Bioinformatics Resource Portal
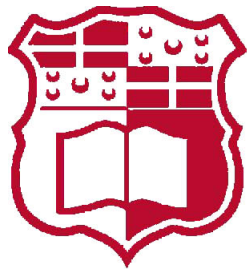
Nigel Alfino

Supervisor(s): Mr. Joseph Bonello and Prof. Ernest Cachia

Faculty of ICT

University of Malta

May 2019

UNIVERSITY OF MALTA
FACULTY/INSTITUTE/CENTRE/SCHOOL_____of ICT_____

DECLARATION OF AUTHENTICITY FOR UNDERGRADUATE STUDENTS

Student's I.D. /Code ____0446998(M)_____

Student's Name & Surname _____Nigel Alfino_____

Course ____B. Sc. I.T (Hons.) in Software Development_____

Title of Long Essay/Dissertation

___Bioinformatics Resource Portal_____

_____

_____

I hereby declare that I am the legitimate author of this Long Essay/Dissertation and that it is my original work.

No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education.

I hold the University of Malta harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

_____   _____
Signature of Student   Date

11.06.2015

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Declaration

Plagiarism is defined as "the unacknowledged use, as one's own, of work of another person, whether or not such work has been published, and as may be further elaborated in Faculty or University guidelines" (University Assessment Regulations, 2009, Regulation 39 (b)(i), University of Malta).

I / We*, the undersigned, declare that the [assignment / Assigned Practical Task report / Final Year Project report] submitted is my / our* work, except where acknowledged and referenced.

I / We* understand that the penalties for committing a breach of the regulations include loss of marks; cancellation of examination results; enforced suspension of studies; or expulsion from the degree programme.

Work submitted without this signed declaration will not be corrected, and will be given zero marks.

* Delete as appropriate.

(N. B. If the assignment is meant to be submitted anonymously, please sign this form and submit it to the Departmental Officer separately from the assignment).

_____Nigel Alfino_____   _____
Student Name   Signature

_____   _____
Student Name   Signature

_____   _____
Student Name   Signature

_____   _____
Student Name   Signature

____ICT3911_____   _____Bioinformatics Resource Portal_____
Course Code   Title of work submitted

_____
Date

Declaration of FYP Submission Size

Submission size is the maximum allowable size of a submitted Final Year Project report measured as the number of pages.

I, the undersigned, submit this FYP report with full knowledge that the recommended number of pages for the dissertation write-up is sixty (60) pages and that the maximum allowable number of pages is seventy-five (75), inclusive of the bibliography and the list of references.

The recommended 60-page limit and the maximum 75-page limit do not include the following:

- front / title pages,
- dedication,
- acknowledgements,
- page count declaration,
- declaration of authenticity,
- abstract,
- table of contents,
- table of figures,
- list of tables,
- glossary,
- list of abbreviations,
- appendices,
- annexes.

I am aware that any case to include pages exceeding the above-mentioned limits must be made through my supervisor at least one month prior to the date of my submission. Failing this, I understand that I will be given a maximum of 3 (three) working days in which to submit changes. I also understand that if I do not submit these changes on time, a 10 (ten) percent penalty may be applied to my final awarded mark. I also understand that if my final submission does not conform to the above-mentioned limits, it may not be accepted for marking.

I, the undersigned, also declare, that I am aware of the FYP submission guidelines as listed on the Faculty of ICT web-site:
https://www.um.edu.mt/__data/assets/pdf_file/0017/208160/Final_Year_Project_Harmonisation_Guidelines.pdf
and that my work conforms to these harmonised guidelines.

Work submitted without this signed declaration will not be corrected, and will be given zero (0) marks.

_____	Nigel Alfino	___ICT3911___
Date of submission	Student's full name	Study-unit code

____Bioinformatics Resource Portal_____

_____

 Title of submitted work (two above lines available)

_____ Student's signature

Abstract:

Bioinformatics is the field of study that applies computational techniques in order to understand and organise biological data. An obstacle found in performing research in bioinformatics is finding the right tool or dataset to use for that research. This obstacle rises from a lack of effort made in research to make tools or datasets reusable. This issue is also present due to a lack of an indexable resource of tools and datasets, as well as a lack of assessment of their reusability. The FAIR guiding principles provide four principles to measure a tool or dataset's Findability, Accessibiltiy, Interoperability and Reusability, as well as guidelines on how to perform such measurements. This study aims to create a searchable portal of tools, as well as providing an automated assessment tool which provides a FAIR score for tools or datasets based on a set of metrics. This will allow the user to see where the resource struggles to perform and can also refine the score with further information. The FAIR score obtained can provide researchers with a level of trust in the resources they use, as they will be able to properly gauge how Findable, Accessible and Reusable the tool is, and if it can Interoperate with different systems.

Acknowledgements:

I would like to express my deepest gratitude and appreciation to all those who helped and supported me along this journey in completing my dissertation.

Firstly, I would like to thank my supervisors, Mr. Joseph Bonello and Prof. Ernest Cachia, for their time and guidance, without which this dissertation would not have been possible.

I would also like to thank my friends and family for their constant support, reassurance and encouragement throughout my study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Definition

In recent years there has been an increase in the number of researchers performing studies in the field of bioinformatics, generating a large amount of tools to solve a wide array of problems in biology (Cannata, Merelli, & Altman, 2005). Due to this increase in tools, it has become difficult for researchers to track which tools, datasets and methods are available and useful for their area of study, especially due to the fact that there is no comprehensive index of tools, datasets, and literature annotated with proper metadata (Cannata et al., 2005).

The increase in the development of tools allowed researchers to be able to process this large amount of data being generated. The ability to process this large amount of data revolutionised research made in the biological field (Zakrisson & Kronfoth, 2017). However, there is limited literature on these tools, which also provides limited information based on the state of the tool when it was created. If updates were made to the tool, changing how a feature works, or increasing the tool's capabilities, these changes are not be covered in the literature (Cannata et al., 2005). The literature is also not classified by some standard based on the tools' features and capabilities, so finding the right tool can be difficult (Cannata et al., 2005). Furthermore, the continuous development of tools has led to an increase in this literature, making it even harder for researchers to find the tool they need (Grivell, 2002).

In order to ensure the quality of data gathered and used by researchers, a standard must be used to provide a quality metric for the tools used and research being developed. Research that can be of "good quality" can only be achieved through proper data management (Wilkinson et al., 2016). Data must be properly managed in order to produce research of better quality when that data is re-used in new research, either on its own or combined with newly generated data (Wilkinson

et al., 2016). The FAIR Guiding Principles (Wilkinson et al., 2016) provides four foundational principles that can be used as guidelines by researchers in order to ensure that the data provided in their research is not only properly obtained but also properly managed in order to be reused in future research, as well as making it findable to allow for timely generation of data (Wilkinson et al., 2016). The four principles provided by FAIR are as follows:

- Findability

- Accessibility

- Interoperability

- Reusability

1.2   Approach

The approach taken in this study involves the implementation of a web application that acts as a searchable repository for bioinformatics tools. A tool will also be developed to automate as much of the assessment of the FAIR score of tools in the repository as possible. The user can then further improve the score by providing any information which the tool might not have been able to find.

The web application will also allow the user to create pipelines of tools, allowing the user to see the FAIR score of the individual tools as well as an overall FAIR score of the pipeline. The user can then identify which tools are negatively impacting the FAIR score of the pipeline.

Evaluation is performed by applying the assessment tool on a set of bioinformatics tools and comparing the FAIR score obtained against the FAIR scores provided for the same tools on FAIRshake (*FAIRshake*, 2018), which will be used as a benchmark.

A number of case studies will then be performed using tools which were not evaluated by FAIRshake (*FAIRshake*, 2018), which will be evaluated using the tool and then manually checked to ensure that the results obtained are as expected.

1.3 Aims and Objectives

The main aims of this project are as follows:

- To conduct an extensive literature review to determine which tools are used in various bioinformatics pipelines;

- To build a model to describe the features and capabilities of bioinformatics tools;

- To build a tool that automates the assessment of a tool's Findability, Accessibility, Interoperability and Reusability (FAIR) and produces a FAIR score.

These aims will be attained through the completion of the following objectives:

- To obtain a comprehensive overview of tasks in bioinformatics and which tools are commonly used in the task;

- To obtain a searchable archive of tools that can be filtered by task, feature and capability;

- To enhance the archive with a FAIR score indicating the Findability, Accessibility, Interoperability and Reusability of the tool to aid researchers make decisions on which tools to use.

1.4 Research Questions

This study is aimed to answer the following list of research questions:

- Can the features of bioinformatics tools be modelled in such a way to provide a researcher insights into the tool's capabilities and usage context?

- Are the tools used in bioinformatics findable through simple searches and can a search provide insights into the tool's capabilities?

- Can a FAIR score for a particular tool help the user in finding the appropriate tool for their particular task?

- Can the FAIR assessment be part of an automated process?

From the above research questions, the following hypothesis was formed:
"Using a centralized repository of bioinformatics tools that captures the important features and capabilities of the tools, and by enhancing this information with a FAIR assessment score, users can make better judgements on which tools are appropriate for a particular bioinformatics pipeline."

1.5 Report Layout

This first chapter of the report provides a basic explanation on what the study will be addressing, what aims and objectives it aims to achieve as well as provides a brief explanation on how these aims and objectives will be achieved. The rest of the report is organised in the following manner. The second chapter provides information on the background and literature review on the area. The third chapter provides an analysis on how the application was designed. The fourth chapter provides a detailed description of how the application was implemented. The testing and evaluation of results obtained from the application are presented in the fifth chapter. This is followed by future works and finally the conclusions drawn.

Chapter 2

Background and Literature Review

2.1 Bioinformatics

Bioinformatics is a field that consists of multiple disciplines, namely those of computer science, mathematics, statistics, genetics and molecular biology. It is described as applying computational techniques - obtained from applied mathematics, computer science and statistics - in order to understand information and organise it (Luscombe, Greenbaum, & Gerstein, 2001). As a research discipline, bioinformatics is viewed as biology involving computation (Mayer, 2011). Bioinformatics is the field of study that aims to solve, though the use of a computational approach, large scale biological problems that are considered data intensive (Yousef & Allmer, 2014). Yousef & Allmer (2014) show how a solution in bioinformatics usually follows the procedure listed below:

- Biological data is collected from statistics.

- A computational model is built.

- A computational modelling problem is solved.

- The computational algorithm is tested and evaluated.

Due to the phenomenal rate at which data in biology is being produced, the use of bioinformatics is crucial for researchers to be able to process these large amounts of data, thanks to computers' abilities to process such data with relative ease (Luscombe et al., 2001).

The term bioinformatics was first used as early as 1977 by Paulien Hogeweg, a Dutch theoretical biologist, however the term gained a true meaning during the

"Bioinformatics in the 90s" conference (Attwood, Gisel, Eriksson, & Bongcam-Rudloff, 2011). Researchers' need to access and analyse biological data which was being generated at an increasing rate allowed the field of bioinformatics to grow organically. This need, alongside the development of algorithms and resources that allow for analysis and manipulation of large amounts of data allowed the field of bioinformatics to flourish and gain importance within the community of biological research (Attwood et al., 2011).

2.1.1 Aims of Bioinformatics

Research in the field of bioinformatics is made to achieve three main aims: (Luscombe et al., 2001)

- The simplest aim of bioinformatics is to organise data in a manner that would allow researchers to access and manipulate existing data with ease, as well as publish and access data as they are produced.

- Research in bioinformatics is focused on developing tools and resources that aid in the analysis of the data gathered. Although the gathering of data is a crucial task, this data is practically useless until it is analysed. The development of such tools requires proficiency in computational theory as well as in-depth knowledge in the field of biology.

- The tools developed are used to analyse and interpret the data generated in a biologically meaningful manner. Traditional biological studies analysed systems individually in great detail, however bioinformatics allows "global analyses with the aim of uncovering common principles that apply across many systems and highlight novel features" (Luscombe et al., 2001).

2.1.2 Datasets and Tools

The introduction of computers was the key to handling the data that was being generated (Attwood et al., 2011). Advancements made in computation allowed for the creation of multiple databases and datasets to facilitate the storage and manipulation of this large amount of data. These advancements gave rise to the creation of the most popular databases used in bioinformatics today, namely:

- NCBI Entrez (Maglott, 2004)

- EBI Ensemble (Flicek et al., 2012)

- UCSC Genome Browser (Kent et al., 2002)

- Kegg (Kanehisa, Goto, Sato, Furumichi, & Tanabe, 2011)

Despite great strides made in computation, analysis of biological data still faces many challenges such as noise or incompleteness (Yousef & Allmer, 2014). Inconsistencies can also arise from data stored on different databases; as well as redundancy in large amounts of overlapping data, not only due to the data being stored in multiple databases, but also due to similar data and datasets used to perform research (Yousef & Allmer, 2014; Zvelebil & Baum, 2008). Situations have arisen in past research where multiple datasets would need to be combined, each of them containing large amounts of overlapping data. The resulting dataset would end up being too large, making it inefficient to query, thus invalidating the first fundamental aim of bioinformatics (Zvelebil & Baum, 2008).

On top of all this, each database has its own format on how data is stored and managed (Yousef & Allmer, 2014). Although research has been carried out in order to maintain a standard such as that of the Gene Ontology consortium (Ashburner et al., 2000), until such a standard is employed throughout the multiple databases, researchers will still have to learn the specifics of each database in order to use it efficiently and effectively (Zvelebil & Baum, 2008).

