Bioinformatics Resource Portal

Nigel Alfino
University of Malta
nigel.alfino.16@um.edu.mt

Abstract

Tools and datasets in bioinformatics are being generated at a very fast rate. Meanwhile a lack of effort is being made in making these resources findable and reusable. This has made it difficult for researchers to find the right resource. This paper proposes an automated tool to assess a resource's Findability, Accessibility, Interoperability and Reusability in order to obtain a score that would provide researchers with a degree of trust in that resource.

1. Introduction and Background

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 [1]. 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 [1].

The issue of being unable to find the proper resource rises from the fact that there has been a lack of effort made in the proper storage and organization of this rapidly accumulating data [2]. Furthermore, there is a lack of effort made in scientific research to make results reproducible [3]. 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 [3]. However, often times scientific findings cannot be reproduced leading to a waste of time and resources [3]. 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 [3].

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 [4]. 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 [5]. Publishers, governmental agencies and funders have started requesting researchers to present "data management and stewardship plans" [5]. Good data management and stewardship will then ensure high quality digital publications which will in turn further simplify future research [5].

In order to ensure this proper data management and data stewardship the FAIR guiding principles provide four foundational principles [5]:

- Findability, which is described by Wilkinson et al. [5] 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." This is a crucial prerequisite to the other three principles presented by FAIR [6].

- Accessibility, which addresses the necessity to make newly generated and previously existing data as well as digital assets more accessible. [6]

- Interoperability, which 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. This can be ensured through the use of a formal, accessible, shared and broadly applicable language for knowledge representation. [7]

- Reusability, which ensures that data and collections have clear usage licenses and provide accurate information on provenance [7]. The optimisation of reuse is the ultimate goal of FAIR and is what separates traditional data management from FAIR data stewardship [6].

This paper will focus on presented a method to automate a FAIR assessment for tools and datasets, providing researchers to gauge the Findability, Accessibility, Interoperability and Reusability of the resources they are using.

2. Aims and Objectives

The main goal of this project is to develop a tool that automates the assessment of a tool or dataset's Findability, Accessibility, Interoperability and Reusability (FAIR) and produce a FAIR score. This goal can be broken down into the following objectives:

- Obtain a comprehensive overview of tasks in bioinformatics and which tools are commonly used in the task.
- Obtain a searchable archive of tools and datasets.
- Enhance the archive with a FAIR score indicating the Findability, Accessibility, Interoperability and Reusability of the resource to aid researchers make decisions on which tools to use.

3. Design

The system presented by this study provides the user with a FAIR score for the tools and datasets within the system. The user will be able to see the score through a portal, which will also allow the user to see where there was a lack of information which would improve the tool or dataset's FAIR score. The user will then be able to refine the scores through the portal. The portal also allows the user to define pipelines of tools and datasets, and the system will automatically calculate the pipeline's FAIR score based on the tools and datasets used.

The FAIR scores for tools and datasets were calculated using metrics obtained from the FAIRshake tool rubric [8]. These metrics were based off of the guidelines presented by Wilkinson et al. [5] and Dumontier et al. [13]. 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 to calculate a dataset's FAIR score were obtained from the FAIRshake dataset rubric [8] and the nectar-rds FAIR Assessment Tool [9].

The metrics used in this study use a priority system to calculate a score to ensure that tools and datasets that are wholly findable, accessible, interoperable and reusable get a better FAIR score to properly represent their better Findability, Accessibility, Interoperability and Reusability when compared to other tools.

3.1 FAIR Metrics – Tools

The following are the metrics used to calculate a tool's FAIR score, along with the priority they were assigned.

Findability:

- Tool is findable on a website.
  - This metric is mandatory, that is, it must be satisfied in order for the system to be able to automate the assessment.
- Tool is freely downloadable.
  - This metric is given a high priority.
- A proper description is provided for the tool.
  - This metric is given a medium priority.
- Previous versions of the tool are made available.
  - This metric is given a low priority.
- Tool has a unique identifier.
  - This metric is given a medium priority

Accessibility:

- Tool can be programmatically accessed through an API.
  - This metric is given a medium priority.
- Tool can be accessed through a set of commands executed in a command line interface.
  - This metric is given a medium priority. This metric is an additional metric presented by this study, in order to provide a more accurate Accessibility score.

