A SET OF MECHANISMS TO COLLECT, DISPLAY AND ANALYZE QUALITY METRICS OF RESEARCH OBJECTS AS A METHOD TO AUGMENT REPRODUCIBILITY AND OPENNESS IN BIOINFORMATICS RESEARCH

by

PITYANOU KONSTANTINA

BSc. Technological Educational Institute of Crete, 2019

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

SCHOOL OF ENGINEERING

HELLENIC MEDITERRANEAN UNIVERSITY

2022

Approved by:

Major Professor
Professor Manolis Tsiknakis

Abstract

Open and community-responsive tool repositories and Workflow Management Systems aid in the automation and repeatability of large-scale data analysis. However, such repositories do not include evaluation methods on those components, which could help scientists to determine which research object could satisfy their requirements. Furthermore, the problem of misattribution expands the gap of ignorance and misguided information, where users do not get acknowledged for their contributions in the scientific field. Moreover, the problem of profile building does not allow researchers to expand their contributions to different platforms and therefore, further assist with their knowledge.

This thesis aimed to find solutions to some vital problems in the domain of bioinformatics. It contributes to a tool and workflow repository, the OpenBio platform, by adding a rich set of both automatically acquired and user-provided quality metrics. Those metrics are displayed in an intuitive user interface, allowing users to quickly explore them in order to make the most suitable and informed decision regarding the optimal component that suits best their analysis. Next, it tried to find a solution for the misattribution issue of bioinformatics tools, where tools were used without being properly cited in academic papers by connecting the OpenBio platform with ORCID. This acted as a counter incentive for authors when it came to submitting tools in open science repositories. Thus, it allowed users to be properly cited for the work that they could submit in open science environments that deviate from the typical “academic journal” setting such as tool repositories, and open Workflow Management Systems. Finally, it provided clear and objective indications of the expertise and the scientific activity of users, acting as an incentive for them to be more open and active by ranking them according to their quality and quantity of contributions.

The User Experience Questionnaire was used in order to evaluate the implementation of those added components. The questionnaire was filled out by 8 participants after they had tested those components on the OpenBio platform. The overall results showed that the user experience regarding the added components produced positive results. The calculation of the confidence interval showed that more participants could be needed in order to express a more stable result. However, the Lambda-2 coefficient per scale showed that the answers that were given in the UEQ had high percentages to be true and valid.

Keywords: OpenBio, OpenBio.eu, bioinformatics, repository, tool, workflow, citation, evaluation, metrics, orcid, misattribution, profile building

Περίληψη

Τα ανοιχτά και κοινωνικά αποθετήρια εργαλείων και τα Συστήματα Διαχείρισης Ροής Εργασίας βοηθούν στον αυτοματισμό και την επαναληψιμότητα της ανάλυσης δεδομένων μεγάλης κλίμακας. Ωστόσο, τέτοια αποθετήρια δεν περιλαμβάνουν μεθόδους αξιολόγησης αυτών των στοιχείων, οι οποίες θα μπορούσαν να βοηθήσουν τους επιστήμονες να προσδιορίσουν το ερευνητικό αντικείμενο που θα μπορούσε να ικανοποιήσει τις απαιτήσεις τους. Επιπλέον, το πρόβλημα της εσφαλμένης απόδοσης συνεισφοράς διευρύνει το χάσμα της άγνοιας και της λανθασμένης πληροφόρησης, όπου οι χρήστες δεν αναγνωρίζονται για τη συνεισφορά τους στον επιστημονικό τομέα. Επιπλέον, το πρόβλημα της δημιουργίας προφίλ δεν επιτρέπει στους ερευνητές να επεκτείνουν τις συνεισφορές τους σε διαφορετικές πλατφόρμες και συνεπώς να βοηθήσουν περαιτέρω με τις γνώσεις τους.

Αυτή η διατριβή είχε ως στόχο να αντιμετωπίσει σημαντικά προβλήματα στον τομέα της βιοπληροφορικής. Συνεισφέρει σε ένα αποθετήριο εργαλείων, δεδομένων και ροών εργασίας, την πλατφόρμα OpenBio, προσθέτοντας ένα πλούσιο σύνολο μετρήσεων ποιότητας που μπορούν να αποκτηθούν αυτόματα ή και να παρέχονται από τον χρήστη. Αυτές οι μετρήσεις εμφανίζονται σε μια διαισθητική διεπαφή χρήστη, επιτρέποντας στους χρήστες να τις εξερευνήσουν γρήγορα προκειμένου να λάβουν την πιο κατάλληλη και ενημερωμένη απόφαση σχετικά με το βέλτιστο αντικείμενο που ταιριάζει καλύτερα στην ανάλυσή τους. Στη συνέχεια, προσπάθησε να βρει μια λύση για το ζήτημα της εσφαλμένης απόδοσης συνεισφοράς στα εργαλεία βιοπληροφορικής, όπου τα εργαλεία χρησιμοποιήθηκαν χωρίς να συνδέονται σωστά με την ακαδημαϊκή συνεισφορά του επιστήμονα, συνδέοντας την πλατφόρμα OpenBio με το ORCID. Αυτό λειτούργησε ως κίνητρο για τους συγγραφείς όταν επρόκειτο να υποβάλουν εργαλεία σε ανοιχτά αποθετήρια επιστήμης. Έτσι, επέτρεψε στους χρήστες να αναφέρονται σωστά για την εργασία που θα μπορούσαν να υποβάλουν σε περιβάλλοντα ανοιχτής επιστήμης που αποκλίνουν από την τυπική ρύθμιση του «ακαδημαϊκού περιοδικού», όπως τα αποθετήρια εργαλείων και τα ανοιχτά Συστήματα Διαχείρισης Ροής Εργασιών. Τέλος, παρείχε σαφείς και αντικειμενικές ενδείξεις για την τεχνογνωσία και την επιστημονική δραστηριότητα των χρηστών, λειτουργώντας ως κίνητρο για να είναι πιο ανοιχτοί και δραστήριοι, ταξινομώντας τους ανάλογα με την ποιότητα και την ποσότητα των συνεισφορών τους.

Το User Experience Questionnaire χρησιμοποιήθηκε για να αξιολογηθεί η υλοποίηση αυτών των στοιχείων που προστέθηκαν. Το ερωτηματολόγιο συμπληρώθηκε από 8 συμμετέχοντες

αφού είχαν δοκιμάσει αυτά τα στοιχεία στην πλατφόρμα του OpenBio. Τα συνολικά αποτελέσματα έδειξαν ότι η εμπειρία του χρήστη σχετικά με τα πρόσθετα στοιχεία παρήγαγε θετικά αποτελέσματα. Ο υπολογισμός του διαστήματος εμπιστοσύνης έδειξε ότι θα μπορούσαν να χρειαστούν περισσότεροι συμμετέχοντες για να εκφραστεί ένα πιο σταθερό αποτέλεσμα. Ωστόσο, ο συντελεστής lambda-2 ανά κλίμακα υπέδειξε ότι οι απαντήσεις που δόθηκαν στο UEQ είχαν υψηλά ποσοστά αλήθειας και εγκυρότητας.

Λέξεις-κλειδιά: OpenBio, OpenBio.eu, βιοπληροφορική, αποθετήριο, εργαλείο, δεδομένα, ροή εργασιών, παραπομπή, αξιολόγηση, μετρήσεις, orcid, εσφαλμένη απόδοση, δημιουργία προφίλ

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank Prof. Manolis Tsiknakis, Alexandros Kanterakis, and Lefteris Koumakis for their support and guidance throughout my master thesis. I also want to thank my friends and family for supporting me and always believing in me.