2.1.3 Pipelines

The development of tools in the field of bioinformatics has allowed researchers to make great leaps in their discoveries due to the tools' ability process large amounts of data in a timely fashion (Zakrisson & Kronfoth, 2017). However, certain analyses in bioinformatics research that are of a high-throughput nature require the data being analysed to be transformed multiple times, through multiple tools in order to obtain the required result (Leipzig, 2016). This series of transformations is called a pipeline (Leipzig, 2016). Figure 2.1 shows the basic structure of a pipeline. As can be seen in the figure, a pipeline runs most of its analyses in parallel, as well as involving multiple downstream steps combined with report generation (Leipzig, 2016).
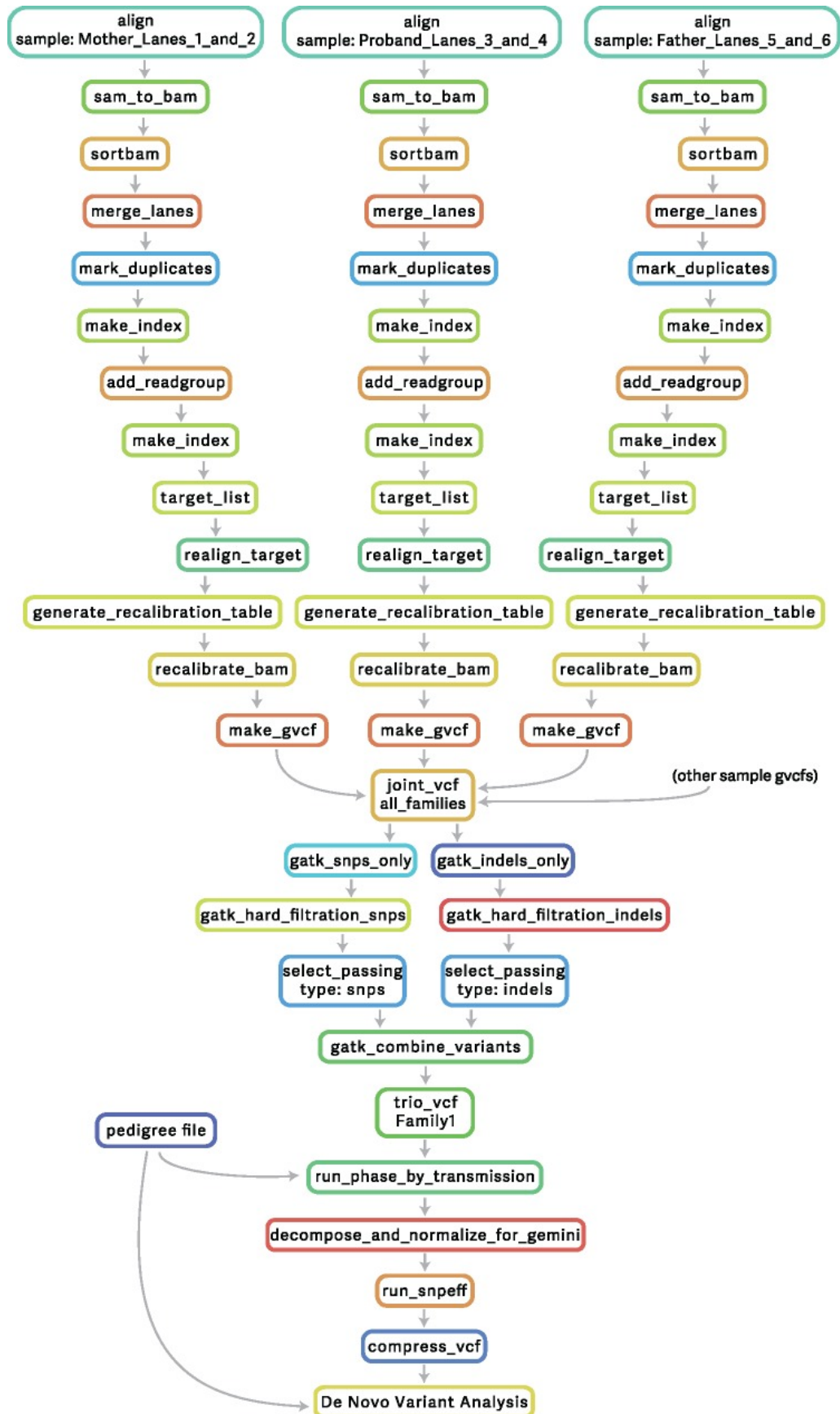
Figure 2.1: A directed acyclic graph depicting a trio analysis pipeline for detecting de novo mutations, reproduced from Leipzig (2016).

Bioinformatics pipelines consist of both parallel and serial steps, using varied software and data types as well as consisting complex dependencies, with both fixed and user-defined variables and parameters (Leipzig, 2016). Modern pipeline frameworks also support more advanced qualities, such as the ability to display the progress of the pipeline in real time, as well as creating environments to run the tool anywhere and providing support for work to be performed on distributed clusters or using cloud technologies (Leipzig, 2016).

2.2 FAIR Guiding Principles

Data in bioinformatics is being generated at a very fast rate; however this doesn't mean that we are gaining an increasing amount of knowledge along with this data (Wittig, Rey, Weidemann, & Müller, 2017). The main problem stems from the fact that there has been a failure in the storage and organization of this rapidly accumulating data, which should have been stored in "rigorous, principled ways," (Attwood et al., 2009) so that the required data can be found easily. This failure has led to exhausting, frustrating, stressful and also expensive experiences when trying to obtain the required data (Attwood et al., 2009).

Furthermore, there is a lack of effort made in scientific research to make results reproducible (ATCC, 2019). In scientific research, one of the fundamental principles is the independent verification of data. Researchers should be able to re-create experiments and arrive at the same conclusions, thus validating and strengthening the original work whilst allowing for further research based on those results (ATCC, 2019). However, often times scientific findings cannot be reproduced leading to a waste of time and resources (ATCC, 2019). This lack of reproducibility makes it difficult to reuse that data which, in turn, lowers the output of scientific research and slows down scientific progress (ATCC, 2019).

The FAIR principles are guidelines that ensure that data can be found and used by machines, in turn supporting data reuse by individuals. This way, more research of better quality can be generated by ensuring that data, algorithms and the tools and workflows that led to these data were designed to be findable, accessible, interoperable and reusable ("FAIR principles for data stewardship", 2016).

The goal that we would like to reach is the ease of data integration and reuse by the community after it is published. In order to achieve this goal it is important that data is managed properly (Wilkinson et al., 2016). Publishers, governmental agencies and funders have started requesting researchers to present "data management and stewardship plans" (Wilkinson et al., 2016). These plans are required to ensure that the assets provided by the research is properly managed and given long-term care - allowing it to be reused with ease in future research, either by

itself or in conjunction with newly generated data. Good data management and stewardship will then ensure high quality digital publications which will in turn further simplify future research (Wilkinson et al., 2016).

It is difficult, however to properly define what good data management is made up of, as there is no proper standard to ensure data is managed properly. As such, the FAIR guiding principles provide the research community with four foundational principles to improve data management and data stewardship that will in turn increase the value gained by digital publishing: (Wilkinson et al., 2016)

- Findability.

- Accessibility.

- Interoperability.

- Reusability.

These principles act as guidelines, and conditions for data sharing, encouraging researchers to ensure that the data being generated can be shared and reused from the outset (Boeckhout, Zielhuis, & Bredenoord, 2018). The FAIR principles highlight the importance of metadata and metadata standards with regards to data stewardship. Metadata is the "information and attributes applying to datasets and the data contained therein" (Boeckhout et al., 2018). These principles stress the need that metadata and metadata standards should be articulated and made readily and publicly available - allowing for further Reusability and reproducibility of data in the scientific field (Boeckhout et al., 2018).

Furthermore, these guidelines ensure that the metadata can be indexed, retrievable and analysed through the use of computers. Automation in data retrieval is important in research made on large-scale programs that use large amounts of data because it allows researchers to obtain data in an easier, more efficient manner. The FAIR principles also ensure that the terms and conditions under which the data can be shared and accessed are "explicit, well-defined and readily available" (Boeckhout et al., 2018). This includes any constraints in gaining or granting access to the data, privacy of the data, and publication rights. Although calls have been made in favour of open data, the FAIR principles instead offer a middle-ground - "the aim is to settle on legitimate and effective means of controlling access while facilitating bona fide research for all data" (Boeckhout et al., 2018).

The FAIR principles aim to achieve a step toward "machine-actionability - where a digital object provides detailed information to an autonomously-acting, computational data explorer" (Wilkinson et al., 2016). This would allow for the data to be obtained in a timely manner, as computers can retrieve data at a much faster rate than humans, while also maintaining some of the semantic importance

that is given by humans in the process of searching for data (Wilkinson et al., 2016).

It is important to note that these principles do not only apply to formal digital publishing and its data, but also to the tools, pipelines and algorithms that led to the creation of this data. The application of these principles to all aspects of research and data generation will be beneficial to all forms of digital research, whether it is data or analytical pipelines, since their availability can ensure transparency, Reusability and reproducibility (Wilkinson et al., 2016).

The FAIR guiding principles benefit a number of stakeholders by ensuring the Reusability of data, mainly researchers who can obtain and publish data in a timely manner, but also publishers, software developers who provide processing services and funding agencies (Wilkinson et al., 2016). However, these stakeholders have differing interests in the effect of FAIR, therefore any metrics that arise from these guidelines need to provide benefits for all the different stakeholders' interests (Wilkinson et al., 2018).

Wilkinson et al. state that a good metric should have the following properties: (Wilkinson et al., 2018)

- Clear: the purpose of the metric can be easily understood by anyone;

- Realistic: the metric should not be complicated for a resource to abide by;

- Discriminating: the metric should measure something that impacts FAIRness; evaluate to what degree that resource has achieved that metric; and provides instruction on how this value can be maximized;

- Measurable: the assessment of the metric can be made in an "objective, quantitative, machine-interpretable, scalable and reproducible manner, ensuring transparency of what is being measured and how.;"

- Universal: the metric must be applicable to all digital resources

If the aims of FAIR were to be realised and implemented, the end result would be more rigorous data management and data stewardship of digital resources, which would provide a benefit to the entire academic community (Wilkinson et al., 2016). To this end, a set of guidelines were created to aid researchers to create digital research artefacts that are more findable, accessible, interoperable and reusable. Although these guidelines and principles are not a standard or specification, they are a step forward in order to ensure FAIRness of data.

In order to fully understand the aims presented by FAIR, it is important to first understand what is truly meant by findability, accessibility, interoperability and reusability, as well as understanding these guidelines that ensure FAIRness of data.

2.2.1 Findability

Findability is the first principle in FAIR. The guidelines presented by the principle of findability focuses on the requirement that the data being generated is made findable - i.e. The data is easy to find. This is a crucial prerequisite to the other three principles presented by FAIR (Wise et al., 2019).

Wilkinson et al. (2016) describe findability as - "data should be identified using globally unique, resolvable and persistent identifiers, and should include machine-actionable contextual information that can be indexed to support human and machine discovery of that data."

To this end, the following set of guidelines were formally given to ensure findability of data:

- Data is properly described using rich metadata;

- Data and metadata are assigned a globally unique identifier;

- Metadata clearly and explicitly includes an identifier of the data it describes;

- Data and metadata are registered and/or indexed in a public searchable resource;

- Data, metadata and their identifiers should be re-findable at any point in time, thus should be persistent;

- Data should contain some basic machine actionable metadata which separates and distinguishes it from other data objects.

Findability of data can then be further enhanced by ensuring the use of standard identifiers and annotations, as well as through the use of standard ontologies and databases (Wittig et al., 2017). Further enhancement can be made through the use of a controlled vocabulary. This will ensure that information is unambiguous - use of identifiers based on standards ensures that certain abbreviations have only one meaning and ensures that there will be no variation in the schemas of metadata (Wittig et al., 2017; Wise et al., 2019). The result would be the production of models that might use different terminologies but can be used simultaneously, which will further build on the aim of ensuring findability of data. The key performance indicators that ensure findability of data include sufficient metadata that will allow users and machines to understand the data, additional documentation as well as the availability of uniform resource identifiers (URIs) or persistent identifiers (PIDs) (Wise et al., 2019).

2.2.2 Accessibility

The second principle in FAIR is that of accessibility. Accessibility addresses the necessity to make newly generated and previously existing data as well as digital assets more accessible (Wise et al., 2019). Accessibility ensures that metadata and data are understandable to both humans and machines as well as ensuring that the data is stored in a trusted repository (LiberEUROPE, 2017). Accessibility is based on three main components: (Wise et al., 2019)

- Access protocol - i.e. Information is provided on how to access the data;

- Access authorisation - i.e. Information is provided on what authorisation (if any) is required to access the data;

- Metadata longevity - i.e. Ensuring that the metadata provided will remain accessible, even if the data itself is not.

Usually, the data generated through research has a defined and finite lifespan, however, the metadata that corresponds with that data should ideally be stored permanently, so that the scientific record of the original data would be permanently stored (Wise et al., 2019). It is also important that the protocols used for access authorisation are made public, making it easy to determine who is authorised to gain access of the data (Wise et al., 2019).

It should be noted that data does not need to be made open to be FAIR. Accessibility dictates only the access and authorisation protocols be clearly defined and use open standards. Privacy requirements can still be observed (Wise et al., 2019).

To ensure accessibility of data, the following set of guidelines were formally presented: (Wilkinson et al., 2016)

- Data and metadata are retrievable by their identifier using a standardised communications protocol;

- The communications protocol used to retrieve data and metadata is open, free and universally implementable;

- The communications protocol allows for authentication and authorisation procedures, where necessary;

- The metadata related to the data remains accessible, even after the data becomes unavailable;

- Both humans and machines must be able to judge the actual accessibility of each data object.

2.2.3 Interoperability

The third principle in FAIR is that of interoperability. The principle of interoperability ensures that data being used can be integrated with other data, as well as being able to interoperate with other applications and workflows used for analysis, storage and processing of data. To ensure interoperability, the metadata should use a formal, accessible, shared and broadly applicable language for knowledge representation (LiberEUROPE, 2017).

The strategies used to achieve interoperability are essentially linked with the other FAIR principles. According to FAIR, data and metadata must be expressed in a formal, accessible, shareable and broadly applicable language (Wise et al., 2019). As such, observing standards when generating data will automatically drive data integration at all levels (Wise et al., 2019).

To achieve interoperability, the following guidelines were formally provided: (Wilkinson et al., 2016)

- Data and metadata use a formal, accessible, shared and broadly applicable language for knowledge representation;

- Data and metadata use vocabularies that follow FAIR principles;

- Data and metadata include qualified references to other data and other metadata;

- Data and metadata are machine-actionable;

- Data and metadata formats use shared ontologies;

- Data and metadata within the data object are both syntactically parseable as well as semantically machine-accessible.

The Interoperability of generated data can be further improved through the use of common exchange formats (Wittig et al., 2017). A widely-used example of such a format in systems biology is SBML - Systems Biology Markup Language (Hucka et al., 2003). Such formats would allow for the automatic and machine-readable exchange of data and facilitate the development of automatic data workflows between data management systems, applications and databases (Wittig et al., 2017).

2.2.4 Reusability

The final principle in FAIR is that of Reusability. The optimisation of reuse of data is the ultimate goal of FAIR. It is what separates traditional data management from FAIR data stewardship (Wise et al., 2019). Reusability ensures that data and collections have clear usage licenses and provide accurate information on provenance (LiberEUROPE, 2017).

Reusability is mainly achieved through well described data and metadata, allowing them to be replicated and combined in different settings (*FAIR Principles*, 2016).

It allows for the re-purposing of data for new needs, applications and user communities (Wise et al., 2019). Reusability allows data to "become more valuable to more people across large organisations, whether open-source communities or private organisations" (Wise et al., 2019).