Interoperability:

- Tool compatibility information is provided.
  - This metric is given medium priority. This metric is an additional metric presented by this study, in order to provide a more accurate Interoperability score. This metric is split into three parts, Windows compatibility, Linux compatibility and Mac compatibility.
- Tool source code is provided.
  - This metric is given medium priority. This metric is an additional metric presented by this study in order to provide a more accurate Interoperability score.

Reusability:

- Tool is hosted in a public repository.
  - This metric is given a high priority.
- Tool uses community accepted ontologies.
  - This metric is given medium priority. If the tool does not use an ontology this metric will not apply and as such will not affect its FAIR score.
- Proper documentation of the tool is provided.
  - This metric is given medium priority.
- Developer contact information is provided.
  - This metric is given low priority.
- Information on how to cite tool is provided.
  - This metric is given low priority.

3.2 FAIR Metrics – Datasets

The following are the metrics used to calculate a dataset's FAIR score, along with the priority they were assigned.

Findability:

- Dataset is findable on a website.
  - This metric is mandatory, that is, it must be satisfied in order for the system to be able to automate the assessment.
- Dataset has a proper description.
  - This metric is given medium priority.
- Dataset has a unique identifier.
  - This metric is given low priority.

Accessibility:

- Dataset is freely downloadable from the website.
  - This metric is given a high priority.
- 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 identifier linking it to the paper that originally presented the dataset, then this metric is automatically satisfied.

Interoperability:

- 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.

Reusability:

- Information is provided on how to cite the dataset.
  - This metric is given low priority.
- Contact information is provided for the creator(s) of the dataset.
  - This metric is given low priority.
- Previous versions of the dataset are made available.
  - This metric is given a low priority.

4. Implementation

The system was implemented using Python [10] along with a few freely available supporting libraries. The portal was built using Django (2018) [11], a web framework for Python. It 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 a tool or dataset's website in order to increase that resource's FAIR score. The user can then manually refine the score by providing any information which the automated assessment tool might have missed.

4.2 FAIR Assessment Tool

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 is limited to 20 links. Once the website is found, the link will be used by Selenium to obtain information and perform the assessment.

4.3 Tool FAIR Assessment

Findability

Once the website is found, the assessment tool will begin to search the website for information based on the metrics for Findability. This is done by searching for information on the website that could provide information to satisfy the metrics, such as looking for download links for the tool. Some assumptions are also drawn when certain information is found, such as if the tool is hosted in a GitHub repository it is automatically assumed that it is freely downloadable, has freely available source code, is found in a public repository and has previous versions available. After finding the required information, a score is calculated. The scores of the Findability metrics for tools are as follows:

- Tool is freely downloadable: 40%
- A proper description is provided for the tool: 25%
- Previous versions of the tool are made available: 25%
- Tool has a unique identifier: 10%

Accessibility

After calculating a score for Findability, the tool moves on to look for information to satisfy the metrics provided for Accessibility. When the required information is found a score is then calculated. The scores of tool Accessibility metrics are as follows:

- Tool can be programmatically accessed through an API: 50%
- Tool can be accessed through a set of commands executed in a command line interface: 50%

Interoperability

The assessment tool then calculates a score for Interoperability by searching for information to satisfy the metrics presented for Interoperability. When the required information is found, a score is calculated. The scores of tool Interoperability metrics are as follows:

- Tool compatibility information is provided: 50%
  - This metric is split into three parts as follows:
    Windows Compatibility: 16.67%
    Mac Compatibility: 16.67%
    UNIX Compatibility: 16.67%
  - If all three parts are satisfied then a full 50% is awarded, otherwise the corresponding fraction of the 50% is awarded based on which OS are supported.
- Tool source code is provided: 50%

Reusability

Finally the assessment tool searches for information to calculate a Reusability score based on the provided metrics. The Reusability metrics have different scores based on whether a tool uses an ontology or not. If a tool does not use an ontology, the scores of tool Reusability metrics are as follows:

- Tool is hosted in a public repository: 50%
- Proper documentation of the tool is provided: 25%
- Developer contact information is provided: 12.5%
- Information on how to cite the tool is provided: 12.5%

If the tool does use an ontology, the scores of Reusability metrics are as follows:

- Tool is hosted in a public repository: 40%
- Tool uses community accepted ontologies: 20%
- Proper documentation of the tool is provided: 20%
- Developer contact information is provided: 10%
- Information on how to cite the tool is provided: 10%