1 Introduction

As a scientific discipline, bioinformatics has advanced over the last few years. It is a relatively modern interdisciplinary field that promises to join a disparate type of resources with the single purpose of advancing healthcare with the help of computer science. The advancement of Deoxyribonucleic acid (DNA) profiling techniques, the availability of HPC (High Performance Computing) infrastructure, and the introduction of big data technologies, has enabled the large scale and whole genome analysis in a personalized fashion. In this large and complicated scientific study field, many tools have been created to analyze the ever-increasing volumes of data. Therefore, large-scale data analysis necessitates the sequential execution of several command line applications. The development of workflow engines aid in the automation and repeatability of these processes. Through their graphical interfaces, systems like Biopipe [1], Taverna [2], Galaxy [3], GeneProf [4], or PegaSys [5] are simple to learn and use. Others, such as Ruffus [6], Pwrake [7], GXP Make [8], and Bpipe [9], use text-based workflow interpretations, which can be useful because workflows can be edited without a graphical interface, and programmers can cooperate on them using source code management tools.

Some of the long-term visions of bioinformatics are to bring genomics to the everyday healthcare clinical practice, to advance personalized medicine and to offer reliable and secure health monitoring from small and accessible devices that surround the individual. A common denominator of these visions is that bioinformatic methods are gradually becoming more critical and vital in healthcare and health monitoring. Healthcare is considered one of the most sensitive and crucial software development areas, where quality, resilience and robustness is of paramount importance. Yet, various surveys and studies in this area have pinpointed that despite the quantity of available tools, various quality indicators, such as the availability, maintainability, support, documentation and availability of tests is the exception rather than the norm [10]. A common method to augment the quality of a large and dispersed set of software components is to include them in a uniform repository that continuously tests it over various metrics. Some examples include the Advanced package tool (APT) interface for Linux systems [11], the Python Package Index [12] for the python language, and the npm for JavaScript [13]. Today, almost all available tool repositories for bioinformatics, such as bio.tools [14] and Datasets2Tools [15], act as plain/flat lists without the ability to actually test the listed components. Therefore, given the importance of

bioinformatics in several biological and biomedical investigations, substantial effort should be made to make computational analyses repeatable and workflow systems and tools maintainable. A scientist should have the option to determine which tool, data, or workflow (known as research objects) could satisfy their requirements. Consequently, the methods of evaluation on any set of components should be investigated.

The term of misattribution refers to the incorrect or lack of acknowledgement of a scientist's contribution to the scientific field. Ignorance or misguided information can lead to incorrect attribution. Although numerous conventional methods are currently being used to evaluate a researcher's scientific career in publications, there is a lack of recognition in a scientist's developed software programs. Linking research data to publications and software, in an open and semantically consistent manner, can be critical for improving academic communication quality and openness. Manual labeling, developed algorithms, and third-party authenticators can assist to the linkage between an author and their published paper. Producers and consumers are still enamored of forms devised during the period of print publishing, and research reward schemes are still based on old delivery techniques. Although research papers and different publications have long benefited from an architecture that makes it simple to cite them, the same values do not apply to software data. Research on this subject is mandatory, in order to find different ways that can semantically connect authors and developers with their publications and created software.

An open tool repository and Workflow Management System should allow its users to build their profile and get acknowledged for their content. Indicators regarding their expertise and activity can help in their evaluation by other users and the increment of their popularity on the platform. Because the maintenance of such platforms and systems is dependent on the contribution of its users, those sites' creators should pay attention to them and look for methods to reward them through different approaches. Since awards can serve as social status indicators, greater prestige values can be given to different users based on their contributions, fulfilling a variety of socio-psychological roles. Therefore, those approaches should be researched as they can be part of strong indicators of a system's long-term worth.

The main purpose of this Master Thesis is to find solutions to some vital problems in bioinformatics, by extending a repository of tools and data that continuously tests and measures the included components. Based on the above-mentioned issues, three research questions were formed:

RQ1: How can we effectively solve the problem of misattribution?

RQ2: How can we approach the evaluation, in terms of effectiveness and efficiency, of a tool or workflow that exists in a repository?

RQ3: How can we evaluate the expertise and contribution of users, and assist them in the problem of profile building?

The different approaches that can answer those research questions will be embedded in the OpenBio platform [16], which is an open and community-responsive platform, that can function as a tool repository and Workflow Management System. OpenBio.eu [17] is developed in the CBML lab of the Institute of Computer Science of FORTH. It currently contains an environment where users can submit a tool or a dataset along with the commands and scripts that install it. Users can also combine these tools into workflows in an online graphics environment. Moreover, OpenBio.eu allows users to grade and comment on existing components. Every action in OpenBio.eu is semantically connected to a user. This master will implement the methods to collect all quality metrics presented on OpenBio.eu platform and display them in a user-friendly User Interface. It will allow users to query tools, data and workflows based on these metrics and will automatically generate a user profile that has all the indications regarding their expertise and popularity of these users. Finally, one of the most important objectives of this thesis will be to connect the platform with ORCID so users can be able to build an online profile and get acknowledged for their content whether that includes publications of created tools and workflows.

2 Systematic Literature Review

The expansion of the research community's ability to collect and transmit data has created major prospects. It is altering traditional scientific research methodologies and allowing the development of new ones. The capacity to identify, verify, obtain, and understand published digital data is essential. Data citations are required for these activities, as well as additional purposes such as credit attribution and authenticity verification. However, unlike references to literature, data references pose unique issues. In the lack of common standards like page numbers or chapters, how can one describe a specific subset of data? In the case of journal articles and other literature, the traditions and good practices for keeping the academic record through accurate citations to a publication are well established and known but assigning acknowledgment to data through bibliographic references is still not widely applied. Acknowledging the need for improved data reference and citation standards, as well as putting up the effort to solve such needs, has developed in various professions and disciplines. Errors and mismatches can obstruct the interchange and use of scientific data as conflicting traditions and practices evolve in different communities. Sharing experiences across groups may be required, or at the very least beneficial to achieve the full capability of public data.

Complex algorithms and data analysis are increasingly being used to create meaningful technological discoveries. These computations may include hundreds of stages, each of which may use several models and data sources developed by different organizations. The construction and maintenance of such highly advanced distributed algorithms has many obstacles, and the increasingly ambitious scientific research is constantly testing the limitations of current technology. Research objects, such as tools and workflows, have evolved as a paradigm for describing and managing these massive scientific computations, while accelerating technological discovery. Scientific workflows may both organize dataflow across relevant data processing and analysis stages and provide the tools to execute them in a dispersed context. However, the assessment of those research objects receives limited attention. The researcher should be able to assess the effectiveness and usability of each research object to determine whether it meets their criteria. Metrics can play a significant role in any research system, particularly an open one, and they can be utilized to accelerate development. As a result, many types of metrics should be investigated since they can help lead qualitative assessments and increase the understanding of the value of study outcomes and research objects.

As technological innovations and platforms are being developed, the software and computer environments react rapidly. The environment of programming languages, techniques, and developments is extremely complicated, with tools and actions adjusting to different factors on a regular basis. The rise of integrated learning environments has altered the perceptions of social relationships, which are always crucial in the learning process, as being especially significant in the development of personal forms of collaboration and environments. Because there is commonly a lack of social communication in these situations, virtual awards and badges can be utilized to increase social context visibility and improve learning frameworks. Those rewards adopt gamification methods that have been utilized to enhance human productivity and involvement in a range of systems and applications. Because users participate without receiving financial rewards, it is difficult for online communities to motivate them to actively engage. Therefore, it is important to research appropriate gamification methods to maintain a system's long-term worth and evaluate the user's contribution and expertise on the platform.

