcome charts with success and failure percentage for each task, aggregated by groups, and time execution charts with the minimum, average and maximum execution time in seconds for each task. The charts presented in the next paragraphs of this subsection were created with the summarization of the raw data from all five participants of the experiment. To clarify, the label of the tasks have a prefix for each tool, where for OpenEasier the prefix is OE, e.g. T01-OE; and for Pentaho CKAN the prefix is PC, e.g. T01-C.

Figure 18 – Chart Outcome Group A



Font: Authors

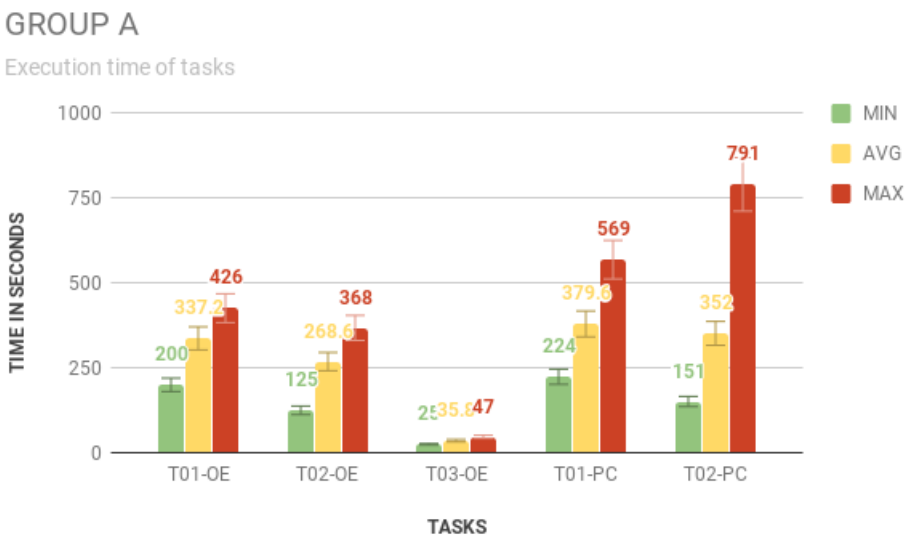
Group A. In the Group A the participants had no training and only the task sheet to follow. As the Figure 18 shows, the participants executed the task T01-OE with 60% of success outcome, this is 3 times more than T01-PC, with only 20% of success outcome. This demonstrate that is more simple to select the wanted data to be published using

OpenEasier than using Pentaho CKAN, without training in both tools. Comparing the task T03-OE with the T02-PC we have a difference of 2.5 times in favor of the T03-OE, where the T03-OE had a 100% of success outcome, and T02-PC had 40%. Therefore, establishing that is more straightforward to schedule the publication of the data using OpenEasier. The task T02-OE had a success rate of 60%, thus, being a good result for the OpenEasier tool, evaluating the feature of data dictionary.

In the Figure 19 it is possible to compare the time that the participants took to execute each task, having a minimum, average and maximum execution time. Comparing the T01-OE with the T01-PC the difference of the average execution time was of 42.4 seconds between the tasks. This was not a significant difference. In contrast, in the task T03-OE the participants performed much better than in the task T02-PC, where in the T02-PC the participant took approximately 960% more time to execute the task than in T03-OE, being a significant result.

As for the analyzes of the video, in both the tools the participants felt lost in the first time they were using it. But some of them were capable of searching, recognizing and selecting the correct table while using OpenEasier, but in the final step had difficulty to describe the resource being created.

Figure 19 – Chart Execution Time Group A

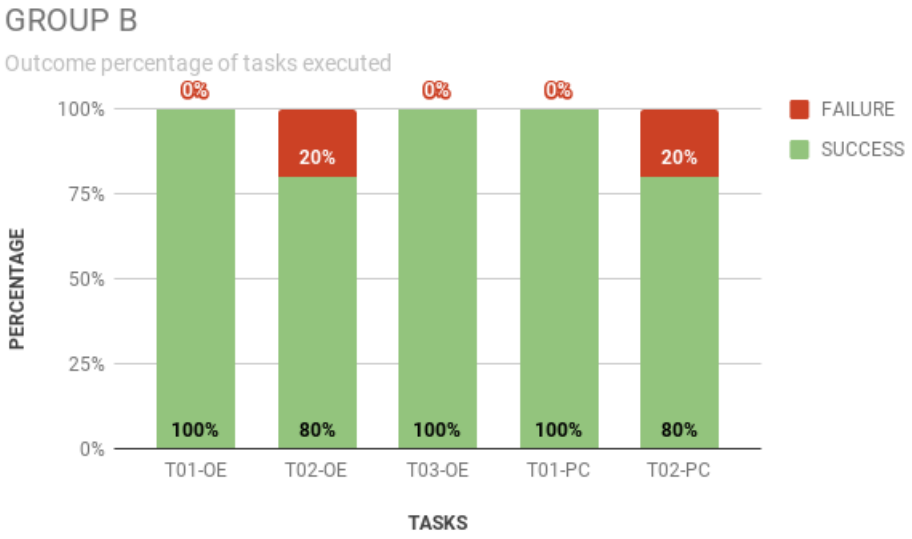


Font: Authors

Group B. In this group of the analyzes, the participants were trained on how to use the tools, and had the task sheet to follow. Thus, the success outcome was high for both tools, only having a 20% of failure in the task T02-OE and T02-PC, as it is shown in the Figure 20. In the task T02-OE one of the participants was not able to properly understand what to do, and did not described the columns in the correct way. While in the

T02-PC, one of the participants used the wrong component to schedule the publication, failing in the task. This demonstrate that with a simple training both tools have the same degree of difficulty. But is important to make clear that the participants had a small gape of time in the use of the tools, thereby, the results could be different if they had to wait days to use the tools again.