To achieve this ultimate goal of Reusability, the following guidelines were formally presented: (Wilkinson et al., 2016)

- Data objects should be compliant with the first three principles of FAIR - i.e They must be findable, accessible and interoperable;

- Data and metadata are richly described with a plurality of accurate and relevant attributes;

- Data and metadata are released with a clear and accessible data usage license;

- Data and metadata are associated with detailed provenance;

- Data and metadata meet domain-relevant community standards;

- Data and metadata should be sufficiently described and rich enough that it can be linked, integrated automatically or with minimal human effort;

- Published data objects should refer to their sources with rich enough metadata and provenance to enable proper citation.

The reusability of data is further enhanced by providing proper descriptive information about its context. This includes any information about the original source where the data was obtained, as well as information about the procedures of how the data was generated. Further information about unique data attributes further enhances the reusability (Wittig et al., 2017). This enhanced reusability allows researchers to fully understand the algorithms, data used and the context with regards to an experiment as well as to fully understand the conclusions drawn from the study (Taylor et al., 2008).

2.3   Conclusion

In this chapter, research made on bioinformatics, the datasets and tools used in bioinformatics was reviewed in order to understand the background of the area that this study is being performed. The literature review documented the problems regarding the ever increasing amount of data being generated in the field of biology, as well as the organisation and management issues that arise from them from a bioinformatics standpoint.

It also reviews the aims in the FAIR principles, as well as the guidelines presented along with these principles in order to assure FAIRness of data. It can be noted that some of the advantages presented by the use of these guidelines would allow researchers to overcome the problems found in the processing of data in bioinformatics.

As such, this study focuses on finding a way to automate as much as possible a system that provides a FAIR score for tools and datasets, providing researchers a means to gauge the findability, accessibility, interoperability and reusability of their research.

Chapter 3

Specification and Design

3.1 Overview

The main goal for this study is to create a tool that semi-automates the FAIR assessment of bioinformatics tools and datasets, while allowing users to refine the score through a portal. The portal can also be used to allow the user to search for a particular tool or dataset, view that particular tool or dataset's performance in terms of a FAIR score as well as create pipelines of tools and datasets, automatically calculating the pipeline's FAIR score based on the tools and datasets used.

The metrics that will be used to calculate a FAIR score for tools were obtained from the FAIRshake tool rubric (*FAIRshake*, 2018). Some adjustments and further enhancements, were made to these metrics to be better suited for the tool's performance. Additional metrics were also added in order to ensure a more accurate FAIR score.

The metrics that will be used to calculate the FAIR score for datasets were obtained from the following sources:

- The FAIRshake dataset rubric (*FAIRshake*, 2018);

- The Australian Research Data Commons (ARDC) FAIR self-assessment tool (*nectar-rds / FAIR Assessment Tool*, 2018).

The sets of tools and datasets that will be used to evaluate the system are split as follows:

- Tools from the FAIRshake tool rubric assessments (*FAIRshake*, 2018) will be used to evaluate and match the results obtained by the tool using FAIRshake as a benchmark;

- Datasets from the FAIRshake dataset rubric assessments (*FAIRshake*, 2018) will be used to evaluate and match the results of the tool using FAIRshake as a benchmark;

- Tools and datasets that will be used as case studies obtained from the European Bioinformatics Institute services (*European Bioinformatics Institute*, 2019);

- Pipeline as presented by (Barber, Xu, Pérez-Losada, Jara, & Crandall, 2012) will be used as an additional case study.

3.2   FAIR Guidelines

The principles presented by FAIR are meant to be applied generally to all facets of research and data generation. This of course includes ensuring FAIRness of software artefacts. However, since the guidelines presented by (Wilkinson et al., 2016) *Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing version b1.0*, 2016, are meant to be broad so that they can be applied to different fields as needed, they cannot be used as metrics to calculate a FAIR score in their proposed form (Dumontier et al., 2018).

The main focus of this study is to create a tool that semi-automates the FAIR assessment of tools and datasets. Therefore, these guidelines need to be further expanded to create metrics that can be used to properly assess software and datasets respectively.

As such, in order to find out how the FAIR principles can be properly applied to software the following set of questions must first be answered (Dumontier et al., 2018):

- Which FAIR principles can be directly adopted?

- Which FAIR principles can be adapted for software, and how?

- Are there any guidelines that do not apply to software, and can thus be dismissed?

- What should be added to the principles in order for them to apply for software?

3.2.1 Findability Guidelines

In order to apply the guidelines provided by FAIR to ensure findability in software and datasets, certain adjustments might need to be made to them. Furthermore certain guidelines need to be further defined to be applied for software (Dumontier et al., 2018).

Taking a look at the guidelines presented by FAIR to ensure findability, the following changes and further definitions can be made for the guidelines to account for software artefacts and datasets (Dumontier et al., 2018):

- **Data is assigned a globally unique and persistent identifier.**
  This is one of the guidelines presented by the principle of findability. When applied to certain types of data, such as research papers or books, this can refer to a DOI or an ISBN. However, software does not contain this type of public identifier.

  Tools in bioinformatics, however, are always presented along with a research paper describing the tool's capabilities, features, tests and experiments performed with it. This paper contains its own unique and persistent DOI. As such, for the purpose of this study, the DOI of the paper which presented the tool will be used as its identifier.

- **Data is described with rich metadata.**
  The tool needs to be described properly in terms of its features and capabilities. Although these will be listed and explained in the paper that presents the tool, it is important that they can be easily found by the user.

- **Data is registered or indexed in a searchable resource.**
  By uploading datasets, tools and their metadata to a website, this requirement can be fulfilled through the use of a search engine. This will allow them to be found easily by searching for certain terms that can match either the dataset or tool's name or a tool's capabilities.

  By storing the metadata on the website, a researcher would also be fulfilling the previous requirement.

  Furthermore, only by fulfilling this guideline can the assessment of the tool's accessibility, interoperability and reusability be performed. This guideline ensures that the tool is findable in the most trivial sense. If the tool is not findable at all, then it's accessibility, interoperability and reusability cannot be assessed.

3.2.2   Accessibility Guidelines

Similar to the guidelines to ensure findability, the guidelines presented for the principle of accessibility also needed to be changed or defined further in order to be applied for software.

The following changes were made to ensure that the guidelines for accessibility can account for software and datasets: (Dumontier et al., 2018)

- **Metadata and data are retrievable by their identifier using a standardised communications protocol.**
  Both metadata and data should be retrievable in some manner. Given that the DOI, the metadata of the tool or dataset, in this case represented by the research paper, should be retrievable via HTTP or HTTPS. Furthermore, since the tool or dataset will be hosted on a website, this website should also be accessible and retrievable via HTTP and HTTPS.

- **The metadata related to the data remains accessible, even after the data becomes unavailable.**
  This guideline ties in with one of the guidelines presented for findability. It is important that the DOI of the research paper presenting the tool or dataset is findable due to the fact that, if for some reason, the website containing the data is down, there will be no access for the metadata related to the tool.

  Furthermore, the DOI allows a tool's metadata to be preserved in the case that, for some reason, the tool becomes outdated and as such is no longer supported.

- **Information should be provided on how a tool can be accessed.**
  This is the most important aspect of ensuring the accessibility of a tool. Since the guidelines themselves are intended to be generic so that they can be applied to different aspects of research, there is no specific guideline on how to handle accessibility of software. As such, this is a proposed added guideline to aid in the assessment of software to ensure the generation a proper FAIR score.

  Through this guideline, the metadata related to the tool should be enhanced slightly to include information that would not be included otherwise in other types of data, as for such types it would be irrelevant. The metadata should include with it information on how to access the tool, whether it is through API usage or a sequence of operations performed through a command line interface.

3.2.3 Interoperability Guidelines

The guidelines presented by FAIR to ensure interoperability are mainly focused on ensuring that metadata is properly understandable and shareable. However some of these guidelines are not applicable with regards to tools and datasets.

As such some of these guidelines must be further defined, as well as adding an additional guideline as can be seen below (Dumontier et al., 2018):

- **Metadata and data use a formal, accessible, shared and broadly applicable language for knowledge representation.**
  The metadata related to the tool or dataset must be expressed in a way that can be properly understood and properly follow rules of some knowledge representation language. This applies to the data of a dataset itself. However, in the case of tools, the data refers to the source code of the tool, which cannot be assessed by such a guideline. An argument can be made, however, that following proper coding practices and obeying syntactic and semantic rules of programming languages used can server as an equivalent for a knowledge representation language. (Dumontier et al., 2018)

- **Data and metadata use vocabularies that follow FAIR principles**
  This guideline states that the chosen language used to represent the metadata and data within the datasets should be openly accessible and follow FAIR principles. In the case of software, since the data is the source code of the tool, this guideline does not apply.

- **Data and metadata include qualified references to other data or metadata.**
  As a minimum requirement to fulfil this guideline, the software should use well-known vocabularies to express what inputs and what outputs are accepted.

  Furthermore, proper referencing must be made with regards to what other tools or research was used in order to create the tool or dataset.

As can be seen above, the guidelines to ensure interoperability are not applicable when the data to be assessed is a software artefact. As such, the following guideline is proposed so that the interoperability of software can be properly assessed:

- **Metadata must properly indicate how the data, in this case the software, interacts with the system it is being used on.**
  This guideline states that software metadata can be further defined to also include how the software interacts with the system it is being executed on. This includes information on compatibility with different operating systems, as well as availability of source code.

3.2.4 Reusability Guidelines

Given that reusability is the ultimate goal of FAIR, it is important to ensure that the guidelines presented by this principle are further defined to be applicable for software.

As such, the following definitions were provided to ensure that software's reusability can be properly assessed (Dumontier et al., 2018):

- **Data and metadata are richly described with multiple accurate and relevant attributes.**
  Similar to the guidelines presented by findability, ensuring that the tools and datasets are properly described through their metadata will ensure that the data can be findable through their metadata. Furthermore, being provided with the proper metadata will ensure that the tools and datasets can be re-used in the future, as it will allow the user to understand them and how to use them.

- **Data and Metadata are released with a clear and accessible data usage license.**
  It is important to make sure that the license of the metadata is properly understood. Furthermore you must make sure that you have a license for your software.

  The license for both the data and metadata must both be properly presented so that the evaluator can properly assess the tool.

- **Data and metadata are associated with detailed provenance.**
  Information on when the software was first created, along with the first experiments and tests that were executed on it should be made available. Furthermore, although the metadata - which will be preserved in the research paper - should not change, if amendments are made to the paper, the original paper should still be made available along with the amendments.

Although these guidelines cover some aspects of a software's reusability, in order to ensure that the software's reusability is properly assessed, the following guideline is proposed:

- **Data and metadata are available in a public resource.**
  This guideline, similar to the guideline provided by the principle of findability, states that the data and metadata can both be found publicly. With the metadata being made available in a public resource, the information is available for when the software is to be re-used. Moreover, if the data is stored in a public repository, it will automatically ensure that the data can be found and be made readily available for re-use.

3.3 FAIR Metrics - Tools

After further defining the FAIR guidelines to ensure that they can be applied to software, the metrics that will be used within our system to evaluate the FAIR score of bioinformatics tools will be defined. Most of these metrics are based on research previously made by FAIRshake (2018) with a few more metrics being proposed to ensure the generation of a more accurate FAIR score.

3.3.1 Tool Findability Metrics

For the purpose of this study, the metrics provided by the FAIRshake (2018) tool rubric proved to be enough to cover the findability score of a tool, along with an additional metric to better adhere to the adjusted guidelines as proposed by (Dumontier et al., 2018).

Furthermore, certain metrics were given a higher priority to ensure that tools that are wholly findable get a better score to properly represent their better Findability when compared to other tools.

The metrics used in the system to assess findability are as follows:

- **Tool is findable on a website.**
  This is a metric that must be satisfied. If a website is not found for the bioinformatics tool, the FAIR assessment tool will not be able to automate the assessment. Satisfying this metric ensures that the tool adheres to the guideline, i.e., data is registered in a publicly searchable resource (Wilkinson et al., 2016).

- **Tool is freely downloadable.**
  This metric is given a high priority, since if a tool is freely downloadable, its findability is greatly increased.

- **A proper description is provided for the tool.**
  This metric is given a medium priority. Satisfying this metric ensures that the tool adheres to the guideline, i.e., data is described by rich metadata (Wilkinson et al., 2016).

- **Previous versions of the tool are made available.**
  This metric is given a low priority. Satisfying this metric ensures that the tool adheres to the guideline, i.e., data should be re-findable at any point in time and thus should be persistent (Wilkinson et al., 2016).

- **Tool has a unique identifier.**
  This metric is given a medium priority. This metric states that the tool must

have a unique identifier, in this case a DOI linked to the paper that presented the tool. Satisfying this metric ensures that the tool adheres to the guideline, i.e., data is assigned a globally unique identifier (Wilkinson et al., 2016).

3.3.2   Tool Accessibility Metrics

The metrics provided by the FAIRshake (2018) tool rubric do not cover much in terms of accessibility, providing only one metric which needed to be adjusted slightly in order to be automated. Together with this metric another metric is proposed to ensure a proper assessment of the tool's accessibility.

The metrics used in the system to assess accessibility are as follows:

- **Tool can be programmatically accessed through an API.**
  This metric is given a medium priority. Satisfying this metric ensures that the tool adheres to the proposed guideline, i.e., information should be provided on how the tool can be accessed.

- **Tool can be accessed through a set of commands executed in a command line interface.**
  This metric is given a high priority. Satisfying this metric also ensures that the tool adheres to the proposed guideline. However, it is given a higher priority as if it can be accessed through a command line interface then it can be accessed by any system.

3.3.3   Tool Interoperability Metrics

The FAIRshake (2018) tool rubric does not provide any metrics for interoperability. As such, in order to be able to assess the interoperability of the tools, the following two metrics are proposed:

- **Tool compatibility information is provided.**
  This metric is given medium priority. This metric states that proper information is provided on which operating systems this tool is compatible with. Satisfying this metric ensures that the tool adheres to the proposed guideline. Metadata must properly indicate how the tool interacts with different operating systems. The assessment of this tool is split into 3 parts:

  - Windows compatibility,
  - Linux compatibility, and
  - Mac compatibility.

- **Tool source code is provided.**
  This metric is given medium priority. This metric assumes that if a tool's source code is provided then it can be executed provided that the system it is executed on has a source-compatible compiler for the programming language in which the tool was written. Satisfying this metric ensures that the tool adheres to the proposed guideline that metadata must properly indicate how the tool interacts with different operating systems.