This section represents a systematic literature review that was conducted to find possible solutions to the problems of misattribution, research objects evaluation, and profile building and user evaluation. The review analyses the issues of misattribution in publications and software and proposes solutions. The importance of nanopublication and microattribution is also described and reviewed. Additionally, the evaluation metrics for research objects are examined, the proposed research on workflow systems is mentioned, and statistics about the maintenance and execution of tools are presented. The various principles that exist for research objects are discussed, the importance of evaluation metrics in open science is analyzed, and the three different categories of evaluation metrics (traditional, usage, altmetrics) are examined. Finally, the gamification methods for profile building are introduced, the reward systems on Github [18] and Stack Overflow [19] are analyzed, and the user evaluation through virtual rewards is reviewed.

2.1 Search Strategy

In order to find the articles that were relevant with the research purpose of this systematic literature review, a search strategy based on the PRISMA guidelines [20] was conducted. The strategy included electronic database searching, references list checking and citations searching. The main databases on which relevant papers were searched are the PubMed [21], ScienceDirect

[22], ACM Digital Library [23], and SpringerLink [24] databases, and the Google Scholar [25] web search engine.

To be able to keep the most relevant results based on the search string, at each database, the articles in the first one-two pages were collected, as they were sorted by relevance. The search was conducted at the databases by using different keywords relating to misattribution, microattribution, nanopublication, evaluation metrics, research objects, profile building, and user evaluation in bioinformatics. The search diary that was maintained (Table 2.1), details the keywords that were used, the names of the databases that were searched, and the search scope. In order to minimize the chance of missing pertinent studies, the electronic search was augmented with reference list checking and citation searching. For each paper that was collected from the electronic search, its references were checked, and the relevant papers were also collected. Moreover, for the citations searching, the Google Scholar search engine was used to look for papers that cited the already collected articles.

Table 2.1: Search Diary

Search String	Database	Scope
((“misattribution” OR “microattribution” OR “nanopublication”) AND (“repository” OR “citation” OR “publications”) AND (“bioinformatics”)) OR ((“research object” OR “tool” OR “workflow”) AND (“evaluation metrics”)) OR ((“profile building” OR “user evaluation”) AND (“virtual rewards” OR “gamification”) AND (“Q&A system”))	PubMed	Title, Keyword, Abstract
((“misattribution” OR “microattribution” OR “nanopublication”) AND (“repository” OR “citation” OR “publications”) AND (“bioinformatics”)) OR ((“research object” OR “tool” OR “workflow”) AND (“evaluation metrics”)) OR ((“profile building” OR “user evaluation”) AND (“virtual rewards” OR “gamification”) AND (“Q&A system”))	ScienceDirect	Title, Keyword, Abstract
((“misattribution” OR “microattribution” OR “nanopublication”) AND (“repository” OR “citation” OR “publications”) AND (“bioinformatics”)) OR ((“research object” OR “tool” OR “workflow”) AND (“evaluation metrics”)) OR ((“profile building” OR “user evaluation”) AND (“virtual rewards” OR “gamification”) AND (“Q&A system”))	ACM Digital Library	Title, Keyword, Abstract
((“misattribution” OR “microattribution” OR “nanopublication”) AND (“repository” OR “citation” OR “publications”) AND (“bioinformatics”)) OR ((“research object” OR “tool” OR “workflow”) AND (“evaluation metrics”)) OR ((“profile building” OR “user evaluation”) AND (“virtual rewards” OR “gamification”) AND (“Q&A system”))	SpringerLink	Title, Keyword, Abstract
((“misattribution” OR “microattribution” OR “nanopublication”) AND (“repository” OR “citation” OR “publications”) AND (“bioinformatics”)) OR ((“research object” OR “tool” OR “workflow”) AND (“evaluation metrics”)) OR ((“profile building” OR “user evaluation”) AND (“virtual rewards” OR “gamification”) AND (“Q&A system”))	Google Scholar	Title, Keyword, Abstract

2.2 Selection Criteria

In order to be certain that only relevant studies were included in the review, inclusion and exclusion criteria were set on the collected articles [26]. To be considered for the review, the searched articles had to focus on solutions to the stated research questions about misattribution, research object metrics and evaluation, and profile building and user evaluation. The results were restricted to English or Greek language; However, the final selected papers were all written in English. The third inclusion criteria referred to the date of publish. The initial thought was to include articles that were written in the last five years, as so to keep the review up to date with the current studies. However, the relevant articles that were collected were not enough to formulate a comprehensive review and therefore, it was decided to include articles that were published from 2005 to 2022.

Although only articles that were published from 2005 and onward were accepted, if older articles were found important to be analyzed and presented, they would also be accepted as an exception. Those four inclusion criteria were applied by reviewing every article's title, abstract, and keywords. The articles that met the inclusion criteria were obtained and then analyzed as to their content and were evaluated.

2.3 Search Results

The literature research on PubMed, ScienceDirect. ACM Digital Library, Springer, and Google Scholar gave a result of 394 different publications. 125 records were excluded because they were found irrelevant based on their title, keywords or abstract. 52 reports failed to be retrieved and 200 reports were excluded because they failed to meet the inclusion criteria. Therefore, a total of 17 publications from the literature search were deemed eligible for review.

Apart from the literature search, the references of each eligible paper were also examined, and 32 relevant publications were collected. Furthermore, the Google Scholar web search engine was utilized to search for publications that cited the selected articles. From the citation searching 36 articles were collected. Out of the 68 reports, 9 failed to be retrieved and 51 failed to meet the inclusion criteria. Therefore, 8 publications were deemed eligible for review.