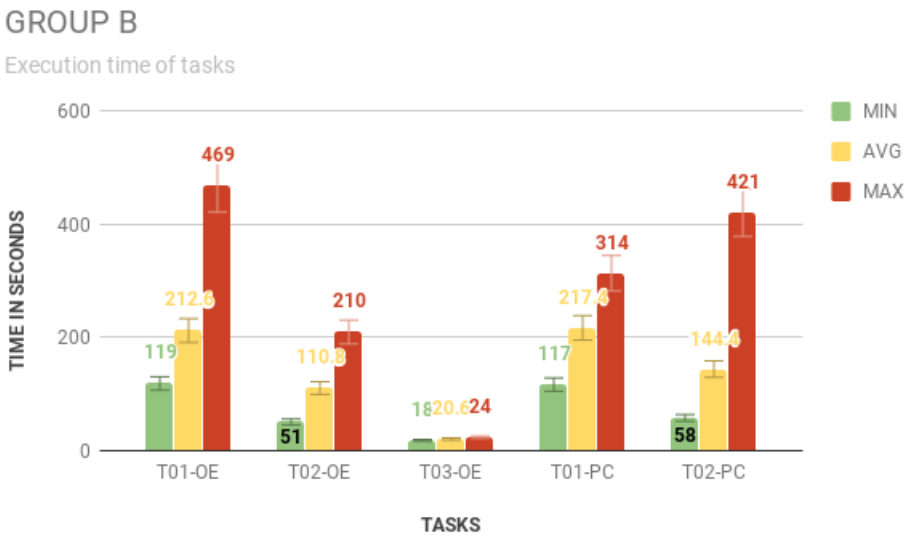
Figure 20 – Chart Outcome Group B



Font: Authors

The results of the execution time of Group B are presented in the Figure 21. It is possible to see that the execution time was similar between T01-OE and T01-PC, but one of the participant still was lost to select the table. This was the reason of the 469 seconds of the max time in the T01-OE, however, the participant was able to complete the task with success. While in T03-OE and T02-PC the difference was still big, where the participants took approximately 600% more to complete T02-PC than T03-OE. While analyzing the videos, the participants shown a more precise execution of the tasks, since they had training before executing the tasks from the Group B.

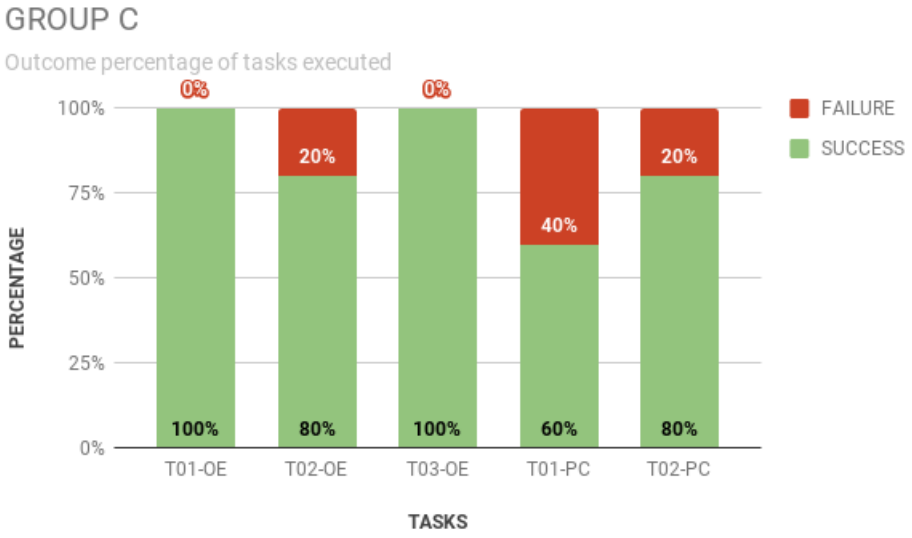
Figure 21 – Chart Execution Time Group B



Font: Authors

Group C. In the Group C, the last group, the participants had no task sheet to follow, relying only in what they learned while using the tools. The Figure 22 shows the outcome from the Group C. Here the participants were able to accomplish the tasks T01-OE and T03-OE with 100% of success outcome. And in the T01-PC and T02-PC the participants had a outcome of 60% and 80% of success respectively. This results shows that using OpenEasier is without the task sheet is less complicated than using Pentaho CKAN.