3.3.4 Tool Reusability Metrics

For the purpose of this study, the metrics provided the by the FAIRshake (2018) tool rubric prove to be enough to provide a proper assessment of a tool's reusability score. As such, the same metrics were adopted to be used for the system.
The metrics used in the system to assess reusability are as follows:

- **Tool is hosted in a public repository.**
  This metric is given a high priority. This metric states that the tool must be hosted on a public repository such as GitHub. Satisfying this metric ensures that the tool adheres to the proposed guideline that data is available in a public resource.

- **Tool uses community accepted ontologies.**
  This metric is given medium priority. This metric states that if the tool uses one or more ontologies, these ontologies must be stated and must adhere to some standard. If the tool does not use an ontology, this metric will not apply and as such will not affect its FAIR score. Satisfying this metric ensures that the tool adheres to the guideline, i.e., data and metadata meet domain-relevant community standards (Wilkinson et al., 2016).

- **Proper documentation of the tool is provided.**
  This metric is given medium priority. This metric states that proper documentation on how to use the tool must be provided. Satisfying this metric ensures that the tool adheres to the guideline, i.e., data and metadata are richly described with a plurality of accurate and relevant attributes (Wilkinson et al., 2016).

- **Developer contact information is provided.**
  This metric is given low priority. This metric increases the tool's reusability in such a way that if an issue arises during use of the tool, one can easily contact the developer with their issues in order to have them resolved.

- **Information on how to cite the tool is provided.**
  This metric is given low priority. Satisfying this metric ensures that the tool adheres to the guideline, i.e., published data objects should refer to their sources with rich enough metadata and provenance to enable proper citation (Wilkinson et al., 2016).

3.4 FAIR Metrics - Datasets

The aim of this study is to create a tool that automates the FAIR assessment for tools and datasets. After defining the metrics that will be used to assess the FAIR score of tools, the metrics that will be used to assess datasets can be defined. Most of these metrics are adopted directly from, or based on the research previously made by FAIRshake (2018) and the nectar-rds FAIR Assessment Tool (2018).

3.4.1 Dataset Findability Metrics

For the purpose of this study, the metrics provided by the FAIRshake (2018) dataset rubric prove to be enough to provide a proper assessment of a dataset's findability score. As such, the same metrics were adopted to be used for the system.
The metrics used in the system assess findability are as follows:

- **Dataset is findable on a website.**
  Similar to the metric used to find tools, this is the most important metric as if the dataset fails it, the FAIR assessment tool will not be able to automate the assessment.

- **Dataset has a proper description.**
  This metric is given medium priority. Satisfying this metric ensures that the dataset adheres to the guideline, i.e., data is described by rich metadata (Wilkinson et al., 2016).

- **Dataset has a unique identifier.**
  This metric is given low priority. This metric states that the dataset must have a unique identifier, in this case a DOI linked to the paper that originally presented the dataset. Satisfying this metric ensures that the dataset adheres to the guideline, i.e., data is assigned a globally unique identifier (Wilkinson et al., 2016).

3.4.2 Dataset Accessibility Metrics

For the purpose of this study, metrics were adapted from both the FAIRshake (2018) dataset rubric, as well as from the nectar-rds FAIR Assessment Tool (2018). The metrics used in the system to assess accessibility are as follows:

- **Dataset is freely downloadable from the website.**
  This metric is given a high priority. If the dataset can be freely downloaded, its accessibility is improved greatly.

- **Metadata will still be available even if data is no longer available.**
  This metric is given a low priority. Assuming that the dataset has a unique identifer linking it to the paper that originally presented the dataset, then this metric is automatically satisfied.
  Satisfying this metric ensures that the dataset adheres to the guideline, i.e., the metadata related to the data remains accessible, even after the data becomes unavailable (Wilkinson et al., 2016).

3.4.3 Dataset Interoperability Metrics

The FAIRshake (2018) dataset rubric does not provide any metrics for the analysis of interoperability of datasets. The nectar-rds FAIR Assessment Tool (2018) provides one metric which can be directly adopted into the system.
As such, the metric used in the system to assess interoperability is as follows:

- **Information is provided on which format(s) or file format(s) the dataset is available in.**
  This metric is given medium priority. In the case of datasets, interoperability is improved when the dataset comes in formats that can be used in conjunction with multiple varying tools and datasets. As such, having the information on which formats the dataset is available in will greatly improve the dataset's interoperability.
  Satisfying this metric ensures that the dataset adheres to the guideline, i.e., data and metadata use a formal, accessible, shared and broadly applicable language for knowledge representation (Wilkinson et al., 2016).

3.4.4 Dataset Reusability Metrics

For the purpose of this study, the metrics provided by the FAIRshake (2018) dataset rubric are be enough to provide a proper assessment of a dataset's reusability. As such, the same metrics were adopted for the system.

The metrics used in the system to assess the reusability are as follows:

- **Information is provided on how to cite the dataset.**
  This metric is given low priority. Satisfying this metric ensures that the tool adheres to the guideline, i.e., published data objects should refer to their sources with rich enough metadata and provenance to enable proper citation (Wilkinson et al., 2016).

- **Contact information is provided for the creator(s) of the dataset.**
  This metric is given low priority. The ability to contact the creator or creators of the dataset in case something is unclear or some further information is need improves the reusability of the dataset.

- **Previous versions of the dataset are made available.**
  This metric is given a low priority. Satisfying this metric ensures that the dataset adheres to the guideline, i.e., data should be re-findable at any point in time and thus should be persistent (Wilkinson et al., 2016).

3.5 Gathering of Data

The main function of the system is that given a name of a dataset or a tool, search the web for information about the tool or dataset and try to gather as much information from the metadata found in order to generate a FAIR score for that tool or dataset. This search can be further narrowed down if the URL itself is given. Furthermore, the system should be able to calculate the FAIR score of a pipeline that is defined by the user. In order to be able to test such a function, multiple sets of data were gathered to ensure that the evaluation of the system was done to verify and validate the system's capabilities.

To evaluate the system, the sets of data needed to include sets of tools and datasets to evaluate the system's ability to generate FAIR scores. The set of tools consists of:

- Tools obtained from the FAIRshake (2018) tool rubric assessment.
  This is a set of tools on which assessment was performed by the FAIRshake (2018) tool. This assessment will serve as a benchmark against which the system's score will be compared. Although some of the metrics may be different and there may be some new metrics involved, the system should be able to provide a similar FAIR score to that of the assessment made by the FAIRshake (2018) tool.

- Tools obtained from the EMBL-EBI tool service (*European Bioinformatics Institute*, 2019).

  The EMBL-EBI is a service that hosts some of the most commonly used tools (*European Bioinformatics Institute*, 2019). Given that these tools are popular and used often, it is expected that these tools should generate a favourable FAIR score. Some of the tools will be assessed by the system, and then a manual assessment will also be performed on them. These will serve as case studies to ensure that the tool performs as expected.

- Datasets obtained from the FAIRshake (2018) dataset rubric assessment.

  Similar to the set of tools that was obtained from the tool rubric assessment, this is a set of datasets on which an assessment was performed using the FAIRshake (2018) dataset rubric. The assessment will serve as a benchmark against which the system's score will be compared. Similarly, although some of the metrics may be different, the system is expected to provide a similar FAIR score to that of the assessment made by the FAIRshake (2018) dataset rubric assessment.

- Datasets obtained from the EMBL-EBI dataset service (*European Bioinformatics Institute*, 2019).

  The EMBL-EBI also hosts some commonly used datasets (*European Bioinformatics Institute*, 2019). An assumption is made where given that the hosted datasets are popular and used quite often, these datasets should perform well when assessed by the system, generating a favourable FAIR score. Therefore, a small set of these datasets will be assessed by the system, and then be manually assessed. Similarly to the tools, these will serve as case studies to ensure that the tool performs as expected.

An extra set of data must be collected in the FAIR assessment of tools. As stated in one of the metrics used to assess a tool's Reusability, a tool must use community accepted ontologies. A list of community accepted ontologies has been compiled. For the purpose of this study, this list of ontologies is compiled from the OBO Foundry. The OBO Foundry provides a table of used ontologies and their status, whether they are active, inactive or obsolete (Smith et al., 2007). For the purpose of this study, the list of active ontologies will be used as the set of community accepted ontologies.

Finally, the system's capability to generate a FAIR score for a pipeline that will be built by the user must be evaluated. In order to evaluate this function, a sample pipeline was obtained from research performed by (Barber et al., 2012). The tools and datasets in the pipeline presented by (Barber et al., 2012) will be passed through the system, having a FAIR score generated, which will then be

double-checked by performing a manual assessment. A FAIR score for the pipeline will then be calculated from the FAIR scores of these tools and datasets.
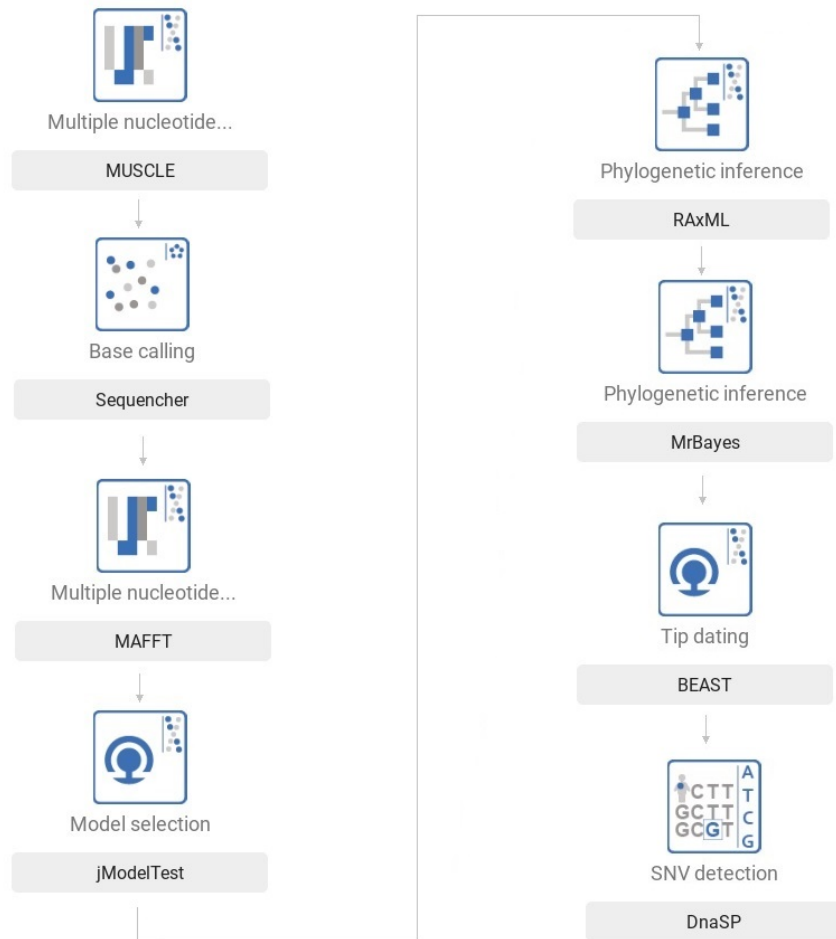
Figure 3.1: A visual representation of the pipeline presented by Barber et al. (2012), reproduced from *Conflicting Evolutionary Patterns Due to Mitochondrial Introgression and Multilocus Phylogeography of the Patagonian Freshwater Crab Aegla neuquensis* (2015)

Chapter 4

Implementation

4.1 System Architecture

This chapter will discuss the computation involved in the implementation of the system, as well as the technologies used to implement it. In this section, a basic overview of the system architecture will be given.

The system is built using Python and its main supporting library Django (2018), a web framework for Python. This allowed for the creation of the FAIR assessment tool and back-end processing of the portal using Python and other supporting libraries. The portal was then built using the various functions provided by Django to design the views and templates while using an HTML front-end.

The FAIR assessment tool was built using web crawling techniques provided by the Selenium package for Python to obtain information about the tools and datasets from online sources, along with additional Python logic to calculate scores for findability, accessibility, interoperability and reusability. The scores obtained from the FAIR assessment tool are then stored in a MySQL database. This database holds the FAIR scores of the various tools and datasets in the system along with some additional information about them. When the user checks for further information about the tool, along with which areas need information to be able to obtain a better FAIR score, summary information about the tool is provided, such as the tool's description, where to download the tool and the DOI of the paper where the tool was presented.

The portal itself acts as a front-end for the user to be able to see what tools and datasets are in the system and their FAIR scores. The user can then see the full details on a tool or dataset, allowing them to find out what can be added to or changed in the tool or dataset's website in order to increase that tool or dataset's FAIR score. The user can then manually refine the score by providing any information which the automated assessment tool might have missed. An

example of this relates to the second metric provided by this study for reusability: a tool must use a community accepted ontology. This metric would need to be refined manually, as there is no way for the automated tool to tell whether a tool uses an ontology or not. As such, it is automatically assumed that the tool uses one, but it was not listed. The user is then told that if the tool uses an ontology, listing which ontologies are used would increase its reusability, but if it does not use an ontology, then the user can refine the score by marking the metric as not applicable to the tool. Once the user finishes adjusting the information, the FAIR score is re-calculated in the back end.

The following sections will discuss what technologies were used, how the FAIR assessment tool was implemented and the additional data to aid in calculating the FAIR assessment was obtained.

4.2 Technologies Used

In this section the main technologies used in the study along with the supporting libraries that were used in order to implement the solution will be discussed.

4.2.1 Development Software

The system was written in Python 3.7 (2018). A portal was then created using Django 2.1 (2018) which is a web framework for Python 3.7 (2018). Thanks to Python's ability to quickly integrate systems together, the FAIR assessment tool was then integrated into the portal.

The IDE used to develop the system was Visual Studio Code (2015). A MySQL database was used which was hosted on an Apache (2018) server started through XAMPP (2017). This allowed for the use of phpMyAdmin 4.8.2 (2018) to manage the database.

4.2.2 Supporting Libraries

In order to implement the system, this study made use of some freely available Python libraries to allow for rapid development. To install these libraries, pip was used, which is a package-management system used to install and manage software packages written in python.

Django 2.1 (2018)	High-level Python Web framework. Provides necessary tools for web development that were used to develop the Portal.
mysqlclient (2018)	Library that allows Python to interact with a MySQL database.
Google API Python Client (2018)	Provides the necessary tools to allow Python to interact with tools provided by Google API. Used to implement website search through Google Custom Search (2006).
Selenium Python (2018)	Provides the necessary tools to perform web crawling using Python. To use this library, the proper driver had to be downloaded to allow Selenium to interface with the browser. In the study the application interacts with Firefox. As such, the driver that had to be installed was geckodriver (2018).
PyYAML (2018)	Library that allows Python to parse YAML files. Used to parse the list of active ontologies obtained from the OBO foundry (Smith et al., 2007), which was obtained in a YAML format.
Crossref API Python Client (2018)	Provides functions that allow Python to iterate through the Crossref API. Used in order to obtain DOI of papers that presented the tools and datasets.