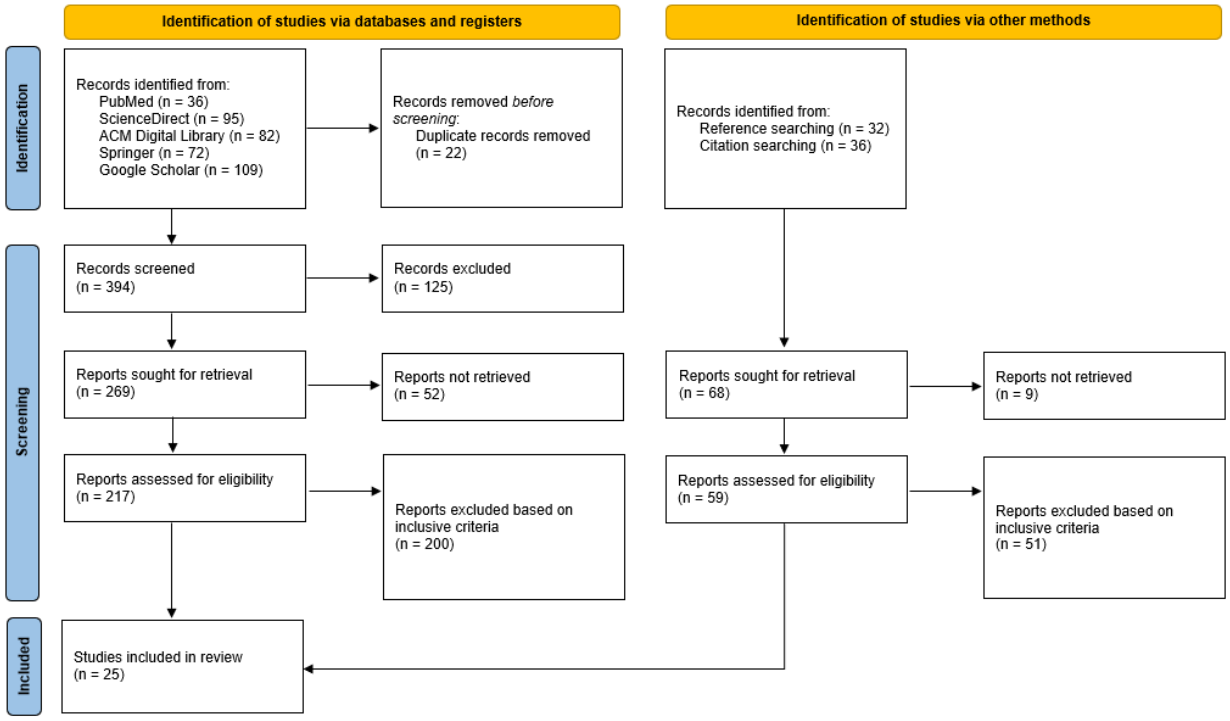

Figure 2.1: The search and screening process based on the PRISMA flow diagram

In order to illustrate the search and screening process, the PRISMA flow diagram was used. The PRISMA flow diagram [20] illustrates the way information moves through the various stages of a systematic review by summarizing the screening process. It displays the number of records found, the number that was included and excluded, as well as the reasons for those exclusions. Figure 2.1 represents the PRISMA flow diagram that was followed for the conduction of this systematic literature review.

For the final review, this method identified 25 publications in total, published between 2004 and 2020. The list of the included studies with their authors, publication year and title are presented below in alphabetical order (Table 2.2).

Authors	Year	Title
P. H. Russell, R. L. Johnson, S. Ananthan, B. Harnke, and N. E. Carlson	2018	A large-scale analysis of bioinformatics code on GitHub
H. Kawashima and H. Tomizawa	2015	Accuracy evaluation of scopus author ID based on the largest funding database in Japan
I. Tahamtan and L. Bornmann	2020	Altmetrics and societal impact measurements: Match or mismatch? a literature review
J. Priem, D. Taraborelli, P. Groth, and C. Neylon	2010	Altmetrics: a manifesto
T. Mutter and D. Kundisch	2014	Behavioral mechanisms prompted by badges: The goal-gradient hypothesis
D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, and I. Rafols	2015	Bibliometrics: The Leiden Manifesto for research metrics
R. C. Gentleman et al.	2004	Bioconductor: open software development for computational biology and bioinformatics
H. Cavusoglu, Z. Li, and K. W. Huang	2015	Can gamification motivate voluntary contributions? The case of StackOverflow Q&A community
S. Mangul et al.	2019	Challenges and recommendations to improve the installability and archival stability of omics computational tools
L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann	2011	Design lessons from the fastest Q&A site in the west
D. Van Dijk, M. Tsagkias, and M. De Rijke	2015	Early detection of topical expertise in community question answering
Z. Lacroix and H. Ménager	2005	Evaluating workflow management systems for bioinformatics
Y. Gil et al.	2007	Examining the challenges of scientific workflows
M. Martone	2014	Joint Declaration of Data Citation Principles
F. Kern, T. Fehlmann, and A. Keller	2020	On the lifetime of bioinformatics web services
N. Immorlica, G. Stoddard, and V. Syrgkanis	2015	Social status and badge design
N. P. Chue Hong et al.	2019	Software Citation Checklist for Authors
N. P. Chue Hong et al.	2019	Software Citation Checklist for Developers

A. M. Smith, D. S. Katz, and K. E. Niemeyer	2016	Software citation principles
D. S. Katz et al.	2016	Software vs. data in the context of citation
A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec	2013	Steering user behavior with badges
B. Giardine et al.	2011	Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach
G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig	2013	Ten Simple Rules for Reproducible Computational Research
Y. Hu, S. Wang, Y. Ren, and K. K. R. Choo	2018	User influence analysis for Github developer social networks
R. Abdalkareem, E. Shihab, and J. Rilling	2017	What Do Developers Use the Crowd For? A Study Using Stack Overflow

Table 2.2: Selected articles included in the review

2.4 Analysis of Literature

2.4.1 Misattribution in Bioinformatics

Although being cited can be extremely important in advancing one's academic and scientific career [27], studies have proved that references in academic literature are commonly inaccurate due to the premise that authors do not check their references or may not even read them, or that credit is often assigned to a review paper rather than the original source of the idea [28]–[31]. This can create the problem of misattribution where an author may not get the proper credit of their contributions to the scientific field. Therefore, it is highly important to find solutions in order to eliminate the incorrect attribution due to ignorance or misguided information.

The conventional methods being used to evaluate a researcher's scientific career may include their publishing track record in international peer-reviewed scientific journals (ISI Impact Factor [32]), the number of citations that each article receives, or the aggregated measure H-index [33] of the total number of citations that each article receives. However, apart from publishing scientific publications, there are additional ways for academics to contribute to the scientific community, one of which is their ability to submit and claim their developed software programs. Linking research data with publications in an open and semantically stable way, has become

important for the improvement of the quality and transparency of academic communication [34]. It is essential for confirming study results and allowing data reuse. The relatively recent practice of citing data begins to solve the long-standing issues that have limited the collective capacity to find data and use it effectively in science. Citations assist the research infrastructure by giving attribution detail, allowing future access, and encouraging cross-collaboration and inquiry, as well as, providing attribution detail, facilitating future access, and promoting cross-collaboration and investigation. Therefore, the scientific community can highly benefit from the widespread use of data citation [35]–[38].

2.4.1.1 Misattribution in Publications

A traditional method to create a linkage between the authors and their publications is the manual labeling by humans. Many studies have been conducted for the evaluation and effectiveness of the algorithms that support that method. Results from those studies showed that manually labeling data can be a difficult task [39]–[42]. Furthermore, labeling selections made by humans may be incorrect, time-consuming or need numerous verification steps [43], [44]. Therefore, manual labeling is frequently insufficient for assessing a clarification task with a high number of name occurrences.

To surpass the weaknesses of the human factor on the manual labeling, researchers developed several algorithms for labeling data. Some studies relied on the published article's additional attributes, such as the co-authors, in order to validate the author [45]. Others, determined the linkage based on the shared e-mail addresses or self-citation [46], [47]. These labeling approaches can determine the linkage of the author and their articles on a vast scale at a short time, yet their correctness is seldom checked. They frequently need negative, like non-matched labels, sets created by heuristic rules, because they are meant to yield positive sets of name instance pairings. For instance, names with distinct strings that do not share co-authors are presumed to correspond to separate authors. An iterative clustering algorithm that triangulates various matching characteristics including co-authors, e-mail addresses, and self-citation has been proposed to solve this problem [48]. However, if those distinguishing qualities are inadequately documented for a particular group of name occurrences, the algorithm's usefulness may be limited.

Other set of studies depended on localized third-party authentication sources to validate the researcher's information. Kawashima H. et al. [49], tested SCOPUS's [50] disambiguation performance on a list of Japanese author names in different publications. They utilized the KAKEN Database [51], which keeps track of a researcher's unique ID number, as well as, a list of their confirmed publications in Japan. By comparing name strings, publication data and affiliations, each KAKEN researcher profile was matched to the author's name occurrence in a SCOPUS-indexed publication and the KAKEN researcher ID was applied to the author's name instance if the match was validated. Similar studies have employed this approach to labeling on Italian and Dutch scholars, utilizing each nation's administrative scholarly databases [52], [53]. NIH ExPORTER [54] and Highly Cited Researchers [55] databases were also used in studies [46], [56]–[58]. These record linking approaches yield large-scale and precise results, but they can produce skewed findings. Names of researchers who are not active in a specified country or discipline, or are not widely referenced, with these methods, they cannot be linked with their publications [57].