Figure 22 – Chart Outcome Group C

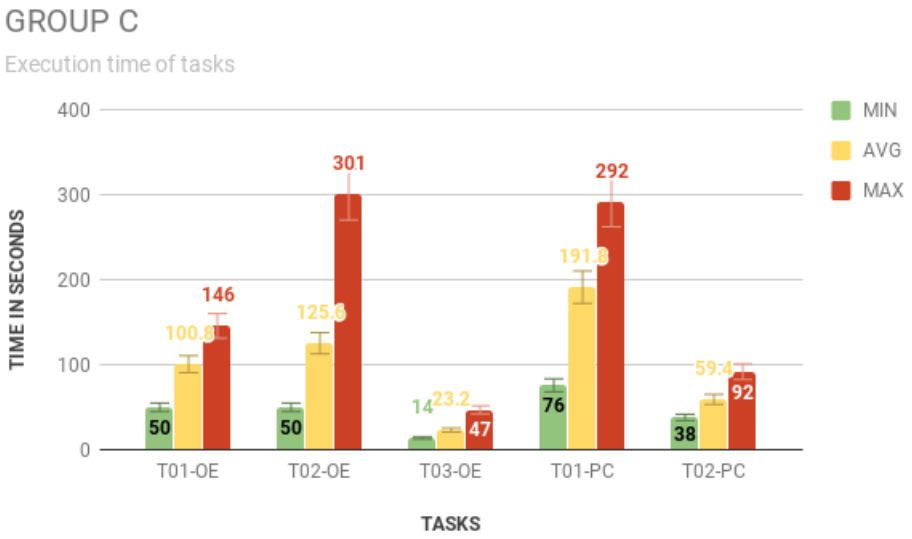


Font: Authors

In the Figure 23 the execution time of the Group C is displayed. Both tools had an improvement in the results of the execution time, but OpenEasier stills goes better. The average time of T01-PC was almost twice compared to the average execution time of T01-0E, and T01-OE had 100% of success outcome, demonstrating that OpenEasier is easier to use and learn. There was also a good improvement in the execution time of T02-PC, decreasing to 59.4 seconds of execution time, compared to the average of 144.4 seconds in the Group B.

When analyzing the the participants expressions and comments in the videos, in general, they appear to be lost while using Pentaho CKAN, not remembering the name of the steps that they had to choose to complete the tasks. But while using OpenEasier, the participants shown more confident and were able to finish the tasks properly.

Figure 23 – Chart Execution Time Group C



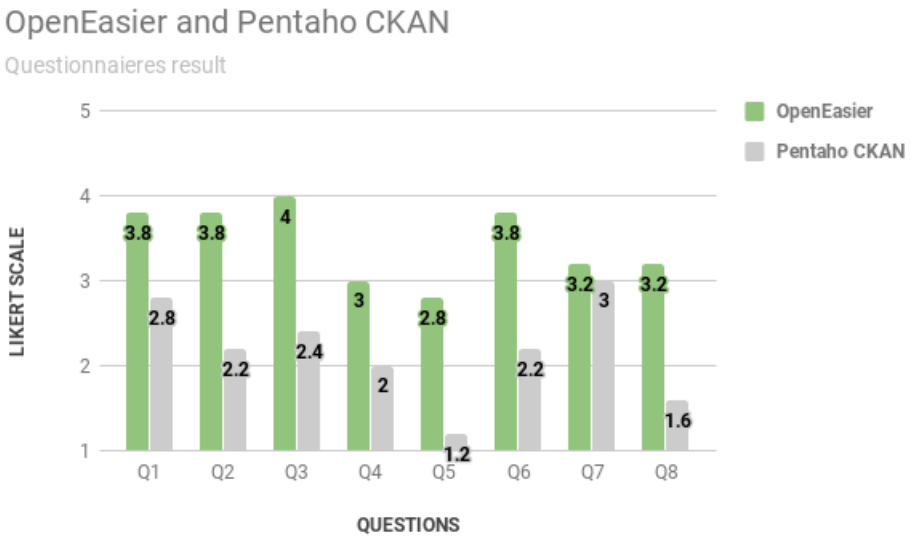
Font: Authors

Questionnaire. The Figure 24 shows the average result of the tools in each question of the post-questionnaire, having a total of eighth questions. We can see an overall score of the tools, so we can be able to compare how the participants experienced the use each tool. We have classified each question of the questionnaire, relating it to one or more aspects of usability that we were evaluating in the tools. The questions Q1, Q6 and Q8 are all related to the aspect of satisfaction. Questions Q2 and Q7 are pertinent to the measure of efficiency. Questions Q3 and Q4 are related to "learnability" and efficiency. And last, Q5 is connected to the measure of the aspect of "learnability". Compared to Pentaho CKAN, OpenEasier was able to score higher in every question of the questionnaire, thereby, having a better acceptance from the participants.

There was also a space in the questionnaire where the participants could write a

free commentary about the tool. Summarizing the answers, the participants wrote that OpenEasier had a good interface, intuitive, and with clear steps to follow. Only one participant said that OpenEasier was a little confuse in the first attempt. As for the answers related to Pentaho CKAN, the participants reported difficulties and a not so intuitive interface. A participant even reported that using Pentaho CKAN was complex and that it was necessary to be trained to use the tool.

Figure 24 – Chart Questionnaires Result



Font: Authors

Discussion of the results. With all the results demonstrated, it is reasonable to conclude that, when the participants used OpenEasier, they had a higher success rate and took less time that than when using Pentaho CKAN. Of course that the profile of the participants must be taken in consideration, since the participants had little knowledge in IT, and Pentaho CKAN is a tool that requires technical knowledge to use it. However, for the experiment, we have abstracted most of the technical knowledge existent to use Pentaho CKAN. For example, the SQL query was provided to the participants, thus, there was no need for the participants to write the query to select the data. But in OpenEasier the participants had to select each column using the features of OpenEasier, giving more autonomy to the participants. Another important conclusion is that to properly use OpenEasier it is essential to have a good understand of Open Data. Some of the failures of the participants while using OpenEasier was due to the lack of understand of the concept of Open Data, and not because the interface was complex to use. Even with the introduction given in the begin to the participants, it was still not enough for some participants properly understand the concepts of Open Data. Now, answering the research usability questions elaborated in the section 6.4:

- **ANSWER RUQ-01** - Comparing the effectiveness of both tools, the participants had a better performance outcome in the tasks while using OpenEasier, in exception of the Group B that both tool scored equal. Therefore, it is more efficient to use OpenEasier to publish Open Data than using Pentaho CKAN.

- **ANSWER RUQ-02** - Regard the aspect of efficiency, the average execution time while using OpenEasier was lower than while using Pentaho CKAN. Moreover, the participants replied that using OpenEasier had a lesser human effort, being less complex to use. Thus, operating OpenEasier to publish Open Data is faster than using Pentaho CKAN.