Table 4.1: List of supporting libraries.

4.3 FAIR Assessment Tool

In this section the implementation of the assessment tool will be discussed. The process begins by having the user input the name of the tool or dataset to be assessed. Using the Google Custom Search API, a list of links is found using the name of the tool or dataset as a search phrase. For the purpose of this study, this list is limited to a maximum of 20 links. These links are then filtered to attempt to find the official website for the tool or dataset to begin crawling for information. This is done by comparing the data found in the website with a list of terms that are related to and commonly used in bioinformatics listed in listing 4.1.

```
phrases = ["protein", "dna", "bioinformatics", "genome",
        "nucleotide", "biological", "biotechnology",
        "alignment", "amino acid", "autoradiography",
        "autosomal", "blotting", "blots", "cell",
        "computational biology", "sequencing",
        "similarity", "cluster", "cytoplasm", "enzyme",
        "gene", "genomics", "immunoglobuline",
        "lead", "mitosis"]
```

Listing 4.1: List of most commonly used phrases in bioinformatics.

Once a website is found, the link will be used by Selenium to obtain information and perform the assessment. A slight drawback in the implementation of the system is that a website can be found that would not be the official website of the tool or dataset, however it would contain one or more phrases from the list of phrases most commonly used in bioinformatics and would be used to try and obtain information. The information gathered would not be relevant to the tool or dataset on which the assessment would be performed, resulting in an incorrect assessment.

Given this drawback, Selenium proves to be the perfect tool to obtain information in order to perform the assessment. Although there are more efficient tools for web crawling, Selenium allows you to see which websites the program is accessing and which links it is following. This ensures that the website being crawled is the correct one. When integrated with the portal, the user is given the option to also provide the link for the tool's official website, further ensuring that the assessment is performed correctly.

4.3.1   Tool FAIR Assessment

This section will consider the computation involved in performing a FAIR assessment for a tool. This starts by considering the most important metric presented in this study for Findability, i.e., the tool is findable on a website. If no website is found after filtering the links found from the Google Custom Search API, then the assessment will automatically fail, as there is no way to gather information to automate the assessment.

Findability

The assessment tool then goes on to search for information based on the second metric, that the tool is freely downloadable. This is implemented by first checking if the website is hosted on a GitHub repository, in which case it is automatically assumed that the tool is freely downloadable and marked as such. Otherwise, the website is searched for a link marked "Download". If such a link is found, then the score is given, otherwise it is given a zero score. This metric was given a high priority equal to 40% of the total findability score.

The tool then searches for information for the third metric, i.e., a proper description is provided for the tool. This is implemented by first searching for a link that would redirect the tool to a page providing information about the tool. If no such link is found, the tool then searches for areas on the page marked either as "Introduction", "About" or "Description". If no such area is found, the tool then searches by XPath to see if an area in the HTML would be marked as "desc". If the information required is found, the description of the tool is saved into the database
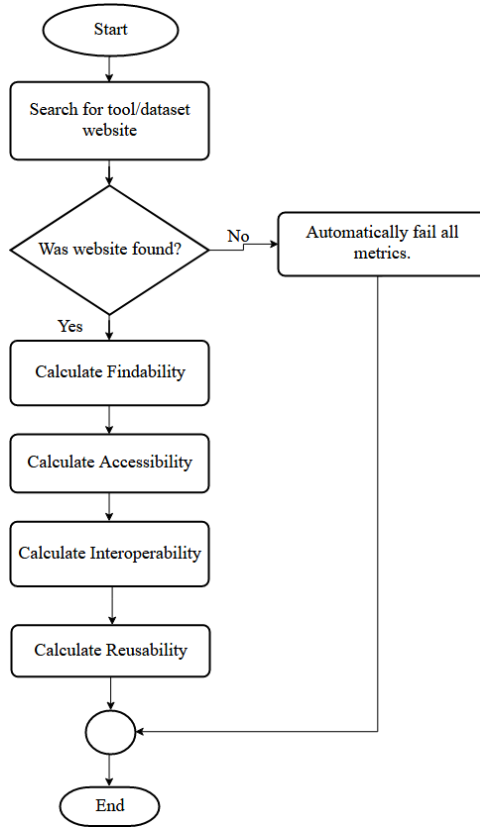
Figure 4.1: Flowchart showing start of FAIR assessment.
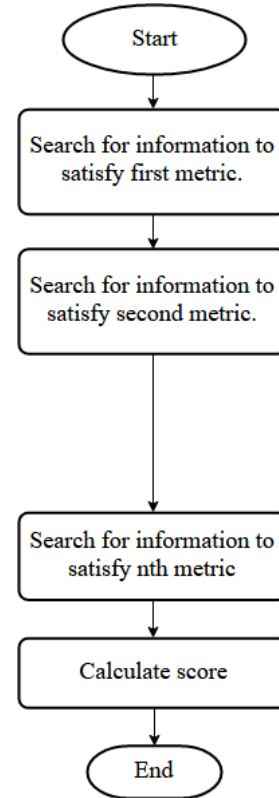
Figure 4.2: Sample flowchart of an assessment.

and the score for this metric is given. This metric was given a medium priority, equal to 25% of the total findability score.

The assessment tool then searches for information based on the fourth metric presented by the study, i.e., previous versions of the tool are made available. This is implemented by searching the page for links or areas labelled as either "Older versions" or "Previous versions". If such an area is found then the tool is given points for this metric. If the driver is in a GitHub link, however, it is assumed that previous versions can be accessed within the repository, therefore a full score would be awarded. This metric carries 10% of the total findability score, as it was given a low priority.

The assessment tool then goes on to search for information based on the final metric presented in this study, that the tool has a unique identifier. This is implemented by first searching the page for a link containing a DOI, which could be used as a direct link to the paper that presented that tool. If such a link is not found, the tool interfaces with Crossref using the Crossref API to search for the paper that presented the tool by searching for the tool's name. If a paper is found, the DOI is directly extracted from the paper and saved into the database, awarding points to the tool for the metric. This metric was given a medium priority and a

score of 25% of the total findability score is assigned.

```
tool_doi = item['DOI']
```

<div align="center">Listing 4.2: Extracting DOI from paper.</div>

Accessibility

After calculating the tool's findability, the assessment tool starts crawling the website for information to calculate its accessibility. It first begins by searching for information to cover whether the tool can be programmatically accessed through an API. This is done by first searching for a link marked with the text "API", which would link you to a page providing information on how to access the tool from an API. If such a link is not found, the page is instead checked for areas marked with the text "API". If such a link is found then the score for this metric is awarded. Since this metric is marked as medium priority, it carries 50% of the total accessibility score.

The assessment tool then searches for information based on the second metric, i.e., the tool can be accessed through a set of commands executed in a command line interface. This is implemented by searching through the web page for areas containing the text "command line" or "CLI". If the required information is found then the points for this metric are awarded. This metric is marked as medium priority and carries 50% of the total accessibility score.

Interoperability

The following step in the FAIRscore calculation process calculates the tool's interoperability. This starts by considering the first metric presented by this study, whether the tool's compatibility information is provided. As a whole, this metric was given medium priority and carries 50% of the total interoperability score. This metric is split into 3 parts: Windows compatibility, Linux compatibility and Mac compatibility. These are implemented as follows:

- The assessment tool searches for areas on the website marked with the text "Windows" or "windows". If the text is found, then $\frac{1}{3}$ of the 50% carried by this metric is awarded.

- The assessment tool then searches for areas on the website marked with the text "Unix" or "UNIX". If the text is found, then $\frac{1}{3}$ of the 50% carried by this metric is awarded.

- Finally the assessment tool searches for areas on the website marked with the text "Mac". If this text is found, then $\frac{1}{3}$ of the 50% carried by this metric is awarded.

If all three areas are found, then the full 50% are awarded, otherwise the corresponding fraction of the 50% is awarded based on which OS are supported.

The assessment tool then searches for information to cover the second metric presented by this study for interoperability, that the tool's source code is provided. This is implemented in different ways. If the website is hosted on a GitHub repository, then it is automatically assumed that the source code is provided, awarding full points for this metric. Otherwise, the assessment tool searches for a link marked as source code, which would be assumed to be a download link for the source code. If the link is not found, then the website is searched for an area marked as "source code", assuming that information on how to obtain the source code is provided. If this area is found, then the points for this metric are awarded. This metric is given medium priority, and carries 50% of the total Interoperability score.

Reusability

Finally, the assessment tool searches for information to calculate the tool's reusability. This starts by searching for information on whether tool is hosted in a public repository. This is implemented in different ways. If the website is hosted on a GitHub repository, then it is already in a public repository. As such, full points would be awarded for this metric. Otherwise, the assessment tool would search the website for a link containing link text "www.github.com". If such a link is found, then it is assumed the tool is hosted in a public GitHub repository and full points would be awarded. This metric carries 40% of the total reusability score as it is a high priority metric. However, if the second metric is marked as not applicable for this tool, then this metric will carry 50% of the total reusability score.

The assessment tool then searches for information based on the second metric of whether the tool uses community accepted ontologies. This is done by having the tool search the web page for text containing "ontology" and comparing it to a list of active ontologies obtained from the OBO foundry (Smith et al., 2007). These ontologies are then saved into the database and full points are awarded for this metric. If the ontologies found do not match any of those found on the OBO foundry (Smith et al., 2007), then half the points are awarded. This is because although the tool does not use community accepted ontologies, the developers still provided information on which ontologies were used. This metric is given medium priority and carries 20% of the total reusability score. Since this metric will not apply to every tool, it can be marked as so during the user-refinement stage. If this metric is marked as inapplicable for the tool, then it will affect the percentage carried by the other metrics and have its marks redistributed to the other metrics based on the priority they are given.

The assessment tool will then search for information on whether the tool provides proper documentation. If the tool is hosted on a GitHub repository, then it will be assumed that documentation is provided within the repository, and full points will be awarded for this metric. Otherwise, the assessment tool will search

the web page for a link marked "Documentation", which is assumed to link to a page containing the documentation of the tool. If a link is not found, then the page will be searched for areas marked as "documentation", which if found is assumed as areas containing documentation for the tool. This metric is given medium priority. If the second metric is marked as not applicable by this tool, then this metric carries 25% of the total reusability score, otherwise it carries 20% of the total score.

The assessment tool then goes on to search for whether the developer's contact information is provided. This is implemented by first searching the web page for a link marked "Contact". If a link is not found, then the assessment tool will search the page for an area by XPath containing the text "email". If this link or area is found, then it is assumed that the developer's email is given, which would satisfy this metric. This metric is given a low priority. If the second metric is marked as not applicable for this tool, then this metric carries 12.5% of the total reusability score, otherwise it carries 10% of the total score.

Finally the assessment tool will search for whether information on how to cite this tool is provided. This is done by having the tool search for an area in the web page by XPath containing the text "Citing" or "References". If this area is found, it is assumed that information on how to cite or reference the tool is provided. This metric is given a low priority. If the second metric is marked as not applicable, then this metric carries 12.5% of the total reusability score. Otherwise, if the second metric is applicable to this tool, then it carries 10% of the total score.

Figure 4.3 shows a flowchart of the calculation of a tool's Reusability
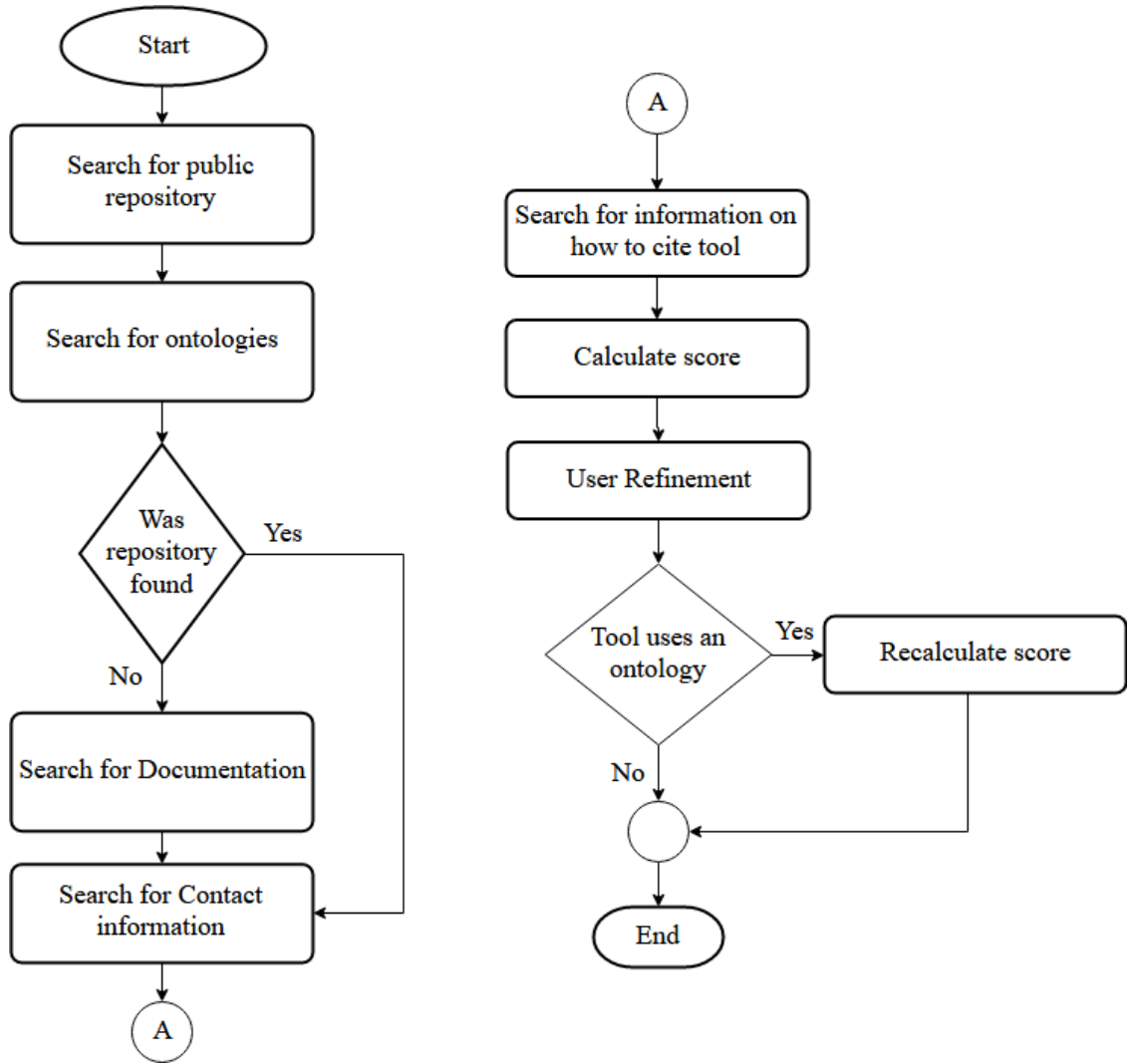
Figure 4.3: Flowchart showing calculation of Reusability.