Numerous studies have started using ORCID [59] as the globalized third-party authentication source that can link authors with their publications [48], [60], [61]. Various studies provide information on how ORCID could be used for authority control across different repositories and digital libraries [62]–[64]. ORCID is an open platform with millions of researcher profiles that include information about their authorship, education and employment [64]. Like other authority sources, using ORCID can result in large-scale linking data [61]. However, unlike other third-party authentication sources, profiles in ORCID are not restricted to certain disciplines, geographic locations, organizations, or high-profile researchers. Therefore, ORCID can identify the identities of researchers from a wide range of backgrounds, overcoming the weaknesses of existing authentication sources.

2.4.1.2 Misattribution in Software

New knowledge is generated because of research and scholarship. The transmission of this information has a significant influence on how society evolves and progresses, while also feeding back to better future study and scholarship. The internet is altering the way things function in the scientific community, as it does in several other areas of everyday life. It offers up potential for

new procedures that can speed the evolution of learning, including the establishment of new ways to communicate that information among scholars and the wider community. The potential for considerably more effective higher education has been revealed by two decades of emerging and more prevalent information technology.

However, the application of this technology is still restricted as research methods and distribution of research findings have yet to completely embrace the web's and other digital media's possibilities. Producers and consumers are still enamored of forms devised during the period of print publishing, and research reward schemes are still based on old delivery techniques. Although research papers and different publications have long benefited from an architecture that makes it simple to cite them, the same values do not apply to software data. During the last decade, the scientific community and open science required the increment of accessibility, openness, and reproducibility, as well as the support of FAIR-aligned data repositories. The concept of FAIR data is a set of criteria that assure that data are findable, accessible, interoperable and reusable [65]. To support the full understanding and distribution of research, software is just as important as a journal article. Citing is an important part of the research and academic discourse process in all disciplines. Therefore, just like other papers or books, software should also be mentioned and referenced.

FORCE11 [66] is a group of academics, publishers and research funders that worked together in order to promote the transition toward better knowledge production and sharing. Their aim is to revolutionize academic communication via the integration of information technology, both individually and collectively. The digital publishing of papers facilitates efficient scholarly collaboration, which can include stronger data connections, the dissemination of software applications, mathematical models, protocols, workflows, and research interaction via social media channels [67]. They support that software should be cited in the same way as a paper is referenced in a journal article and therefore, have developed guidance for software citation.

In 2016 [68], a working group aimed to establish a unified set of citation guidelines based on a survey of existing community standards in order to encourage widespread adoption of a general protocol for software reference across domains and platforms. Their work was presented as a collection of software citation standards and reviews about the motives for generating the principles, the existing community procedure, and the obligations these principles would impose on various stakeholders. Their work was motivated due to prior research [69] that presented a list

of guidelines for data in academic literature or research object in order to support and encourage ethical behavior. These Data Citation Principles addressed the aim, function, and qualities of citations. They emphasized the need of developing citation methods that would be both intelligible by humans and usable by machines.

Since software can be an executive tool that operates on data, another study [70] provided instances and references to results in the form of citations to highlight the distinctions among software and data in the context of citation. They recognized that software, while comparable to data in regards of not being typically mentioned in publications, is nevertheless distinct from data. The term "data" is used in research to refer to electronic recordings of observations made during a research project (raw data), information obtained from such observations via some type of processing (processed data), and the result of modeling software (simulated data). The term "data" has a broader meaning in electronics and information systems, because it refers to everything that can be processed by a computer and that contributes to the confusion over the separation between software and data.

Future research [71], provided a standard checklist that authors of academic work, such as papers, books, conference abstracts, etc., could use to ensure that they are referencing and citing software they have used, whether they have created it for their research or they obtained it from other sources, according to best practices. Furthermore, the same research could also be used as the foundation for more specific instructions for writers and reviewers by journal editors, publishers, or conference chairs. On the same note, the "Software Citation Checklist for Developers" [72] provided a simple, standard checklist that software developers (both open source and closed source) may use to ensure the adoption of best practices when it comes to software citation. This would aid developers in receiving credit for their work, as well as promote transparency, repeatability, and reuse.

Citing software enhances research by allowing other researchers to access it in order to support proper citation and credit, enable peer-review, affirmation and reproducibility of findings, support association and reuse and promote building on the work of others [73]. Software citation raises software to the status of a first-class object in the academic environment, reflecting its current importance.

2.4.1.3 Nanopublication and Microattribution

A nanopublication does not have to be linked to a complete scientific paper (though that is also a possibility), but it might instead refer to a repository from which it was produced. Furthermore, nanopublications can be represented in Resource Description Framework (RDF) and shared across computers using the Extensive Markup Language (XML). Nanopublications may now be processed, searched, and retrieved easier through the Internet, as well as subjected to computer reasoning, which is not possible with conventional papers [74]. Most crucially, because nanopublications may be acknowledged and cited in the same manner as conventional papers, they could motivate potential data contributors to publish their data in the public domain for everyone to openly access and exploit [75]. This could have a big influence on future academic publication methods since it makes data mining and data sharing easier, which raises the chances of a publication being found and referenced.

Microattribution can be used for the academic acknowledgement of a specific author's small contributions. The objective of microattribution is to create a publication procedure that is available to all journals and that depends on the knowledge of all individuals with a commitment in the unambiguous linkage of data to their contributors via a unique identifier [76]. Many scientific fields allow for data sets to be compiled from multiple sources. It may appear that listing every inventor is inconvenient, but technological advancements can alleviate this view and enable equal attribution methods. While in the past these sources were simply ignored in the academic contribution, today a table detailing the data and source serves as a contribution and is provided as supplemental data. Microattributions can also be important for giving credit to nanopublications. The components of assertion and attribution can be used to attribute these sorts of publications. In 2011, Giardine B. et al. [77], concluded that microattribution significantly increased the reporting of different human variations, resulting in a complete online resource for comprehensively characterizing human genetic variation.

Contributions to the construction and preservation of collected data are frequently overlooked. This is due to a lack of formal data citation processes, as well as the low precision of several contributions and the difficulty of distinguishing author names and identities [78]. Recent approaches to this problem include internet identification frameworks such as OpenID [79], name-authority databases such as VIAF [80], scholar and professional profile networks such as

ResearchGate [81], and bottom-up and top-down researcher identifier databases such as ORCID [59] and ISNI [82].

ISNI is an ISO-certified worldwide standard number (ISO 27729:2012) [83] for distinguishing the millions of people who contribute to creative works and those who are involved in their dissemination, such as researchers, inventors, authors, artists, visual designers, entertainers, producers, publishers, aggregators, and others. It is member of a community of international standard identifiers, such as DOI, ISBN, ISSN, etc. that contain identifiers for works, recordings, products, and copyright owners in all contexts. The ISNI International Agency's mission is to assign a persistent unique identifying number to the name of a researcher, inventor, writer, artist, performer, publisher, etc. in order to solve the problem of name confusion in search and discovery; and to spread each designated ISNI across all forms in the global market so every published work can be unequivocally attributed to its creator wherever that research is mentioned. By meeting these objectives, the ISNI will serve as a cross-domain identifier and a major element in Linked Data and Semantic Web services.

ORCID built a database for researchers aiming to give a solution to the name - ambiguity problem in academic research by providing each researcher with a unique ID. It has created an author self-registration service, as well as an author claim system for publications, and is working with participants such as publishers, research organizations, and funders to connect digital research to this database [84]. Most major scientific publications, CrossRef [85], and the Wellcome Trust [86] are among the adopters. Such identifier systems could benefit microattribution, because it can allow researchers to keep track of their advancements in science in real time and exceed the conventional publication list [87].