- **ANSWER RUQ-03** - As for the aspect of "learnability", the comparison in the Group A shows that is more easy to learn how to use OpenEasier, since the success rate was higher without training. And in the Group C, the participants also had a higher success rate using OpenEasier, demonstrating that is more simple to learn and remember how to execute the steps using OpenEasier. Therefore, using OpenEasier is more easier than using Pentaho CKAN without training.

- **ANSWER RUQ-04** - Concerning the aspect of satisfaction, analyzing the participants feedback from the videos and questionnaires, it is possible to conclude that the participants felt more satisfied when using OpenEasier compared to Pentaho CKAN. Additionally, the participants also reported that felt more autonomy when using OpenEasier rather than using Pentaho CKAN. Thus, overall, OpenEasier satisfied more the participants.

Concluding, the experiment covered the four aspect of usability intent in the evaluation of the tools. Showing that OpenEasier, in general, is a better option to use to publish Open Data, in the situation where the user is the business technician.

7 Final Considerations

This Chapter will detail the conclusions of the research work presented. The section 7.1 presents the conclusions. The section 7.2 discuss the limitations of the work. And last, the section 7.3 discuss the future work of the research.

7.1 Conclusions

The Open Data movement is becoming increasingly popular in the past years, this helps to highlight the benefits and obstacles of the movement, showing how important it is to our society. The publication and management of the data being open is the obstacle discussed in this study. We attempted to understand the existing tools used to publish and manage Open Data, and to accomplish that, we created a taxonomy to proper evaluate and understand those tools. We concluded that the existing tools are exclusively of use of those who have technical knowledge in IT, being a factor that hampers the adoption of the Open Data movement. Thus, this research aimed to improve the current methods of publishing and managing Open Data, focusing in allowing the non-IT technician to publish and manage Open Data.

As an effort to accomplish the goals of this research, we have designed and developed a tool, named OpenEasier, to allow non-IT technicians to publish and manage Open Data. The design of the tool begun based on the taxonomy created in the evaluation of the existing tools, and comprehension of the Open Data concepts in the Brazilian context. This was crucial to design a tool that attempt to the necessities of the Brazilian's institutions. After the design of the tool, we established the technologies based on the requirements and started the development of the tool. We have concluded a first version of the tool, with the most important features to attain what was proposed in this work. To validate the developed tool we have design and performed an usability experiment, defining and measuring usability aspects to compare OpenEasier with a chosen tool from the work executed to understand the existing tools. The results of the experiment made possible to establish that OpenEasier attempted the goals of this study, showing that is possible to the business technician select and schedule the data to be open.

This work has its importance because Open Data is a concept that provides good results for our society, and as it was demonstrated in this study, the existing tools hamper the Open Data movement. When providing such tool as the proposed one, the institutions will have a less complex and costly option to open the data. Thereby, allowing more institutions to open their data, looking to the benefits that Open Data provides to our society.

7.2 Limitations

Concerning the limitations of the study, we have identified a few. The first is that it is necessary a context to use the tool, which means that there must exist an Information System to have the data extracted. Moreover, the users of the OpenEasier must be the ones that use the Information System of the institution, and knows about the data. This limits the use of the tool, since not every institution has an Information System. It is highly important that the user has a good understand of what is Open Data. This is crucial so that the user can be able to properly manage and publish the data.

A factor that can be considered as limitation is that OpenEasier is only capable to properly present the data to the users if the database is well structured. Tables must have names that indicate its purpose and primary keys and foreign keys must be well defined. Without that, OpenEasier is not able to properly map the tables and present to users to select the data. This could cause a misleading in the data being published, being something to be careful when using the tool.

There is also a limitation of Information Systems with the quality attribute of multi-tenancy, where multiple users use the same environment (system) to store the data. Right now, OpenEasier has no option to configure a filter to solve this problem, and this could lead to the wrong data being published. Therefore, a concern that should be take care of in future versions of the tool.

Last, the evaluation of the tool had its limitations. Because of the lack of time and availability, we could not properly test the tool in a real environment, which should be the ideal case. For example, in the experiment, none of the users knew what is Open Data, and this clearly affected the performance of the participants when using OpenEasier. Therefore, a more broader public of participants should be ideal to better evaluate the tool.

7.3 Future Work

As for the future work, we have identified a few important activities and features to work on. One of those activities was the evaluation. In spite the fact of the proven success of OpenEasier through the evaluation, it should be more accurate to evaluate the tool in a real environment. This is important to have real insights of the use of the tool, allowing to improve the tool through those insights. This was not done because of the lack of time and also necessaries cares during the opening of the data.

Another import feature to work on is the implementation of support to multi-tenancy. This feature is important to make possible to extract data from multi-tenancy systems. The IT technician should be able to configure specific filters to each business

technician, allowing them to only select and publish the data related to the institution where the business technician works.

Also a limitation to work on is the data source and data formats that OpenEasier can handle. Extracting the data from only one data source (relational database) limits the cases where OpenEasier can be used, therefore, an important feature is to be capable of extracting data from other sources, such as CSV files. The data formats output of OpenEasier is also another feature to improve, enabling the IT technician to configure other data formats outputs such as JSON and XML. This will give more possibilities to manipulate the data for the ones that will consume the resources of the dataset.

And last, an important aspect to improve the quality of the data being published is the Data Quality. In the current version, OpenEasier only gives a simple evaluation of the data being published, e.g. missing data and data types variation. Therefore, it would be relevant for the study to work on ways to allow the IT technician to configure metrics of Data Quality to be evaluated while publishing the data. This will enable more customization and help to improve the quality of the data being open, benefiting those who will consume the data.