4.4 Dataset FAIR Assessment

Findability

Once the dataset's website is found, the assessment tool begins to search the website for information based on the metrics for Findability. The scores of dataset Findability metrics are as follows:

- Dataset has a proper description: 60%
- Dataset has a unique identifier: 40%

Accessibility

Once a score is calculated for a dataset's accessibility, the assessment tool starts searching for information based on the metrics for dataset Accessibility. If a dataset has a unique identifier, it is also automatically assumed that the metadata will be available when the metadata is no longer available. The following are the scores for dataset Accessibility metrics:

- Dataset is freely downloadable from the website: 80%
- Metadata will still be available even if the data is no longer available: 20%

Interoperability

The assessment tool will then calculate the dataset's Interoperability by searching for information based on the metric for dataset Interoperability. The score for the dataset's Interoperability metric is as follows:

- Information is provided on which format(s) or file format(s) the dataset is available in: 100%

Reusability

The assessment tool then finally begins searching for information based on the metrics for dataset Reusability. The scores of dataset Reusability metrics are as follows:

- Information is provided on how to cite the dataset: 33.3%
- Contact information is provided for the creator(s) of the dataset: 33.3%
- Previous versions of the dataset are made available: 33.3%

4.5 Pipeline FAIR Assessment

The user can define pipelines made from tools and datasets, and the system will automatically calculate a FAIR score based on the FAIR scores of the individual tools and datasets in it.

Findability

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

Accessibility

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

Interoperability

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

Reusability

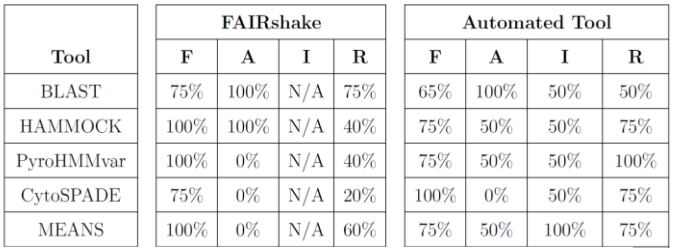
$$\frac{Reusability_{T1} + Reusability_{D1} + ... + Reusability_{Tn} + Reusability_{Dn}}{N_T + N_D}$$

5. Results and Evaluation

To evaluate the system the following sets of data were gathered:

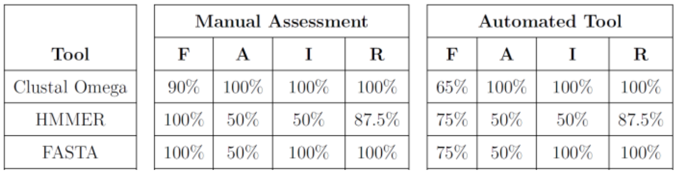
- Tools obtained form FAIRshake tool rubric assessment [8]. These will serve as a benchmark against which the system's scores will be compared.
- Tools obtained from EMBL-EBI tool service [12]. Given that these tools are popular and used often, it is expected that these tools should generate a favourable FAIR score.
- Datasets obtained from the FAIRshake dataset rubric assessment [8]. These will serve as a benchmark against which the system's scores will be compared.
- Datasets obtained from EMBL-EBI dataset service [12]. Given that the hosted datasets are popular and used quite often, they are expected to generate a favourable FAIR score.
- Pipeline of tools obtained from Barber et al. [14]

- ## 5.2 Results



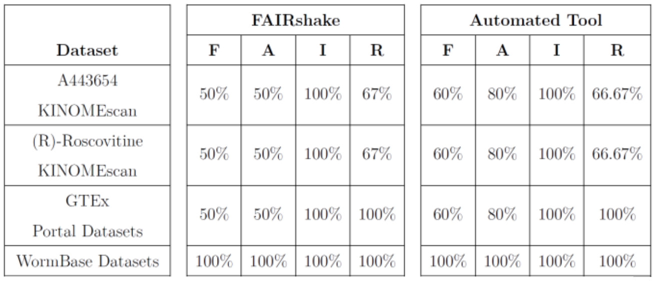
Tool	FAIRshake				Automated Tool			
	F	A	I	R	F	A	I	R
BLAST	75%	100%	N/A	75%	65%	100%	50%	50%
HAMMOCK	100%	100%	N/A	40%	75%	50%	50%	75%
PyroHMMvar	100%	0%	N/A	40%	75%	50%	50%	100%
CytoSPADE	75%	0%	N/A	20%	100%	0%	50%	75%
MEANS	100%	0%	N/A	60%	75%	50%	100%	75%

Figure 1: Comparison of results between FAIRshake assessment on tools and assessment made by automated assessment tool.