2.4.2 Evaluation Metrics for Research Objects

Due to the introduction of new technology and lower costs related with sequencing equipment, the availability and creation of new data has expanded significantly over the previous decade. One of the major problems in biomedical research is the integrated analysis of massive and increasingly complex sets of data in order to comprehend biological mechanisms, such as the genesis and course of human illnesses. To detect and interpret the biological insights accessible from this amount of data, it is critical to have efficient and useable sets of research objects, such

as tools and workflows. Usability, according to the International Organization for Standardization (ISO), is the degree to which a product can be used by specific people to achieve specific goals with efficiency, effectiveness, and satisfaction in a specific context of use (ISO 9241-110:2020) [88]. Therefore, research requires computer-assisted technique in order to bring new solutions for maintaining scientific quality standards and repeatability.

2.4.2.1 Research on Workflow Systems

In bioinformatics, the workflow concept is gathering steam as the preferred method of capturing the phases of computer experiments [89]–[92]. Using workflow formulation and implementation tools like Taverna [93] and Galaxy [3], as well as workflow sharing platforms like myExperiment [94] and CrowdLabs [95], it allows scientists to outline the phases of a complicated analysis and expose them to peers. Data outputs are created from data inputs in a standard workflow by a group of (possibly distributed) computational operations that are coordinated according to a workflow design. Workflows, on the other hand, do not give a complete solution for gathering all data and meta-data required to fully comprehend the environment of an operation. As a result, sometimes it can be difficult (or impossible) for scientists to reuse or adapt existing procedures for their own research. A recent report [96] of Taverna workflows on myExperiment identified inadequate meta-data as one of the primary reasons of workflow degradation, because they were unable to execute it again after it was created.

Table 2.3: Ten rules for reproducibility of computational research, derived from Sandve G.K. et al. [97]

Rule #	Rule
1	For Every Result, Keep Track of How It Was Produced
2	Avoid Manual Data Manipulation Steps
3	Archive the Exact Versions of All External Programs Used
4	Version Control All Custom Scripts
5	Record All Intermediate Results, When Possible, in Standardized Formats
6	For Analyses That Include Randomness, Note Underlying Random Seeds
7	Always Store Raw Data behind Plots
8	Generate Hierarchical Analysis Output, Allowing Layers of Increasing Detail to Be Inspected

9	Connect Textual Statements to Underlying Results
10	Provide Public Access to Scripts, Runs, and Results

A few research works [98], [99] discussed the issue of study reproducibility. Reproducibility criticality impacts a wide range of scientific domains to varying degrees [98]. Given the importance of bioinformatics in several biological and biomedical investigations nowadays, substantial effort must be made to make computational analyses repeatable [97], [100], [101]. The short service of bioinformatics software, the intricacy of pipelines, the uncontrollable effects generated by changes in system frameworks, the inconsistent data or inaccuracy in workflow specification, and other factors may all contribute to reproducibility concerns in bioinformatics. Sandve G.K. et al. [97] proposed ten good practice criteria for the creation and use of a computational workflow aiming to address reproducibility difficulties in bioinformatics (Table 2.3). The Bioconductor [102] project, that offers version control for a significant number of genomics/bioinformatics packages, complies with several of Sandve's principles. Users can access previous versions of any Bioconductor package in this way. Nevertheless, Bioconductor does not support all phases of any bioinformatics process; for instance, in an RNAseq workflow, fastq trimming and alignment are often executed with tools that are not available in Bioconductor [103]. Both professional and open-source cloud systems, such as BaseSpace [104] and Galaxy [105], also substantially fulfill Sandve's objectives. Furthermore, in such systems, procedures cannot be significantly changed; for example, BaseSpace [106] has tight restrictions for application submission. Additionally, cloud applications must address legal and ethical concerns [107].

2.4.2.2 Maintenance and Execution of Tools

With the mass production of scientific studies and publications accompanied by the development of bioinformatic software tools little attention is being given to the importance of maintaining bioinformatic tools. Scientific journals require data and code sharing, but none currently require authors to guarantee the continuing functionality of newly published tools. Multiple studies have identified the deterioration of long-term archival stability of published software tools [108]–[112].

Russell P. et al. [113] investigated the state of source code in the bioinformatics community using Github repositories. In order to find such repositories to investigate, they had to search public

journal articles in bioinformatics topics and look through the high volume of posts about certain projects on the online forum of Biostars [114]. The result was 1720 repositories based on the journal articles, and 23 by Biostars. The conducted research showed that mostly the larger development teams tend to revisit the code and keep it maintained, which also suggests that users tend to choose that software. In another research, Mangul S. et al. [115] estimated the archival stability of computational biology software tools by performing an empirical analysis of the internet presence for 36702 omics software resources published from 2005 to 2017. They found that almost 28% of all resources were not accessible through URLs published in the paper they first appeared in. Lastly, Kern F. et al. [116] provided an up-to-date and comprehensive evaluation of the general availability of web services, by collecting 2727 articles describing 2396 unique tools published by PubMed indexed journals from 2010 onwards and testing their availability over time. Their results analysis generalized the results of the Schultheiss study [117] described 10 years ago, as they found that of the 2396 tools extracted from 2767 articles published in PubMed between 2010 - 2020, 74.3% were still working, whereas 25.7% had gone offline.

Moreover, besides the troubling maintenance of tools, testing them for their installation and general usability seems to be another difficulty. Mangul S. et al. [115] selected 98 software tools to test for their installation usability, which resulted in 51% being deemed “easy to install”, and 28% of the tools having failed to be installed at all due to problems in the implementation. Among those problems are the license restrictions that come with the use of the created code. As Russell P. et al. [113] mentioned, many repositories, despite being made public on GitHub, do not feature explicit licenses, which results in restricting the rights of others to reuse and modify their code. Furthermore, a common wonder is the number of accessible tools compared to the published ones. Kern F. et al. [116] found that of the 2396 tools, only 40.5% is listed in bio.tools [118], one of the largest known sites when it comes to bioinformatics software tools, and 59.5% is not. Most importantly, 3.5% of the listed tools are not accessible by users.

2.4.2.3 Principles for Research Object Systems

There are several software solutions that may be classified as scientific workflow management systems, each with its own set of features. The applications and technology in this field are quite diverse. These systems have many properties, such as being adapted to support

specific processes, or being scientific or commercial [119], [120]. As a result, while choosing software, a scientist must determine which satisfies his requirements the best. To determine how well a piece of software matches a user's demands, related software must be found, and the advantages offered by each must be evaluated. Because of the range and complexity of the existing alternatives, as well as the fact that each user has their own unique demands, comparing various systems can be a difficult process. Therefore, each of the potential criteria should be given a different level of priority.

A set of criteria for assessing various bioinformatics workflow management applications were proposed by various studies [121], [122]. These criteria ought to be simple to evaluate and relate to the software category's requirements. The requirements they suggested included: the description of the workflow's definition; the completion of the prescribed tasks in accordance with the workflow's coordination characteristics; the insurance that the outcomes and effects are accessible; the provision of a secure multi-user framework, which is essential in any commercial software system; the provision of data gathering tools that allow scientists to take advantage of the abundance of resources available; the provision of high processing capabilities that could match the data-intensive processes performed; the maintenance of data traceability and, as a result, establishment of repeatability in a highly dispersed and dynamic context; and the provision of easier system use by maintaining a high level of transparency between workflow design and execution, automatically resolving translation and interoperability difficulties between the many databases and tools utilized.