Bibliography

ABRAN, A. et al. Usability meanings and interpretations in iso standards. *Software quality journal*, Springer, v. 11, n. 4, p. 325–338, 2003. Cited in page 56.

AMBLER, S. W. *The Elements of UML (TM) 2.0 Style*. [S.l.]: Cambridge University Press, 2005. Cited in page 38.

ARAÚJO, N. M. d. *Dados abertos do governo brasileiro: entendendo as perspectivas de fornecedores de dados e desenvolvedores de aplicações ao cidadão*. Dissertação (Mestrado) — UFRN, Brasil, 2017. Cited 2 times in the pages 14 and 16.

BATINI, C. et al. Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, ACM, v. 41, n. 3, p. 16, 2009. Cited in page 25.

BEVAN, N.; CARTER, J.; HARKER, S. Iso 9241-11 revised: What have we learnt about usability since 1998? In: SPRINGER. *International Conference on Human-Computer Interaction*. [S.l.], 2015. p. 143–151. Cited in page 56.

BRASIL. *Decreto 8.777, de 11 de Maio de 2016*. Diário Oficial da união, 2016. Accessed in December 23, 2017. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm>. Cited in page 13.

DIETRICH, D. et al. *Open data handbook*. 2009. Cited in page 20.

DOMINGO, A. et al. Public open sensor data: Revolutionizing smart cities. *IEEE Technology and Society Magazine*, IEEE, v. 32, n. 4, p. 50–56, 2013. Cited in page 12.

JANSSEN, M.; CHARALABIDIS, Y.; ZUIDERWIJK, A. Benefits, adoption barriers and myths of open data and open government. *Information systems management*, Taylor & Francis, v. 29, n. 4, p. 258–268, 2012. Cited 2 times in the pages 12 and 20.

JARDIM, J. M. A lei de acesso à informação pública: dimensões político-informacionais. 2013. Cited in page 13.

KIMBALL, R.; CASERTA, J. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. [S.l.]: John Wiley & Sons, 2011. Cited in page 14.

KNAP, T. et al. Unifiedviews: Towards etl tool for simple yet powerfull rdf data management. In: *DATESO*. [S.l.: s.n.], 2015. p. 111–120. Cited in page 28.

KUCERA, J.; CHLAPEK, D. Benefits and risks of open government data. *Journal of Systems Integration*, Journal of Systems Integration, v. 5, n. 1, p. 30, 2014. Cited 3 times in the pages 12, 14, and 16.

KUCERA, J. et al. Methodologies and best practices for open data publication. In: *DATESO*. [S.l.: s.n.], 2015. p. 52–64. Cited in page 27.

LAKOMAA, E.; KALLBERG, J. Open data as a foundation for innovation: The enabling effect of free public sector information for entrepreneurs. *IEEE Access*, IEEE, v. 1, p. 558–563, 2013. Cited in page 12.

MARCH, S. T.; SMITH, G. F. Design and natural science research on information technology. *Decision support systems*, Elsevier, v. 15, n. 4, p. 251–266, 1995. Cited in page 17.

MATHEUS, R.; RIBEIRO, M. M.; VAZ, J. C. Brazil towards government 2.0: Strategies for adopting open government data in national and subnational governments. In: *Case Studies in e-Government 2.0*. [S.l.]: Springer, 2015. p. 121–138. Cited 2 times in the pages 13 and 17.

MOLLOY, J. C. The open knowledge foundation: open data means better science. *PLoS Biol*, v. 9, n. 12, p. e1001195, 2011. Cited in page 12.

NIELSEN, J. *Usability 101: Introduction to usability*. 2003. Cited in page 56.

OJO, A.; CURRY, E.; ZELETI, F. A. A tale of open data innovations in five smart cities. In: IEEE. *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. [S.l.], 2015. p. 2326–2335. Cited in page 12.

PEFFERS, K. et al. A design science research methodology for information systems research. *Journal of management information systems*, Taylor & Francis, v. 24, n. 3, p. 45–77, 2007. Cited in page 17.

PIPINO, L. L.; LEE, Y. W.; WANG, R. Y. Data quality assessment. *Communications of the ACM*, ACM, v. 45, n. 4, p. 211–218, 2002. Cited in page 25.

RUBIN, J.; CHISNELL, D. *Handbook of usability testing: howto plan, design, and conduct effective tests*. [S.l.]: John Wiley & Sons, 2008. Cited 2 times in the pages 56 and 60.

SAYOGO, D. S. et al. Going beyond open data: Challenges and motivations for smart disclosure in ethical consumption. *Journal of theoretical and applied electronic commerce research*, SciELO Chile, v. 9, n. 2, p. 1–16, 2014. Cited in page 14.

SEFFAH, A. et al. Usability measurement and metrics: A consolidated model. *Software Quality Journal*, Springer, v. 14, n. 2, p. 159–178, 2006. Cited in page 56.

SOMMERVILLE, I. et al. *Engenharia de software*. [S.l.: s.n.], 2010. v. 8. Cited in page 38.

UBALDI, B. Open government data: Towards empirical analysis of open government data initiatives. *OECD Working Papers on Public Governance*, Organisation for Economic Cooperation and Development (OECD), n. 22, p. 0_1, 2013. Cited 2 times in the pages 14 and 21.

VAISHNAVI, V. K.; KUECHLER, W. *Design science research methods and patterns: innovating information and communication technology*. [S.l.]: Crc Press, 2015. Cited 2 times in the pages 17 and 18.

VASSILIADIS, P. A survey of extract–transform–load technology. *International Journal of Data Warehousing and Mining (IJDWM)*, IGI Global, v. 5, n. 3, p. 1–27, 2009. Cited in page 23.