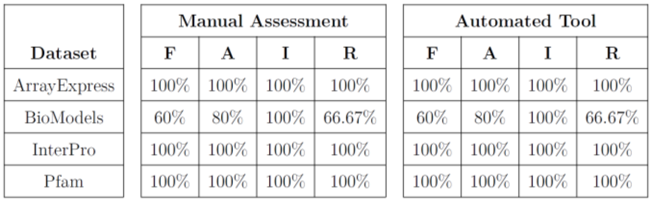
Tool	Manual Assessment				Automated Tool			
	F	A	I	R	F	A	I	R
Clustal Omega	90%	100%	100%	100%	65%	100%	100%	100%
HMMER	100%	50%	50%	87.5%	75%	50%	50%	87.5%
FASTA	100%	50%	100%	100%	75%	50%	100%	100%

Figure 2: Comparison of results between a manual assessment made on EMBL-EBI tools and assessment made by automated assessment tool.



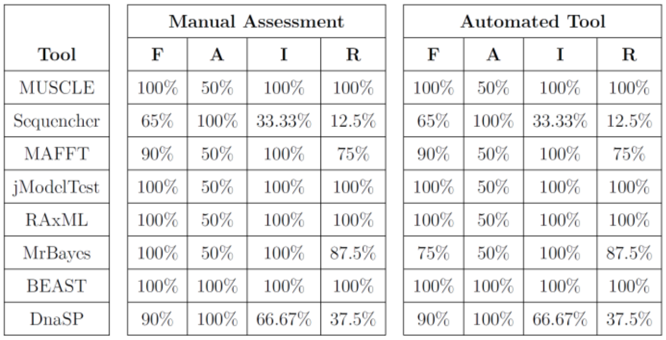
Dataset	FAIRshake				Automated Tool			
	F	A	I	R	F	A	I	R
A443654 KINOMEscan	50%	50%	100%	67%	60%	80%	100%	66.67%
(R)-Roscovitine KINOMEscan	50%	50%	100%	67%	60%	80%	100%	66.67%
GTEx Portal Datasets	50%	50%	100%	100%	60%	80%	100%	100%
WormBase Datasets	100%	100%	100%	100%	100%	100%	100%	100%

Figure 3: Comparison of results between FAIRshake assessment on datasets and assessment made by automated assessment tool.



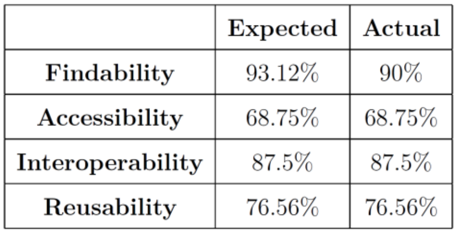
Dataset	Manual Assessment				Automated Tool			
	F	A	I	R	F	A	I	R
ArrayExpress	100%	100%	100%	100%	100%	100%	100%	100%
BioModels	60%	80%	100%	66.67%	60%	80%	100%	66.67%
InterPro	100%	100%	100%	100%	100%	100%	100%	100%
Pfam	100%	100%	100%	100%	100%	100%	100%	100%

Figure 4: Comparison of results between manual assessment made on EMBL EBI datasets and assessment made by automated assessment tool.



Tool	Manual Assessment				Automated 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%

Figure 5: Comparison of results between manual assessment made on tools in pipeline presented by Barber et al. (2012) [14] and assessment made by automated assessment tool.



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

Figure 6: Comparison of pipeline expected FAIR score against actual FAIR score calculated by system.

5.3 Evaluation

Tools

In figure 1 and figure 2 a few differences can be noted in the results for Findability. This difference arose from an inherent difficulty in obtaining a tool's unique identifier. In the majority of cases, the tool was unable to find a DOI linking to the paper that presented the tool.

The other differences in results for Accessibility and Reusability in figure 1 all arose from the fact that the metrics presented in this study use a different weighting system than that of FAIRshake.