4.3.2   Dataset FAIR Assessment

This section will consider the computation involved in performing a FAIR assessment for a dataset. This starts by considering whether a dataset is findable on a website. If the website for the dataset is not found after filtering the links found from the Google Custom Search API, then the assessment will automatically fail, as there will be no way to gather information to automate the assessment.

Findability

The assessment tool then searches for information based on the second metric presented by this study for findability, that the dataset has a proper description. This is implemented by first searching for a link marked as "About", with the assumption that it would redirect you to a page providing information about the dataset. If no such link is found, the tool then searches for areas on the page named either as "Introduction", "About" or "Description". If no such area is found it then searches using XPath to see if an area in the HTML would be marked as "desc".

If the information required is found, the description of the dataset is saved into the database and the points for this metric are awarded. This metric was given a medium priority. It carries 60% of the dataset's total findability score.

The assessment tool then goes on to search for information based on the second metric presented in this study, i.e., dataset has a unique identifier. This is implemented by first searching the page for a link containing a DOI, which could be used as a direct link to the paper that presented that dataset. If a link is not found, the tool interfaces with Crossref using the Crossref API to search for the paper that presented the dataset by searching for the dataset's name.

```
w2 = works.query(title=dataset_name).filter
        (type=''journal-article").sort("relevance")
```

Listing 4.3: Searching Crossref API for paper presenting the dataset.

If a paper is found, the DOI is directly extracted from the paper and saved into the database, awarding points to the dataset for the metric. This metric was given a medium priority, equal to 40% of the dataset's total findability score.

```
dataset_doi = item['DOI']
```

Listing 4.4: Extracting DOI from paper.

Accessibility

After calculating the dataset's findability, the assessment tool gathers information to calculate its accessibility. This process first begins by searching for information based on the first metric presented by this study, that the dataset is freely downloadable from the website. This is implemented by having the tool search the website for a link marked as "Download". If a link is found, it is assumed that it will take you to a download page where the dataset can be downloaded. This metric is given a high priority, equal to 80% of the dataset's total accessibility score.

The second metric presented by the study is that metadata will still be available even if data is no longer available. This is covered by the previous process to search for the paper presenting this dataset. If a paper is found, it is automatically assumed that the metadata will be permanently available in that paper, awarding full points to the dataset for this metric. This metric carries 20% of the total accessibility score as it is given a low priority.

Interoperability

The assessment tool then searches through the website for information to calculate the dataset's interoperability. This starts by considering the metric presented by this study to calculate the dataset's interoperability, information is provided on what format(s) or file format(s) the dataset is available in. This is implemented by searching the web page for areas containing the text "file format" or "format". If

this area is found, it is assumed that this area provides the necessary information to fulfil this metric. This metric is given medium priority, however it is the sole metric that is used to calculate interoperability and carries 100% of the interoperability score.

Reusability

Finally the assessment tool calculates the dataset's reusability. This starts by searching for information based on the first metric presented by this study, i.e., information is provided on how to cite the dataset. This is implemented by searching for an area in the web page by XPath containing the text "Cite", "Citation", "Citing" or "References". If this area is found, it is assumed that this area provides information on how to cite or reference the dataset. This metric is given a low priority and carries 33.3% of the total reusability score.

The assessment tool then searches for information based on the second metric presented by this study for reusability, that contact information is provided for the creator(s) of the dataset. This is implemented by first searching the web page for a link marked as "Contact". If the link is not found, then the assessment tool will search the page for an area by XPath containing the text "email". If an email link or area is found, then it is assumed that the email of the creator of the dataset is given. As such the criteria for this metric would be satisfied and the points for this metric would be awarded. This metric is given a low priority and carries 33.3% of the total reusability score.

The assessment tool then goes on to search for information based on the final metric presented in this study for reusability, i.e., previous versions of the dataset are made available. This is implemented by searching through the web page for a link marked as "Previous versions" or "Older versions". If a link is found, it is assumed that previous versions of the dataset are available, fulfilling this metric. This metric carries 33.3% of the dataset's total reusability score as it is given a low priority.

4.3.3 Pipeline FAIR Assessment

After calculating the FAIR score for the various tools and datasets, the user can then create a pipeline out of them. The pipeline itself can have a FAIR score which is calculated based on the FAIR score of the various tools and datasets in it.

Pipeline Findability:
This is calculated by adding the findability of the various tools and datasets and dividing them by the total number of tools and datasets in the pipeline.

$$\frac{Findability_{T1} + Findability_{D1} + ... + Findability_{Tn} + Findability_{Dn}}{N_T + N_D} \tag{4.1}$$

Pipeline Accessibility:
This is calculated by adding the accessibility of the various tools and datasets and dividing them by the total number of tools and datasets in the pipeline.

$$\frac{Accessibility_{T1} + Accessibility_{D1} + ... + Accessibility_{Tn} + Accessibility_{Dn}}{N_T + N_D} \tag{4.2}$$

Pipeline Interoperability:
This is calculated by adding the interoperability of the various tools and datasets and dividing them by the total number of tools and datasets in the pipeline.

$$\frac{Interoperability_{T1} + Interoperability_{D1} + ... + Interoperability_{Tn} + Interoperability_{Dn}}{N_T + N_D}$$
$$\tag{4.3}$$

Pipeline Reusability:
This is calculated by adding the reusability of the various tools and datasets and dividing them by the total number of tools and datasets in the pipeline.

$$\frac{Reusability_{T1} + Reusability_{D1} + ... + Reusability_{Tn} + Reusability_{Dn}}{N_T + N_D} \tag{4.4}$$

4.4 Obtaining Data

Along with the sets of tools and datasets required to test the system, some other sets of data were first gathered in order for the system to be implemented. The first set of data was the list of most commonly used phrases in bioinformatics. The data in this set was taken from the Bioinformatics glossary presented by Biosynthesis (*Bio-Informatics Services*, 2010). Not all the terms presented in this glossary were

used. Instead, the list was created using the terms presented in the glossary which were most commonly used in websites of tools and datasets.

The second set of data that needed to be obtained was the list of active ontologies. This list would be used to be able to fulfil the metric presented for reusability of tools, i.e., the tool uses a community accepted ontology. In this case, the active tools would serve as community accepted ontologies. This list of ontologies was obtained from the OBO Foundry (Smith et al., 2007) in a YAML format. The following code was executed in order to obtain this list of ontologies.

```python
active_ontologies = []
with open("ontologies.yml", 'r') as ont:
    content = yaml.load(ont)
    for item in content.items():
        x = item[1]
        for i in range(len(x)):
            if x[i]["activity_status"] == "active":
                active_ontologies.append
                (x[i]["title"].lower())
```

Listing 4.5: Obtaining list of ontologies from YAML file.

The following is a list of some of the active ontologies used in this study:

- Xenopus Anatomy Ontology;

- Zebrafish anatomy and development ontology;

- Human Disease Ontology;

- Gene Ontology;

- Basic Formal Ontology

- Evidence Ontology;

- Drug-drug Interactioon and drug Interaction Evidence Ontology;

- Human phenotype ontology.

A complete list of all the active ontologies used in this study can be found in Appendix A.

Chapter 5

Testing and Evaluation

In this chapter, the testing of the system will be discussed, and the results obtained from the system will be evaluated.

5.1 Specification

The system presented in this study is aimed at providing users with a way to obtain an assessment of a tool's findability, accessibility, interoperability and reusability. As such, the tests performed on this system are meant to achieve two main goals.

Firstly, the tests performed on the system are aimed to answer one of the research questions presented in the first chapter of this study: Can the FAIR assessment be part of an automated process? In order to be able to answer this question the tests were performed in the following manner. The assessment tool is first applied to a list of tools and datasets which were assessed by the FAIRshake tool assessment and the FAIRshake dataset assessment respectively. The results obtained were then compared with those found in the assessments made by FAIRshake to ensure that the automated assessment tool was able to find the information required to fulfil the metrics. The assessment tool will then be applied to a list of tools hosted by EMBL EBI (*European Bioinformatics Institute*, 2019), which will first have to be manually assessed. This manual assessment gives us a list of results that should be expected from the assessment tool, to ensure that it is working as expected. This is further assured through the use of Selenium (2018), which allows us to see which areas of a tool's website the tool is checking in order to obtain the required information.

The second goal of these tests is to ensure that the assessment provided by this study is justifiable and truly FAIR compliant. This is done by comparing the values of the results obtained from the automated assessment tool with the results presented by the FAIRshake tool and dataset assessments. A difference between the total scores of findability, accessibility, interoperability and reusability

44

is expected due to the fact that different metrics are given different priorities in this study, whereas in the FAIRshake assessments, each metric is given the same priority. However, although this difference in totals is expected, this difference is not expected to be significant.

5.2 Results

In this section, the results from the tests performed on each set of tools and datasets will be discussed. Since the system presented in this study uses a priority system to distinguish the tools that are truly findable from those that fulfil only some minor criteria, the total result of each metric of findability, accessibility, interoperability and reusability will not be evaluated. The focus will instead be on the individual metrics, comparing how the tool compares to the results from the FAIRshake tool and dataset assessments (*FAIRshake*, 2018) to see if the objective presented by this study holds (i.e. to automate as much of the FAIR assessment as possible).

5.2.1 FAIRshake Tools

The evaluation considers the tools which were already assessed by the FAIRshake tool assessment (*FAIRshake*, 2018) and compares the results obtained from the tool with the results from FAIRshake. It should be noted that the results given for Reusability have been refined to indicate whether a tool uses an ontology or not. This is to ensure that the provided FAIR score is not negatively affected by a metric that would otherwise not be applicable to the tool.

The evaluation considers the total scores of findability, accessibility, interoperability and reusability of the FAIRshake assessment and the results obtained from the automated assessment tool. This is done to ensure that although metrics are given different priorities, there is not a significant difference between the score obtained by the assessment tool and those given by FAIRshake. By checking which scores have significant differences and checking the results obtained in the metrics presented by that principle, it is possible to check where differences in the results rose from.

- Findability:

  - The main reason for the differences in findability scores was that the automated assessment tool was unable to find a DOI linking to the paper that presented the tool, as can be seen in the table 5.1 for BLAST, HAMMOCK, PyroHMMvar, CytoSPADE, MEANS, poretools, miRPathDB and Breakpointer. This resulted in lower scores, but when the information is provided, the score is more accurate and richer.

Tool	FAIRshake				Automated Tool			
	F	**A**	**I**	**R**	**F**	**A**	**I**	**R**
BLAST	75%	100%	N/A	75%	65%	100%	50%	50%
HAMMOCK	100%	100%	N/A	40%	75%	50%	50%	75%
PyroHMMvar	100%	0%	N/A	40%	75%	50%	50%	100%
CytoSPADE	75%	0%	N/A	20%	100%	0%	50%	75%
MEANS	100%	0%	N/A	60%	75%	50%	100%	75%
poretools	100%	0%	N/A	50%	75%	0%	100%	75%
miRPathDB	75%	0%	N/A	80%	65%	0%	0%	50%
CHRONOS	36%	100%	N/A	90%	90%	50%	83.33%	100%
PhosphoPICK	75%	0%	N/A	40%	90%	0%	0%	25%
Breakpointer	100%	0%	N/A	80%	75%	50%	66.67%	87.5%

Table 5.1: Comparison of results between FAIRshake assessment on tools and assessment made by the automated assessment tool.

- Accessibility:

  - The main reason for a difference in accessibility scores was the fact that another metric was presented in this study, which is not provided by FAIRshake, as can be seen in table 5.1 for HAMMOCK, PyroHMMvar, CHRONOS and Breakpointer. This resulted in lower scores when the automated tool found the same information as FAIRshake but did not find information to satisfy the metrics presented in this study as can be seen for HAMMOCK and CHRONOS. On the other hand, this resulted in a higher score when the FAIRshake metric was not satisfied but the automated tool found information to satisfy the metric presented by this study as can be seen for PyroHMMvar and Breakpointer.

- Reusability:

  - One of the reasons for a difference in reusability scores was the fact that this study uses different priorities for its metrics, so although it was able to find the same information as FAIRshake, the final result was different. Another reason why there was a difference was that for HAMMOCK, PyroHMMvar and CytoSPADE, a public repository was added after the assessment was performed by FAIRshake. Furthermore, although the difference may seem very large, this can be explained by the fact that the first metric of a tool's Reusability, i.e., the tool is hosted in a public

Tool	Manual Assessment			
	F	A	I	R
BLAST	90%	100%	50%	50%
HAMMOCK	100%	100%	50%	75%
PyroHMMvar	100%	100%	50%	100%
CytoSPADE	100%	0%	50%	75%
MEANS	100%	100%	100%	75%
poretools	100%	50%	100%	75%
miRPathDB	90%	50%	0%	50%
CHRONOS	90%	50%	83.33%	100%
PhosphoPICK	90%	0%	0%	25%
Breakpointer	100%	50%	66.67%	87.5%

Table 5.2: Manual assessment on tools from FAIRshake tool assessment.

repository, was given a high priority. This metric carries 50% of the total reusability score, which is the reason for a seemingly large difference in scores in the case of BLAST, miRPathDB and PhosphoPICK, as can be seen in table 5.1.

5.2.2 EMBL EBI Tools

Moving on to the tools hosted by the EMBL EBI (*European Bioinformatics Institute*, 2019), the evaluation can now compare the results obtained from a manual assessment with those obtained from the automated assessment tool. The manual assessment was performed using all the metrics presented by this study, including the metrics for accessibility and interoperability that were not provided by FAIRshake and providing a score based on their priority. Similar to the previous set of tools, the results given for reusability have been adjusted to indicate whether a tool uses an ontology or not.