VENABLE, J.; PRIES-HEJE, J.; BASKERVILLE, R. A comprehensive framework for evaluation in design science research. In: SPRINGER. *International Conference on Design Science Research in Information Systems*. [S.l.], 2012. p. 423–438. Cited in page 18.

WINN, J. et al. Open data and the academy: An evaluation of ckan for research data management. 2013. Cited in page 13.

Appendix

APPENDIX A – Term of Consent

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

Termo de Consentimento Livre e Esclarecido – TCLE

BASEADO NAS DIRETRIZES CONTIDAS NA RESOLUÇÃO CNS Nº466/2012, MS.

Você foi convidado(a) a participar de uma pesquisa de avaliação de duas ferramentas de publicação de Dados Abertos. Esta pesquisa está sendo desenvolvido através do Programa de Pós-Graduação em Engenharia de Software (PPGSW) da UFRN/IMD, pelo mestrando Jonas Jordão, sob a orientação do Prof. Dr. Frederico Lopes e Prof. Dr. Nélio Cacho. Seu objetivo é analisar a eficiência das ferramentas OpenEasier e Pentaho CKAN, realizando uma comparação entre ambas, com o intuito de identificar qual é a ferramenta mais apropriada para a publicação e gerência de Dados Abertos. É importante destacar que não estamos interessados em avaliar seu conhecimento nessa área. Apenas buscamos investigar como os sistemas se fazem capazes de ser entendido e utilizados pelo usuário. Gostaríamos de ouvir suas opiniões e comentários sobre os sistemas, pois seu ponto de vista é muito importante para nós. Por estas razões, solicitamos seu consentimento para observarmos a realização das tarefas solicitadas, bem como para realização de uma breve entrevista sobre sua experiência. Vamos gravar o que for falado, gravar sua interação com o sistema e coletar suas respostas de identificação. Para tanto, é importante que você tenha algumas informações adicionais: Os dados coletados destinam-se estritamente a atividades de pesquisa. Somente os pesquisadores envolvidos terão acesso aos dados brutos; A divulgação dos resultados desta pesquisa pauta-se no respeito à sua privacidade e ao seu anonimato em quaisquer documentos elaborados; O consentimento para participar desta pesquisa é uma escolha livre, feita mediante a prestação de todos os esclarecimentos necessários sobre a pesquisa; Você tem toda liberdade para interromper as atividades e desistir de participar da pesquisa. Neste caso, os pesquisadores se comprometem a descartar os dados coletados com sua contribuição; Você pode entrar em contato conosco pelo e-mail *jonasjordao452@gmail.com*.

De posse das informações acima, gostaríamos que você se pronunciasse acerca da sua participação nesta pesquisa.

( ) Dou meu consentimento para sua realização.
( ) Não autorizo sua realização.

Natal, ____ de _____ de 2018.

Participante
Nome:
Assinatura:

Pesquisador(a)
Nome:
Assinatura

APPENDIX B – Participants Guideline

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

ROTEIRO DO PARTICIPANTE

Olá e muito obrigado por aceitar participar da Pesquisa de Avaliação de Usabilidade. Suas contribuições serão muito importantes para a avaliação das ferramentas.

POR FAVOR, LEIA ESTE ROTEIRO COM CUIDADO E ATENÇÃO

Após a leitura você pode sanar todas as suas dúvidas com os pesquisadores. Porém, após o início da atividade, nenhuma pergunta será respondida, pois as dificuldades com que você compreende as tarefas e as interfaces é parte do que estamos avaliando nas ferramentas e no procedimento de teste.

Importante frisar que são as ferramentas que estão sendo avaliadas, e não você. Não hesite em executar as tarefas da maneira que entendeu no enunciado e na interface.

Caso ocorra algum problema técnico no funcionamento do software, não se preocupe. Os pesquisadores lidará com a situação, suspendendo a atividade até o restabelecimento da condição da avaliação. Se necessário, a avaliação será remarcada conforme a sua disponibilidade.

Muito importante!

Durante a atividade:

- Execute uma tarefa por vez, na ordem apresentada;
- Pedimos que fale em voz alta o que está pensando ou o motivo de estar executando cada ação. Caso o resultado não seja o esperado, fale o que esperava que acontecesse;
- Não espere respostas do moderador para suas indagações durante as tarefas;
- Quando finalizar uma tarefa, avise aos pesquisadores que considera a tarefa concluída;
- Se estiver com dificuldade na realização da tarefa pedida, você pode desistir a qualquer momento. Avise aos pesquisadores, informando o motivo.

Mais uma vez, muito obrigado pela participação!

APPENDIX C – T01-PC

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

Pentaho CKAN - TAREFA 01

INTRODUÇÃO:

A instituição que você trabalha precisa tornar os seus dados públicos, buscando atender a Lei de Acesso à Informação (LAI). Portanto, foi elaborado um Plano de Dados Abertos, e passado a você a tarefa de abertura de um conjunto de dados específico, que faz parte do seu contexto de trabalho. O conjunto de dados solicitado a você foi os dados relacionados aos agrotóxicos. Tais dados estão armazenados no Sistema CERES, no qual você usa em seu ambiente de trabalho. Para selecionar os dados a serem publicados você deve utilizar a ferramenta Pentaho CKAN, indicada e pré configurada pelos Técnicos de TI da sua instituição. Na ferramenta você vai criar uma Transformação que fará o trabalho de extrair os dados do Sistema CERES e publicá-los.

1º PASSO:
Para realizar a publicação dos dados é necessário que se crie primeiro uma nova Transformação, onde será feito o trabalho de extração e publicação dos dados.