2.4.2.4 Evaluation Metrics in Open Science

Even though open science has widespread acceptance throughout scientific and technological organizations, institutional and cultural hurdles remain. The existing structure of scientific research across disciplines is insufficiently favorable to information sharing between professions and among researchers and the general public. The way colleges are divided into departments and separated from society might make information, expertise, and data more difficult to obtain. Local initiatives to promote open science may be limited by an underinvestment in education and data systems, and therefore, open data and open access publication may be affected by fund limitations and the use of private data and tools in research [123]–[125]. The institutional

systems of academic research, which frequently fail to recognize, appreciate, and reward attempts to open-up the scientific procedure, are one of the most major impediments to the objectives of open science. As a result, if researchers embrace innovative methods of working and publishing rather than conforming to old procedures, their career development may be inhibited [126]–[128]. In any research system, especially an open one, evaluation metrics can be critical. It is crucial to recognize however, the inherent limitations of them as they can be employed in ways that could slow down the development rather than speed it up.

Some studies claim that metrics contribute to colleges' increasingly bureaucratic and control-driven cultures [129], [130]. Three main problems are frequently raised. First, managers and funders are focusing their attention on items that can be assessed rather than those that cannot. Second, there is a decline in variety when a concentration on certain metrics or ranking tables leads universities to embrace similar strategic goals, and individual academics focus on reduced, progressive work directed at larger publications [126]. Third, there is a misalignment of incentives, which exacerbates issues with research quality, integrity, and reproducibility [131], [132].

However, metrics help in keeping track of the scientific system's progress toward openness at all levels, and in measuring performance, in order to recognize and reward better methods of functioning in groups and individuals. Such objectives need the creation of new indicators as well as the more responsible application of current measurements. Several projects to solve these concerns have been made, including DORA [133], which urged for research to be evaluated on its own qualities and for journal impact factors to be eliminated from funding, hiring, and promotion choices; the Leiden Manifesto [126], which lays out ten principles for using key measures in research evaluation; The Metric Tide [127], which established a framework and specific suggestions for appropriate metrics by conducting an objective review of the use of metrics in scientific evaluation and monitoring.

2.4.2.5 Traditional Evaluation Metrics

Research papers, primarily journal publications, are measured using traditional measures. The number of publications and the citation count an article obtains are the two most fundamental types of metrics. These variables can be combined at various levels. A single article, a scholar, a research unit, or an organization can all be evaluated.

More advanced metrics have been developed from the data of publication and citation counts including the H-index [134], the Journal Impact Factor (JIF) [135], the SCImago Journal Rank indicator (SJR) [136], source normalized citation indicators [137], the Eigenfactor [138], the Source Normalized Impact Per Paper (SNIP) [139], and CiteScore [140]. When items from various fields or aggregation levels are compared, it is necessary to standardize indicators since publication and citation practices differ substantially. Several studies complain that the JIF is misused as a measure of an article's impact. JIF is an average for the whole journal, and therefore does not truly reflect the citation significance of different articles. However, neither the Leiden Manifesto nor the Metric Tide dismiss the JIF or other metrics outright; rather, they recommend that they should be utilized appropriately [141]–[144]. Although the h-index has been criticized for biasing towards older academics and failing to capture the influence of highly cited papers, it is still commonly utilized. Bibliometricians suggest that it is critical not to depend just on numbers and indicators, but to combine them with qualitative evaluation of the object of evaluation [126], [145]–[148].

2.4.2.6 Usage Evaluation Metrics

Usage metrics are frequently generated by the count of views or downloads of an object. Because many users, such as students, policymakers, and the general public, read articles or utilize data without ever publishing, usage metrics differ from citations [149], [150]. Furthermore, even though a study may be included in a paper, it might not get acknowledged [151]. Attention and acceptance are evaluated with usage metrics, such as the usage impact factor [152] or libcitations [153].

Usage metrics are extremely important in open science, not just for measuring traditional publishes (posts, blogs), but also for the re-use of open data and software. Open access and commercial publishers give usage information for different articles. Several organizations, such as ACM [23], provide the number of downloads of a particular article from their platform. In collaboration with Mendeley [154], Elsevier's ScienceDirect [22] offers researchers the number of downloads of their articles on the ScienceDirect platform.

2.4.2.7 Altmetrics

Altmetrics were created to cover further areas of influence that traditional and usage metrics do not address. With the creation of the web technologies, new opportunities for evaluating the effect of research papers have emerged, including published journals, books, reports, data, and other non-traditional publication types. Altmetrics have evolved into a tool for assessing the sociocultural effect of scientific research [155]. In 2009 [156], the term "article-level metrics" was proposed, whereas in 2010 [157], the word "Scientometrics 2.0" was introduced. Priem J. et al. [158], further explored and refined the idea of alternate ways to measure academic activities, and the term "altmetrics" was used as an abbreviation for alternative metrics. Blogs, Twitter [159], ResearchGate [81], and Mendeley [154] are among the most popular social media applications for Altmetrics.

Using altmetric signals has numerous benefits including that they are quick in comparison to citations; they cover a wide range of objects, including datasets, code, exploratory research, nanopublications, blog entries, comments, and tweets; and they are diverse, because they provide a variety of indicators for the same object, such as downloads, likes, and comments [158]. Although they are typically thought of as purely quantitative measures, they also allow for qualitative examination of academics and beneficiaries, such as by analyzing the content of user profiles or comments [155]. As a result, they can contribute to the guide of qualitative analyses, expanding our knowledge on the significance of study results.

2.4.3 Gamification for Profile Building

Several online Q&A sites exist to exchange knowledge for a variety of disciplines ranging from technology to art. Because the maintenance of such platforms is dependent on the contributions of its users, those sites' creators pay attention to them and look for methods to reward them by building and expanding their profiles. Therefore, to encourage and reward user engagement, many systems employ gamification methods. Gamification is the application of game concepts and tactics to non-game settings to encourage and attract people in a variety of tasks. It recognizes the accomplishments of users, it motivates them as they go through stages, and ultimately engages them mentally in the intended behavior [160]. The gamification methods may include reputation score, upvotes, downvotes, rewards, and badges, that assist the Q&A platforms

in attracting new users, and ensure their long-term viability, while also motivating users to continue learning [161], [162]. Several developer social network platforms have adopted those methods including Github [18] and Bitbucket [163], where software developers that register up for an account, connect with other users and share information, such as making contributions to software application projects or repositories, or answering questions.

2.4.3.1 Reward System on Github

Github is among the most widely used social networking platforms because of its features and usefulness. Users may quickly browse through vast quantities of code with Git, fork content from other members, and create project branches. The vast volume of data generated by its social development activities can be mined for social and collaborative aspects [164], [165]. It offers social network capabilities such as following people, starring, and forking projects in addition to code hosting and maintenance. Users may use Github to follow different users, star, fork, commit or create projects, open issues, and make pull requests. The user's potential influence is transformed into real effect through those activities. More active engagement implies that a person has a higher degree of interest and contribution, and regular interaction can improve mutual trust. These are important qualities of social power. Furthermore, the likelihood of a user being recognized by others is proportional to his or her degree of activity. For example, an active social media person is more likely to be recognized by other users. As a result of activities, the user's online impact increases. Hu Y. et al. [166], conducted a study to analyze the data that describe the users' sociability on Github. The results showed that changes in the number of followers may indicate a user's impact over time, as users with a lot of influence have a greater effect on the platform.

2.4.3.2 Reward System of Stack Overflow

The technical Q&A platform Stack Overflow [19], which is part of the Q&A community, has piqued the interest of developers and has risen to the top of the Stack Exchange network. Developers get a social purpose on the platform by submitting questions or reading and responding to them. Users engage with each other in the community and create crowd-sourced content through

their posting behaviors [167]. Furthermore, posting allows users to improve their programming abilities and expertise. Users can improve or acquire new skills by engaging with other coders in the community, as the complexity and knowledge level of queries varies. Furthermore, they can build a programming expertise portfolio because of their social engagement and involvement to aid their job research [168]. In general, the human component in software development is one of the most important factors impacting the industry's long-term viability and evolution [169].