By comparing the results obtained, we can see whether the automated assessment tool is working as expected. In cases of differences we can see that differences can be observed in findability scores. All of these occurrences were due to the fact that the automated assessment tool was unable to find a DOI linking to the paper that presented the tool, as can be seen in table 5.2 for Clustal Omega, HMMER, FASTA and InterProScan. The paper presenting the tools were found through extensive searches through literature in order to find the papers that presented the tools. After the assessment is performed and a score is provided, the user will be

Tool	Manual Assessment				Automated Tool			
	F	**A**	**I**	**R**	**F**	**A**	**I**	**R**
Clustal Omega	90%	100%	100%	100%	65%	100%	100%	100%
HMMER	100%	50%	50%	87.5%	75%	50%	50%	87.5%
FASTA	100%	50%	100%	100%	75%	50%	100%	100%
Phobius	90%	0%	0%	50%	90%	0%	0%	50%
InterProScan	100%	50%	66.67%	87.5%	75%	50%	66.67%	87.5%

Table 5.3: Comparison of results between a manual assessment performed on EMBL EBI tools and assessment performed by the automated assessment tool.

able to refine the score and provide any missing information.

5.2.3 FAIRshake Datasets

We can then see and compare the results from assessments made on datasets by FAIRshake and compare them with the results obtained from the automated assessment tool.

Dataset	FAIRshake				Automated Tool			
	F	**A**	**I**	**R**	**F**	**A**	**I**	**R**
A443654 KINOMEscan	50%	50%	100%	67%	60%	80%	100%	66.67%
(R)-Roscovitine KINOMEscan	50%	50%	100%	67%	60%	80%	100%	66.67%
GTEx Portal Datasets	50%	50%	100%	100%	60%	80%	100%	100%
WormBase Datasets	100%	100%	100%	100%	100%	100%	100%	100%
Gene Ontology Consortium Datasets	50%	50%	100%	67%	60%	80%	100%	66.67%
Rat Genome Database Datasets	50%	50%	100%	67%	60%	80%	100%	66.67%

Table 5.4: Comparison of results between FAIRshake assessment on datasets and assessment performed by the automated assessment tool.

The differences that can be observed in these results arise from the fact that the assessment tool gives different priorities to different metrics. However, all the information obtained by the automated tool for assessment is the same as that for FAIRshake.

5.2.4 EMBL EBI Datasets

Moving on to the datasets hosted by EMBL EBI (*European Bioinformatics Institute*, 2019). The following table compares the results obtained through a manual assessment with those obtained from the automated assessment tool. The manual assessment was performed using the metrics presented by this study and providing a score based on their priority.

Dataset	Manual Assessment				Automated Tool			
	F	A	I	R	F	A	I	R
ArrayExpress	100%	100%	100%	100%	100%	100%	100%	100%
BioModels	60%	80%	100%	66.67%	60%	80%	100%	66.67%
InterPro	100%	100%	100%	100%	100%	100%	100%	100%
Pfam	100%	100%	100%	100%	100%	100%	100%	100%

Table 5.5: Comparison of results between manual assessment performed on EMBL EBI datasets and assessment performed by the automated assessment tool.

The results show that the tool managed to obtain all the data necessary for an assessment as would be expected in a real-world scenario.

5.2.5 Pipeline Tools

Finally, we can compare the results obtained from manually assessing the tools in the pipeline presented by Barber et al. (2012) with the results obtained from the automated assessment tool. The manual assessment was performed using the metrics presented by this study and providing scores based on their priority.

A difference in Findability can be observed in table 5.5 for MrBayes, as the automated assessment tool was unable to find a DOI linking to the paper that presented the tool. The total findability, accessibility, interoperability and reusability of the pipeline were then calculated by the system.

The difference in Findability observed in table 5.5 for MrBayes affected the total Findability of the pipeline, resulting in a 3.12% loss in findability.

	Manual Assessment				Automated Tool			
Tool	F	A	I	R	F	A	I	R
MUSCLE	100%	50%	100%	100%	100%	50%	100%	100%
Sequencher	65%	100%	33.33%	12.5%	65%	100%	33.33%	12.5%
MAFFT	90%	50%	100%	75%	90%	50%	100%	75%
jModelTest	100%	50%	100%	100%	100%	50%	100%	100%
RAxML	100%	50%	100%	100%	100%	50%	100%	100%
MrBayes	100%	50%	100%	87.5%	75%	50%	100%	87.5%
BEAST	100%	100%	100%	100%	100%	100%	100%	100%
DnaSP	90%	100%	66.67%	37.5%	90%	100%	66.67%	37.5%

Table 5.6: Comparison of results between manual assessment performed on tools in pipeline presented by Barber et al. (2012) and assessment performed by the automated assessment tool.

	Expected	Actual
Findability	93.12%	90%
Accessibility	68.75%	68.75%
Interoperability	87.5%	87.5%
Reusability	76.56%	76.56%

Table 5.7: Comparison of pipeline expected FAIR score against actual FAIR score calculated by system.

5.3 Evaluation of Results

In this section, the results obtained will be further evaluated, by taking into consideration the two main goals which these tests aimed to achieve.

5.3.1 Justifiable FAIR Score

The first consideration is given to the secondary goal of these tests, i.e., are the metrics used for this study justifiable and truly FAIR?

In order to ensure that the metrics used in this study and the priority system used are truly FAIR, we must compare the total result obtained from the automated assessment tool with that provided by FAIRshake. Taking a look at table 5.1, we can see that although there are some differences in the results, even though both results have the same amount of passing metrics, the difference is not too large. For example, in the case of miRPathDB, both results have passed the same amount

of metrics, however the metric that failed is given more priority in this study. This difference in priority has not led to a significant difference between the results.

Furthermore, this goal is further achieved through the additional metrics provided in order to ensure a more accurate FAIR score. As can be seen in table 5.1, some of the accessibility scores are very different. This is due to the fact that FAIRshake provides only one metric to calculate a tool's accessibility. In this study, a second metric was provided, in order to allow for a more refined and a more accurate FAIR score. This is also shown in the introduction of two metrics for the Interoperability of tools. This provides a more accurate and refined FAIR score over that of FAIRshake, which do not consider the Interoperability of tools altogether.

A similar observation can be made when looking at the results obtained by the automated assessment tool in the case of datasets. The results of the assessment tool are not too different from those presented by the FAIRshake data assessment, even though the metrics are given a different priority. In the case of datasets however, there are still very few metrics implemented to calculate their FAIR score.

5.3.2 Answering the Hypothesis

The main goal of the tests run on this system was to try and answer one of the research questions presented in this study, i.e., can the FAIR assessment be automated?

Taking into consideration the first set of tools which the assessment was performed on, a few differences can be noted in some of the results. As can be seen in table 5.1, quite a few results were different. However some difference was expected due to the fact that the assessment tool presented in this study uses a different priority system. This is the case for the majority of the differences in accessibility and reusability. In the cases for accessibility, there are two metrics presented by this study, which results in a lower FAIR score if only the metric presented by FAIRshake applies. Furthermore, the difference in priorities for metrics presented by reusability account for the difference in almost all the tools, with the exception of HAMMOCK, PyroHMMvar and CytoSPADE, where a public repository was added after the assessment was performed by FAIRshake. On the other hand, a difference can be noted in the Findability of a majority of the tools as shown in table 5.1. Further inspection showed that the differences all rose from the same metric, i.e., the tool has a unique identifier. This can be seen for BLAST and MEANS in table 5.8 below.

A set of tests were then performed on tools hosted by EMBL EBI (*European Bioinformatics Institute*, 2019). These tests were performed using all the metrics

BLAST	FAIRshake assessment	Automated Assessment
Findability	**75%**	**65%**
Tool is freely downloadable.	25%	40%
Tool has a proper description.	25%	25%
Previous versions of tool are available.	0%	0%
Tool has a unique identifier.	25%	0%
MEANS		
Findability	**100%**	**75%**
Tool is freely downloadable.	25%	40%
Tool has a proper description.	25%	25%
Previous versions of tool are available.	25%	10%
Tool has a unique identifier.	25%	0%

Table 5.8: Comparison of Findability metrics for BLAST and MEANS.

provided by this study, taking into consideration the developed priority system as well as the metrics which were not considered by FAIRshake. As can be seen in table 5.2, due to these considerations the differences that were found in accessibility, interoperability and reusability were no longer present. This further shows that the differences that arose in these areas were all from the different weightings given in the system presented by this study. The differences in the scores of findability, however, were still present. Further inspection of the results showed that almost every tool, with the exception of Phobius, failed the same metric as the first set of tools, i.e, the tool has a unique identifier.

From the manual assessment of tools, some difficulty was found in locating the papers that presented the tools. This was because the websites do not follow a "pattern" that would allow for easily findable information. Some of the required information was scattered around the web page, which made it very difficult to find. This was then further shown through the automated assessment tool in its inability to find the required unique identifier for the majority of the tools.

On the other hand, the results obtained for datasets contain very few differences. As can be seen in table 5.3, most of the differences between the total results are small, all of which could be attributed to the fact that the system presented in this study uses a different priority system than that of FAIRshake. A comparison of metrics for A443654KINOMEscan is shown in table 5.9 below. The results of the second set of datasets on which the automated assessment tool was applied further indicate this trend. The results obtained from the manual assessment and

the automated assessment tool were the same, which means that for the metrics presented for datasets, the FAIR assessment can be sufficiently automated.

A443654KINOMEscan	FAIRshake assessment	Automated Assessment
Findability	**50%**	**60%**
Information about dataset is provided.	100%	60%
Dataset has a unique identifier.	0%	0%
Accessibility	**50%**	**80%**
Dataset is freely downloadable.	50%	80%
Information about dataset will persist after dataset becomes unavailable.	0%	0%
Interoperability	**100%**	**100%**
Information is given on which file formats the dataset is available in.	100%	100%
Reusability	**67%**	**66.67%**
Information on how to cite dataset is provided.	33%	33.33%
Dataset developer's contact information is provided.	0%	0%
Previous versions of dataset are available.	34%	33.33%

Table 5.9: Comparison of FAIR metrics for A443654KINOMEscan.

Finally, the results for the set of tools in the pipelines show that the same issue in the first two sets of tools is also present in this context. This issue of not being able to find a DOI linking to the paper that presented the tool, however, is not as apparent as in the other two sets, as it only affected one of the tools in this set, i.e., MrBayes. Similar to the tools in the other sets, it was very difficult to find the DOI of the paper presenting this tool, which in turn resulted in the automated assessment tool being unable to find it. In this scenario, it did not have a large effect on the pipeline's total findability score, as it was only 3.12% lower than the expected score. However, the majority of the tools which were tested suffered from this issue. If a pipeline in bioinformatics follows the same pattern, then the system will provide a poor findability score, even though a unique identifier exists. The DOI can be given manually by a user at a later stage.

In conclusion, in the case of the datasets, the tests proved that the information required is quite easily findable, resulting in a proper FAIR score as expected. Therefore in the case of datasets, the hypothesis of whether or not the assessment can be automated passes. In the case of tools, however, the assessment might not be fully automate-able. Firstly, the second metric presented for reusability, i.e., the tool uses a community accepted ontology, cannot be automated at all, as there is no way to tell when a tool doesn't use ontologies altogether. In such cases, it is up to the user to refine the result to provide a more accurate FAIR score. Furthermore, a common problem was found in calculating the findability for tools in that the unique identifier is not always made findable. In these cases, the findability score would be lower than it should be simply because the information is not findable by automated means.

Chapter 6

Future Work

Future works in this study concern further improvements being made to the assessment metrics, specifically those for tool accessibility and dataset metrics. In its current state, there are very few metrics that relate to a dataset's findability, accessibility, interoperability and reusability, and very little research available on what metrics can be used in order to calculate such a score. An improvement can be made by increasing these metrics in order to obtain a more accurate FAIR score.

Furthermore, in its current state, the system is slow at calculating a tool's or dataset's FAIR score. In the case where a large pipeline needs to be assessed, it can take quite some time to calculate the FAIR score of each individual tool and dataset. As such, an improvement can be made by implementing multi-threading techniques in the assessment in order to speed up the process.

Another improvement that can be made that would improve the system's efficiency would be to use a different technology for web crawling. For the purpose of this study, Selenium (2018) was used as it allowed us to better visualize the web scraping process. However, this slows down the FAIR assessment quite significantly. Using a different system to for web crawling will therefore improve the speed at which information is obtained from the web pages, leading to the faster calculation of a tool or dataset's FAIR score.

Chapter 7

Conclusions

Data in the biological field will still continue to be generated at a very fast rate, along with tools that will aid in the processing of this data. It is important to ensure that the data being generated is reproducible, to allow researchers to reproduce the results obtained to strengthen their own findings (ATCC, 2019) This increase in data will continue to make it difficult for researchers to find the resources they need for their research which will slow down scientific progress. Studies have presented guidelines and metrics in order to evaluate a tool's Findability, Accessibility, Interoperability and Reusability in order to help users find the appropriate resource for their task.

This study aims to provide researchers with a searchable archive of tools and datasets, as well as provide a means to automate the calculation of FAIR scores for these resources. These aims were achieved through the use of web crawling techniques which were used to obtain the information required to calculate a FAIR score in an automated way. Furthermore, the use of web frameworks and database management systems allowed us to create a portal which acts as a front-end to the user, allowing them to search for the required resource by its features or capability and see the respective FAIR scores. This will allow the user to see what areas in Findability, Accessibility, Interoperability or Reusability the tool is lacking in, and what is required in order to improve that resource's FAIR score.

Sets of tools and datasets obtained from different sources were used to test the system. A set of tools and datasets obtained from the FAIRshake tool and dataset assessments were used to ensure that the FAIR assessment can be automated, by comparing the results obtained from the tool with those obtained from FAIRshake. The automation of the assessment was further shown through manual assessments of a set of tools and datasets obtained from EMBL EBI (*European Bioinformatics Institute*, 2019) and comparing the results from the manual assessments with those obtained from the automated tool.

The evaluation provided further insight and answers to the proposed specifi-

cations. By comparing the results of the individual metrics from the tools and datasets obtained from FAIRshake with those obtained from the system presented by this study, it can be observed that the system was able to obtain the information necessary in order to fulfil the metrics presented and provide a proper FAIR score. These observations were then further proven through the manual assessments of the tools and datasets obtained from EMBL EBI (*European Bioinformatics Institute*, 2019) and the tools used in the pipeline presented by Barber et al. (2012). The manual assessments were performed using the metrics presented in this study, taking into considerations the additional metrics that were not considered by FAIRshake, as well as the different priority system used in this study. The results obtained show that the tool was able to obtain the same results as those obtained by the manual assessments, proving that the tool performs as expected, with a few exceptions in the case of tools where the tool struggled to find a tool's unique identifier.

It can be concluded that the aims and objectives presented by this study have been achieved. The results of the individual metrics obtained from testing are consistent with those obtained from FAIRshake, showing that it is possible for the FAIR assessment to be part of an automated process. The portal then allows users to search for their required tool and provides them with a FAIR score that can aid them to decide on which tool they should choose.