2º PASSO:
Para realizar a extração dos dados é necessário configurar o ***Step*** de ***Table input***, encontre-o na aba de ***Design*** este ***Step*** e arraste-o para a tela da Transformação .

3º PASSO:
Agora você deve editar o ***Step*** de ***Table input*** selecionado, e nele informar o ***SQL*** para extrair os dados.

Digite o seguinte comando:
SELECT * FROM public.iv_agrotoxico;

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

4º PASSO:
Após configurar o ***Table input***, procure pelo ***Step*** do ***CKAN DataStore Upload*** e arraste-o para a tela da Transformação .

5º PASSO:
Agora você deve editar o ***Step*** de ***CKAN DataStore Upload*** selecionado, e nele informar os dados do CKAN e do recurso.

Domain:
http://ckan.opendataprocessor.com/

API Key:
12460910-c32f-4a6a-9102-f4f99eef8f32

Package ID:
agrotoxicos

Resource Title:
Agrotóxicos RN

Description:
Recurso com os agrotóxicos do RN.

6º PASSO:
Agora você deve ligar o ***Step*** de ***Table input*** com o ***CKAN DataStore Upload***, criando uma interação entre ambos, onde os dados serão extraídos pelo ***Table input*** e publicados pelo ***CKAN DataStore Upload***.

7º PASSO:
Por fim, você deve salvar o arquivo de Transformação.

APPENDIX D – T02-PC

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

Pentaho CKAN - TAREFA 02

INTRODUÇÃO:

Ao criar a Transformação para publicação de dados, foi solicitado que você realize o agendamento da publicação dos dados. O agendamento é importante pois é nele que você vai informar a frequência de atualização desse dado. A frequência de atualização deve ser definida pelo Plando e Dados Abertos. Esse agendamento é importante pois vai manter os dados que estão abertos sempre atualizados, trazendo um maior enriquecimento dos dados, beneficiando a quem foi utiliza-lós. Para realizar o agendamento, você deve fazer uso da ferramenta Pentaho CKAN, criando um ***Job*** na ferramenta para realizar o agendamento.

1º PASSO:
Na aba de ***Design***, você deve procurar pela ***Entries*** de ***START***, e deve arrastar o mesmo para a tela de ***Job.***

2º PASSO:
Agora você deve editar a ***Entries*** de ***START*** selecionado, e nele informar os seguintes dados:

Repeat:
MARCAR

Type:
Weekly

3º PASSO:
Após configurar o ***START***, procure pelo ***Entries*** de Transformação arraste-o para a tela do ***Job.***

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

4º PASSO:
Agora você deve editar ***Entries*** de Transformação, selecionando a Transformação *que você criou na tarefa passada. Você deve ir em Browse e procure pelo arquivo que está em:*
C:\Users\Jonas\Desktop\transformation_ckan.krt

5º PASSO:
Agora você deve ligar o ***Entries*** de ***START*** com o Transformação, criando uma interação entre ambos, onde assim será agendado a Transformação de publicação dos dados de acordo com as configurações realizadas.

6º PASSO:
Por fim, você deve salvar o arquivo ***Job***.

APPENDIX E – T01-OE

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

OpenEasier - TAREFA 01

INTRODUÇÃO:

A instituição que você trabalha precisa tornar os seus dados públicos, buscando atender a Lei de Acesso à Informação (LAI). Portanto, foi elaborado um Plano de Dados Abertos, e passado a você a tarefa de abertura de um conjunto de dados específico, que faz parte do seu contexto de trabalho. O conjunto de dados solicitado a você foi os dados relacionados aos agrotóxicos. Tais dados estão armazenados no Sistema CERES, no qual você usa em seu ambiente de trabalho. Para selecionar os dados a serem publicados você deve utilizar a ferramenta OpenEasier, indicada e pré configurada pelos Técnicos de TI da sua instituição.

1º PASSO:
Realizar o login no sistema OpenEasier.

Credenciais:
Login: user
Senha: openeasier452

2º PASSO:
Iniciar a publicação dos dados. Deve ser pesquisado pelos dados referentes aos agrotóxicos escolhendo a tabela que contém os dados de acordo com o seu contexto de trabalho.

Sistema de origem dos dados:
SISTEMA CERES

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

3º PASSO:
Selecionar as colunas primárias da tabela de agrotóxicos.

Colunas primárias para serem selecionadas:
Nome
Corrosivo
Inflamável
Status
Orgânico

4º PASSO:
Após selecionar as colunas primárias, deve ser lecionado as colunas secundárias para cada tabela relacionada aos dados de agrotóxicos, sendo estes dados considerados como dados secundários a tabela principal.

Colunas secundárias para serem publicadas:

Referente a classificação do agrotóxico, você deve selecionar as colunas:
Nome
Sigla
Classificação

Referente ao registrante do agrotóxico, você deve selecionar a colunas
Nome

Referente a formulação do agrotóxico, você deve selecionar a colunas
Nome
Sigla

5º PASSO:
Finalizar a publicação do recurso escolhendo o conjunto de dados de agrotóxicos. Você deve informar o nome do recurso (que seja um nome sugestivo) e uma descrição sobre o recurso. Tais informações serão vista por quem vai utilizar os dados, ajudando a fazer sentido sobre o que são os dados.

APPENDIX F – T02-OE

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

OpenEasier - TAREFA 02

INTRODUÇÃO:

Após a criação do recurso realizada na tarefa anterior, é necessário que você crie o dicionário de dados. No dicionário você deve dar uma descrição geral sobre o conjunto de dados de agrotóxicos publicado, e uma descrição sugestiva para cada coluna do conjunto de dados. O dicionário de dados será utilizado pelas pessoas que tiverem interesse em consumir os dados, permitindo com que eles possam entender melhor os dados que foram publicados. Para isso, você deve utilizar a ferramenta OpenEasier.