The rest of the results in figure 2 show that the assessment of tools can be mostly automated.

Datasets

The differences that can be seen in figure 2 all arise from the fact that the assessment tool gives different priorities to different metrics. However, all the information obtained by the automated tool is the same as that for FAIRshake.

The results from figure 4 show that the tool managed to obtain all the data necessary for an assessment as would be expected in a real-world scenario.

Pipeline

The results in figure 5 show that one of the tools obtained a lower Findability score by the automated assessment tool, as it was unable to find a unique identifier for that tool. This resulted in a 3.12% loss in the pipeline's total Findability.

6. Conclusions and Future Work

The results obtained from the evaluation allowed us to reach the following conclusions:

- The results obtained from the automated assessment tool for datasets was consistent with those obtained from FAIRshake, which means that the assessment of datasets can be fully automated.
- The assessment for tools in bioinformatics however is not fully automatable. Firstly, the second metric presented for Reusability, that is, the tool uses a community accepted ontology, cannot be automated as there is no way to tell when a tool doesn't use an ontology altogether. This is rectified through user input.

  Furthermore, the tool suffers an inherent inability to obtain a tool's unique identifier, which would result in a lower FAIR score.
- The rest of the results for a tool's FAIR assessment were consistent with those obtained from FAIRshake, which show that the tool performs as expected
- The calculation of a pipeline's FAIR score also performs as expected. The inability to find a tool's DOI however affects the total Findability of the pipeline, resulting in a much lower Findability score if there are multiple tools for which a unique identifier was not found.

6.1 Future Work

The following issues can be addressed in the future:

- Further research can be made to improve metrics to calculate a more accurate FAIR score.
- Multi-threading techniques could be implemented to the system to improve efficiency of FAIR assessment by calculating FAIR scores of multiple tools at once.
- Improving web crawling techniques to overcome Selenium limitations, further improving the efficiency of FAIR assessments.

7. Bibliography

[1] N. Cannata, E. Merelli, and R. B. Altman, "Time to Organize the Bioinformatics Resourceome," PLoS Computational Biology, vol. 1, no. 7, 2005.

[2] T. Attwood, A. Gisel, N.-E. Eriksson, and E. Bongcam-Rudloff, "Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective," Bioinformatics - Trends and Methodologies, 2011.

[3] ATCC, "Six factors affecting reproducibility in life science research and how to handle them," Nature News, 2019. [Online]. Available: https://www.nature.com/articles/d42473-019-00004-y.

[4] "FAIR principles for data stewardship," Nature Genetics, vol. 48, no. 4, pp. 343–343, 2016.

[5] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016, March). "The FAIR Guiding Principles for scientific data management and stewardship," Scientific Data, vol. 3, p. 160018, 2016.

[6] Wise, J., de Barron, A. G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., … Hedley, V. (2019, April). "Implementation and relevance of FAIR data principles in biopharmaceutical R&D," Drug Discovery Today, vol. 24, no. 4, pp. 933–938, 2019.

[7] LiberEUROPE. (2017). Implementing fair data principles: The role of libraries.

[8] K. C. Team Nitrogen, "FAIRshake," FAIRshake, 2018. [Online]. Available: https://fairshake.cloud/.

[9] "nectar-rds | FAIR Assessment Tool," ands, 2018. [Online]. Available: https://www.ands-nectar-rds.org.au/fair-tool.

[10] Python 3.7. (2018). Retrieved from https://www.python.org/

[11] Django 2.1. (2018). Retrieved from https://www.djangoproject.com/

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

[13] Dumontier, M., Garcia, L., Kusak, M., Lamprecht, A.-L., Martínez, C., Wilkinson, M., & Zaveri, A. (2018). Fairness assessment for software. Retrieved from https://github.com/dbcls/bh18/wiki/FAIRness-assessment-for-software

[14] B. R. Barber, J. Xu, M. Pérez-Losada, C. G. Jara, and K. A. Crandall, "Conflicting Evolutionary Patterns Due to Mitochondrial Introgression and Multilocus Phylogeography of the Patagonian Freshwater Crab Aegla neuquensis," PLoS ONE, vol. 7, no. 6, 2012.

[15] N. Alfino, "Bioinformatics Resource Portal," University of Malta, 2019. (Unpublished)