Appendix A

List of Active Ontologies

- Plant Ontology;

- Xenopus Anatomy Ontology;

- Zebrafish anatomy and development ontology;

- Ontology for Biomedical Investigations;

- Gene Ontology;

- Basic Formal Ontology;

- Phenotype And Trait Ontology;

- Human Disease Ontology;

- Chemical Entities of Biological Interest;

- PRotein Ontology (PRO);

- Ontology of Vaccine Adverse Events;

- Fungal gross anatomy;

- Gazetter;

- Ontology for MIRNA Target;

- Hymenoptera Anatomy Ontology;

- Information Artifact Ontology;

- Rat Strain Ontology;

- Foundational Model of Anatomy Ontology (subset)

- RNA ontology;

- eagle-i resource ontology;

- Genotype Ontology;

- Evidence ontology;

- Social Insect Behaviour Ontology;

- Ontology of Biological and Clinical Statistics;

- Tick Anatomy Ontology;

- Ontology for General Medical Science;

- Agronomy Ontology;

- Drug-drug Interaction and drug Interaction Evidence Ontology;

- Cardiovascular Disease Ontology;

- Biological Imaging Methods Ontology;

- Mouse pathology ontology;

- Pathogen Transmission Ontology;

- Cell Ontology;

- Unified phenotype ontology (uPheno);

- Xenopus Phenotype Ontology;

- Kinetic Simulation Algorithm Ontology;

- Mammalian Phenotype

- Experimental condition ontology;

- Exposure ontology;

- Geographical Entity Ontology;

- Units of measurement ontology;

- FOODON;

- Pathway ontology;

- Homology Ontology;

- Mammalian Feeding Muscle Ontology;

- Spider Ontology;

- Medaka Developmental Stages;

- FlyBase Controlled Vocabulary;

- Planarian Phenotype Ontology;

- Software ontology;

- Informed Consent Ontology;

- Drosophila gross anatomy;

- Common Anatomy Reference Ontology;

- Systems Biology Ontology;

- Platynereis Developmental Stages;

- Ontology of Genetic Susceptibility Factor;

- Ontology of core ecological entities;

- Ontology of Biological Attributes;

- Beta Cell Genomics Ontology;

- MHC Restriction Ontology;

- Ontology of Medically Related Social Entities;

- Genomic Epidemiology Ontology;

- The Prescription of Drugs Ontology;

- Ontology for Biobanking;

- The Oral Health and Disease Ontology;

- Drosophila Phenotype Ontology;

- Clinical measurement ontology;

- human phenotype ontology;

- Pathogen Host Interaction Phenotype Ontology;

- Sequence types and features ontology; Ontology of Microbial Phenotypes;

- Mental Disease Ontology;

- Ontology of Precision Medicine and Investigation;

- EuPathDB ontology;

- NCBI organismal classification;

- Mouse adult gross anatomy;

- The Data Use Ontology;

- C. elegans Gross Anatomy Ontology;

- Protein Modification;

- Flora Phenotype Ontology;

- Relation Ontology;

- Dictyostelium discoideum anatomy;

- Minimum PDDI Information Ontology;

- Mass spectrometry ontology;

- Confidence Information Ontology;

- Non-Coding RNA Ontology;

- Ontology of Adverse Events;

- Emotion Ontology;

- The Statistical Methods Ontology;

- The Drug Ontology;

- Vertebrate trait ontology;

- Mathematical modelling ontology;

- Molecular Interactions Controlled Vocabulary;

- Biological Collections Ontology;

- MIAPA Ontology;

- Mental Functioning Ontology;

- Zebrafish developmental stages ontology;

- Microbial Conditions Ontology;

- Anatomical Entity Ontology;

- Infectious Disease Ontology;

- Mouse gross anatomy and development, timed;

- Vertebrate Taxonomy Ontology;

- Ontology for Parasite LifeCycle;

- Ontology of Host-Microbiome Interactions;

- Mouse Developmental Stages;

- Fission Yeast Phenotype Ontology;

- Zebrafish Phenotype Ontology;

- The Drug-Drug Interactions Ontology;

- Comparative Data Analysis Ontology;

- Human Developmental Stages;

- Biological Spatial Ontology;

- Plant Experimental Conditions Ontology;

- Ontology of Arthropod Circulatory Systems;

- Vaccine Ontology;

- Monarch Disease Ontology;

- Asocmycete phenotype ontology;

- Contributor Role Ontology;

- microRNA Ontology;

- Chemical Information Ontology;

- Plant Phenology Ontology;

- Chemical Methods Ontology;

- Drosophila development;

- Ontology of Prokaryotic Phenotypic and Metabolic Characters;

- Name Reaction Ontology;

- Interaction Network Ontology;

- Ontologized MIABIS;

- Obstetric and Neonatal Ontology;

- Symptom Ontology;

- Mosquito insecticide resistance;

- C. elegans phenotype;

- Uberon multi-species anatomy ontology;

- Measurement method ontology;

- Taxonomic rank vocabulary;

- Cell Line Ontology;

- Ctenophore Ontology;

- Teleost taxanomy ontology;

- Neuro Behaviour Ontology;

- Malaria Ontology;

- NCI Thesaurus OBO Edition;

- Porifera Ontology;

- Antibiotic Resistance Ontolgoy;

- Environment Ontology;

- planaria-ontology;

- Plant Trait Ontology;

- Dictyostelium discoideum phenotype ontology;

- Population and Community ontology;

- Molecular Process Ontology;

- Zebrafish Experimental Conditions Ontology;

- Scientific Evidence and Provenace Information Ontology;

- The Ontolgoy of Genes and Genomes;

- Human Ancestry Ontolgoy;

- Mosquito grass anatomy ontology;

- BRENDA tissue/enzyme source.

References

Apache. (2018). Retrieved from `https://www.apache.org/`

Ashburner, Ball, Blake, Botstein, Butler, Cherry, … Sherlock (2000, May). Gene ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. doi: 10.1038/75556

ATCC. (2019). Six factors affecting reproducibility in life science research and how to handle them. *Nature News.* Retrieved from `https://www.nature.com/articles/d42473-019-00004-y`

Attwood, Gisel, Eriksson, & Bongcam-Rudloff. (2011, November). Concepts, historical milestones and the central place of bioinformatics in modern biology: A european perspective. In *Bioinformatics - trends and methodologies.* InTech. doi: 10.5772/23535

Attwood, Kell, McDermott, Marsh, Pettifer, & Thorne. (2009, December). Calling international rescue: knowledge lost in literature and data landslide! *Biochemical Journal*, *424*(3), 317–333. doi: 10.1042/bj20091474

Barber, Xu, Pérez-Losada, Jara, & Crandall. (2012, June). Conflicting evolutionary patterns due to mitochondrial introgression and multilocus phylogeography of the patagonian freshwater crab aegla neuquensis. *PLoS ONE*, *7*(6), e37105. Retrieved from `https://doi.org/10.1371/journal.pone.0037105` doi: 10.1371/journal.pone.0037105

Berrios, Beheshti, & Costes. (2018). FAIRness and usability for open-access omics data systems. *AMIA … Annual Symposium proceedings. AMIA Symposium*, 232-241.

Bio-informatics services. (2010). Retrieved from `https://www.biosyn.com/bioinformatics.aspx`

Boeckhout, Zielhuis, & Bredenoord. (2018, May). The FAIR guiding principles for data stewardship: fair enough? *European Journal of Human Genetics*, *26*(7), 931–936. doi: 10.1038/s41431-018-0160-0

Cannata, N., Merelli, E., & Altman, R. B. (2005, Dec). Time to organize the bioinformatics resourceome. *PLoS Computational Biology*, *1*(7). doi: 10.1371/journal.pcbi.0010076

Conflicting evolutionary patterns due to mitochondrial introgres-

sion and multilocus phylogeography of the patagonian fresh-water crab aegla neuquensis. (2015). Retrieved from `https://omictools.com/2230449698700efa4e922a237f243b5e-protocol`

Crossref api python client. (2018). Retrieved from `https://travis-ci.org/fabiobatalha/crossrefapi`

Dijk, Jones, Mons, Fankhauser, Muscella, O'Neill, & Niccolucci. (n.d.). *How to implement the fair data principles?* Retrieved from `https://www.eoscpilot.eu/content/how-implement-fair-data-principles`

Django 2.1. (2018). Retrieved from `https://www.djangoproject.com/`

Dumontier, Garcia, Kusak, Lamprecht, Martínez, Wilkinson, & Zaveri. (2018). *Fairness assessment for software.* Retrieved from `https://github.com/dbcls/bh18/wiki/FAIRness-assessment-for-software`

Elixir. (2017). *Elixir position paper on fair data management in the life sciences.* Retrieved from `https://elixir-europe.org/system/files/elixir_statement_on_fair_data_manage`

European bioinformatics institute. (2019). Retrieved from `https://www.ebi.ac.uk/services/all`

Fair guidelines - fair guiding principles consultancy. (2017). Retrieved from `https://thehyve.nl/solutions/fair-guidelines/`

Fair principles. (2016). Retrieved from `https://www.go-fair.org/fair-principles/`

FAIR principles for data stewardship. (2016, April). *Nature Genetics*, *48*(4), 343–343. doi: 10.1038/ng.3544

Fairshake. (2018). Retrieved from `https://fairshake.cloud/`

Flicek, Ahmed, Amode, R., Barrell, Beal, Brent, … Searle (2012, November). Ensembl 2013. *Nucleic Acids Research*, *41*(D1), D48–D55. doi: 10.1093/nar/gks1236

geckodriver. (2018). Retrieved from `https://github.com/mozilla/geckodriver/releases`

Google api python client library. (2018). Retrieved from `https://developers.google.com/api-client-library/python/`

Google custom search. (2006). Retrieved from `https://cse.google.com/cse/all`

Grivell. (2002). Mining the bibliome: searching for a needle in a haystack?: New computing tools are needed to effectively scan the growing amount of scien-

tific literature for useful information. *EMBO Reports*, *3*(3), 200–203. doi: 10.1093/embo-reports/kvf059

Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0. (2016, Oct). Retrieved from `https://www.force11.org/fairprinciples`

Hucka, Finney, Sauro, Bolouri, Doyle, Kitano, … Wang (2003, March). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, *19*(4), 524–531. doi: 10.1093/bioinformatics/btg015

Ivanović, Ivanović, & Layfield. (2019). FAIRness at university of novi sad - discoverability of PhD research results for non-serbian scientific community–. *Procedia Computer Science*, *146*, 3–10. doi: 10.1016/j.procs.2019.01.071

Jointly designing a data fairport. (2014). Lorentz Center. Retrieved from `https://www.lorentzcenter.nl/lc/web/2014/602/info.php3?wsid=602`

Kanehisa, Goto, Sato, Furumichi, & Tanabe. (2011, November). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research*, *40*(D1), D109–D114. doi: 10.1093/nar/gkr988

Kent, Sugnet, Furey, Roskin, Pringle, Zahler, & Haussler. (2002, May). The human genome browser at UCSC. *Genome Research*, *12*(6), 996–1006. doi: 10.1101/gr.229102

Leipzig. (2016, March). A review of bioinformatic pipeline frameworks. *Briefings in Bioinformatics*, bbw020. doi: 10.1093/bib/bbw020

LiberEUROPE. (2017). *Implementing fair data principles: The role of libraries.*

Luscombe, Greenbaum, & Gerstein. (2001). What is bioinformatics? a proposed definition and overview of the field. *Methods of Information in Medicine*, *40*(04), 346–358. doi: 10.1055/s-0038-1634431

Maglott. (2004, December). Entrez gene: gene-centered information at NCBI. *Nucleic Acids Research*, *33*(Database issue), D54–D58. doi: 10.1093/nar/gki031

Matthews. (2017, January). Fairness in scientific publishing. *F1000Research*, *5*, 2816. doi: 10.12688/f1000research.10318.2

Mayer (Ed.). (2011). *Bioinformatics for omics data.* Humana Press. doi: 10.1007/978-1-61779-027-0

mysqlclient. (2018). Retrieved from `https://github.com/PyMySQL/mysqlclient-python`

nectar-rds / fair assessment tool. (2018). Retrieved from `https://www.ands-nectar-rds.org.au/fair-tool`

Neylon. (2016, Sep). *Fair enough? fair for one? fair for all!* Retrieved from

`http://cameronneylon.net/blog/fair-enough-fair-for-one-fair-for-all/`

phpmyadmin 4.8.2. (2018). Retrieved from `https://www.phpmyadmin.net/`

Python 3.7. (2018). Retrieved from `https://www.python.org/`

Pyyaml. (2018). Retrieved from `https://pyyaml.org/`

Rahal, & Havemann. (2019). Science in crisis. is open science the solution? doi: 10.31222/osf.io/3hb6g

Selenium python client. (2018). Retrieved from `https://github.com/SeleniumHQ/Selenium`

Smith, Ashburner, Rosse, Bard, Bug, Ceusters, … et al. (2007). The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, *25*(11), 1251–1255. doi: 10.1038/nbt1346

Taylor, Field, Sansone, Aerts, Apweiler, Ashburner, … et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the mibbi project. *Nature Biotechnology*, *26*(8), 889–896. doi: 10.1038/nbt.1411

Visual studio code. (2015). Retrieved from `https://code.visualstudio.com`

Wilkinson, Dumontier, Aalbersberg, Appleton, Axton, Baak, … Mons (2016, March). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018. doi: 10.1038/sdata.2016.18

Wilkinson, Sansone, Schultes, Doorn, da Silva Santos, L. O. B., & Dumontier. (2018, June). A design framework and exemplar metrics for FAIRness. *Scientific Data*, *5*, 180118. doi: 10.1038/sdata.2018.118

Wise, de Barron, G., Splendiani, Balali-Mood, Vasant, Little, … Hedley (2019, April). Implementation and relevance of FAIR data principles in biopharmaceutical r&d. *Drug Discovery Today*, *24*(4), 933–938. doi: 10.1016/j.drudis.2019.01.008

Wittig, Rey, Weidemann, & Müller. (2017, November). Data management and data enrichment for systems biology projects. *Journal of Biotechnology*, *261*, 229–237. doi: 10.1016/j.jbiotec.2017.06.007

Xampp. (2017). Retrieved from `https://www.apachefriends.org/index.html`

Yousef, & Allmer. (2014). *Mirnomics: microrna biology and computational analysis.* Humana Press.

Zakrisson, & Kronfoth. (2017). Tools in science. *Acta Physiologica*, *220*(1), 3-6. doi: 10.1111/apha.12878

Zvelebil, & Baum. (2008). *Understanding bioinformatics.* Garland Science.