1º PASSO:
Ao acessar o Painel de Recursos, escolha o recurso previamente criado e clique na opção “Dicionário de Dados”.

2º PASSO:
Você deve descrever o dicionário de dados como solicitado, explicando o que é este conjunto de dados.

3º PASSO:
Descreva cada coluna previamente selecionada para publicação, informando um breve texto sugestivo para cada coluna. Por fim, finalize para salvar as informações.

APPENDIX G – T03-OE

Universidade Federal do Rio Grande do Norte
Instituto Metrópole Digital
Programa de Pós-Graduação em Engenharia de Software - PPGSW

OpenEasier - TAREFA 03

INTRODUÇÃO:

Ao criar o recurso e o seu dicionário de dados, foi solicitado que você realize o agendamento da publicação dos dados. O agendamento é importante pois é nele que você vai informar a frequência de atualização desse dado. A frequência de atualização deve ser definida pelo Plano de Dados Abertos. Esse agendamento é importante pois vai manter os dados que estão abertos sempre atualizados, trazendo um maior enriquecimento dos dados, beneficiando a quem foi utiliza-lós. Para realizar o agendamento, você deve fazer uso da ferramenta OpenEasier.

1º PASSO:

Ao acessar o Painel de Recursos, escolha o recurso previamente criado e clique na opção “Agendar”.

2º PASSO:

Agende a publicação para ser realizada na data de hoje (16/05/2018), com a frequência semanal. Por fim, finalize para salvar as informações.

APPENDIX H – Demographic Questionnaire

Perfil dos Participantes

Formulário de participantes para traçar os perfis dos mesmos. Os dados informados aqui serão utilizados na pesquisa SEM A IDENTIFICAÇÃO do participante.

* Required

1. **Nome completo** *

2. **Idade** *

3. **Sexo** *
   Mark only one oval.
   - ◯ Masculino
   - ◯ Feminino

4. **Nível de Educação** *
   Mark only one oval.
   - ◯ Ensino Médio
   - ◯ Ensino Superior
   - ◯ Pós-graduação
   - ◯ Mestrado
   - ◯ Doutorado

5. **Em qual instituição você trabalha?** *

6. **Qual seu cargo na instituição?** *

7. **Qual sua área de atuação na instituição?** *

8. **Qual sua experiência no uso de ferramentas de computador?** *
   Mark only one oval.
   - ◯ Não utilizo
   - ◯ Utilizo ocasionalmente
   - ◯ Utilizo frequentemente

9. **Antes do treinamento, você tinha algum conhecimento sobre Dados Abertos?** *
Mark only one oval.

- ◯ Sim
- ◯ Não

10. **Se sim, já utilizou ou publicou Dados Abertos?**
Mark only one oval.

- ◯ Sim
- ◯ Não

Powered by Google Forms

APPENDIX I – OpenEasier Usability Questionnaire

Questionário de Usabilidade (OpenEasier)

Este é o questionário de usabilidade referente a ferramenta OpenEasier. Você deve responder este questionário de acordo com as suas impressões durante a execução das tarefas utilizando a ferramenta OpenEasier.

* Required

1. **Nome completo** *

2. **Estou satisfeito(a) de modo geral com a ferramenta OpenEasier.** *
   Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

3. **Foi fácil executar as tarefas utilizando a ferramenta OpenEasier.** *
   Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

4. **Os passos para utilizar a ferramenta OpenEasier são intuitivos.** *
   Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

5. **Me senti capaz de realizar as tarefas utilizando a ferramenta OpenEasier sem a ajuda externa.** *
   Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

6. **O nível técnico exigido para utilizar a ferramenta OpenEasier é baixo.** *
   Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

7. **Os termos e mensagens utilizados na ferramenta OpenEasier estão de acordo com os meus conhecimentos.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

8. **Os valores padrões da ferramenta OpenEasier me ajudaram a introduzir os dados que eu precisava.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

9. **A ferramenta OpenEasier retornou respostas claras quando uma ação era realizada com sucesso.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

10. **Comente sobre sua experiência geral utilizando a ferramenta OpenEasier. (pontos positivos e negativos)**

Powered by Google Forms

APPENDIX J – Pentaho CKAN Usability Questionnaire

Questionário de Usabilidade (Pentaho CKAN)

Este é o questionário de usabilidade referente a ferramenta Pantaho CKAN. Você deve responder este questionário de acordo com as suas impressões durante a execução da tarefa utilizando a ferramenta Pentaho CKAN.

* Required

1. **Nome completo** *

_____

2. **Estou satisfeito(a) de modo geral com a ferramenta Pentaho CKAN.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

3. **Foi fácil executar as tarefas utilizando a ferramenta Pentaho CKAN.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

4. **Os passos para utilizar a ferramenta Pentaho CKAN são intuitivos.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

5. **Me senti capaz de realizar as tarefas utilizando a ferramenta Pentaho CKAN sem a ajuda externa.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

6. **O nível técnico exigido para utilizar a ferramenta Pentaho CKAN é baixo.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

7. **Os termos e mensagens utilizados na ferramenta Pentaho CKAN estão de acordo com os meus conhecimentos.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

8. **Os valores padrões da ferramenta Pentaho CKAN me ajudaram a introduzir os dados que eu precisava.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

9. **A ferramenta Petanho CKAN retornou respostas claras quando uma ação era realizada com sucesso.** *
Mark only one oval.

	1	2	3	4	5	
Discordo totalmente	◯	◯	◯	◯	◯	Concordo totalmente

10. **Comente sobre sua experiência geral utilizando a ferramenta Pentaho CKAN:**