Because users contribute without receiving financial reward, it is difficult for online communities to motivate members to engage effectively in them. However, the Stack Overflow platform managed to overcome those problems by creating distinctive participation features, and a reputation system, which is based on social approval and balances the quality and quantity of contribution. Its members review each other's contributions, and the marks of their accomplishments include points, badges, and levels for quality engagement. Stack Overflow used gamification concepts to utilize the potential of the public for meaningful participation. Cavusoglu et al. [170] conducted an empirical study to research the way awarded badges on the Stack Overflow platform can motivate users. Consequently, their findings indicated that users who got badges for giving answers were more likely to answer even more questions.

2.4.3.3 User Evaluation through Virtual Rewards

Stack Overflow has grown to a vast social platform where knowledge seekers and suppliers of all levels and skills connect to solve programming problems [171]. It altered the way developers learn, interact, and collaborate to create information repositories for future use [172]–[174]. It became a vital element of the software development system and developers are increasingly relying on it for their everyday programming needs. Furthermore, users on other platforms, such as Github, actively urge their members to look for answers on Stack Overflow [175]. The platform’s users may discover several high-quality solutions for multiple programming languages, tools, frameworks, or software [176]. Moreover, if what they require is not accessible, they can publish a post and obtain responses almost immediately [177]. Also, users are encouraged to participate to obtain virtual awards, such as reputation points and badges [178]. Most importantly, by gaining those virtual awards, they can demonstrate their knowledge to potential recruits using their profile page that lists their contributions and achievements [179]. As a result, other users and

recruiters can form impressions about their expertise of topics, their programming abilities, skills, and experience. In such a highly competitive environment, users that stand out are those that successfully acquire visible traces to attract attention. One such significant way for users to stand out is by acquiring many reputation points and badges. Consequently, fellow users and recruiters may create opinions about their expertise, programming skills, capabilities, and knowledge. Users who effectively collect visual traces to draw attention are those who stand out in such a competitive setting [180], [181]. Obtaining a big number of reputation points and badges is one important method for users to identify themselves.

Despite having no stated value, virtual awards serve as social status indicators. Some badges are difficult to get because they demand a lot of work. These have a greater prestige value since they differentiate community members. Others are easy to get and serve as motivators and learning tools. On crowdsourced systems, badges fulfill a variety of socio-psychological roles [182]. The function of badges in reward systems has been a frequent topic in research [183]–[186]. According to Immorlica N. et al. [187], the best design utilizes predefined badges, which are awarded only to individuals who have made a certain number of contributions, and to assess the efficacy of such badge systems.

Virtual rewards have been studied qualitatively and quantitatively in a variety of scenarios, including open-source software and data libraries [175], [188], [189]. Anderson A. et al. [190] provided a mathematical model for predicting how badges influence user behavior. Mutter T. et al. [191] proved that when users are close to earning a badge, their participation level increases. First-time badges, which are granted after a user performs a certain activity for the first time, have an indirect effect on user behavior and improves the platform's functionality [192]. Therefore, reputation points and community activity patterns could be strong indicators of a system's long-term worth and a user's evaluation of contribution and expertise on the platform [171].

2.5 Findings

The problem of misattribution has been widely addressed and researchers have underlined the problems that result from not acknowledging an author's contribution to scientific fields, especially, bioinformatics. Manual labeling by humans is a traditional method that creates a linkage between the authors and their publication. However, due to human-errors and time-

consuming processes, researchers tried to surpass those weaknesses by developing algorithms for labeling data. Although fast, those algorithms showed limitations when distinguishing qualities were inadequately documented. In an effort to overcome the limitations of those algorithms, several studies depended on localized or globalized third-party authentication sources to validate each researcher's information. Among many authenticators the open platform of ORCID was most widely used. Various studies provided information on how ORCID could be used for authority control across different repositories and digital libraries. Therefore, ORCID could identify the identities of researchers from a wide range of backgrounds, overcoming the weaknesses of existing authentication sources.

Apart from publishing scientific articles, there are additional ways for academics to help the scientific community, one of which is their ability to submit and claim their developed software programs. Linking research data with publications in an open and semantically stable way, became important for the improvement of the quality and transparency of academic communication. The scientific community and open science required the increment of accessibility, openness, and reproducibility, as well as the support of FAIR-aligned data repositories. Several studies supported that software should be cited in the same way as a paper is referenced in a journal article, and therefore have published guidelines, principles, and checklists for software citation. It has been proved that citing software enhances research, therefore, allowing other researchers to access proper citation and credit, can enable peer-review, affirmation and reproducibility of findings, association and reuse, and promote building on the work of others.

Nanopublication and microattribution are terms that have also been used by researchers in the limitations of supporting authors' contributions to the scientific fields. As research shows, nanopublications should be acknowledged and cited in the same manner as conventional papers, because they could motivate potential data contributors to publish their data in the public domain for everyone to openly access and exploit. Open-access journals, such as Scientific Data [193], or JMIR Data [194] have acknowledged, adopted, and shared the importance of publishing open datasets for analysis and reuse of scientific data. This could have a big influence on future academic publication methods since it could make data mining and data sharing easier and raise the chances of a publication being found and referenced. Microattribution can be used for the academic acknowledgement of a specific author's small contributions. The objective of microattribution is to create a publication procedure that is available to all journals and that

depends on the knowledge of all individuals with a commitment in the unambiguous linkage of data to their contributors via a unique identifier. These technological advancements can enable equal attribution methods and therefore, contributions to the construction and preservation of collected data should not be overlooked. Identifier systems, such as ORCID, could benefit microattribution and nanopublication, because it could allow researchers to keep track of their advancements in science in real time and exceed the conventional publication list.

The integrated analysis of massive and increasingly complicated sets of data is difficult in biomedical research. It's vital to have efficient and usable sets of research objects, such as tools and workflows, to discover and analyze the biological insights available from this massive amount of data. In bioinformatics, the workflow concept is gathering steam as the preferred method of capturing the phases of computer experiments. Using workflow formulation and implementation, several tools and platforms allow scientists to outline the phases of a complicated analysis and expose them to peers. Given the importance of bioinformatics in several biological and biomedical investigations, substantial effort must be made to make computational analyses repeatable and workflow systems and tools maintainable. A scientist should have the option to determine which research object will satisfy their requirements. Various studies have proposed a set of criteria for assessing various bioinformatics workflow management applications. Those criteria, among others, included the insurance that the outcomes and effects are accessible; the provision of high processing capabilities that could match the data-intensive processes performed; the maintenance of data traceability and, as a result, establishment of repeatability in a highly dispersed and dynamic context; and the provision of easier system use by maintaining a high level of transparency between workflow design and execution, automatically resolving translation and interoperability difficulties between the many databases and tools utilized.

Despite growing support of open science and research across scientific and technology institutions, institutional and cultural barriers still exist. The current framework of scientific research across disciplines is not conducive to information exchange within professionals, as well as between researchers and the general public. Evaluation metrics can help in keeping track of the scientific system's progress toward openness at all levels, and in measuring performance, in order to recognize and reward better methods of functioning in groups and individuals. Three types of evaluation metrics exist. The most fundamental traditional metrics measure the number of publications and the citation count that an article obtains. The usage metrics are frequently

generated by the count of views or downloads of an object. Altmetrics assess the sociocultural effect of scientific research as they cover a wide range of objects, including datasets, code, exploratory research, nanopublications, blog entries, comments, and tweets. All these types of metrics can contribute to the guide of qualitative analyses, expanding our knowledge on the significance of study results and research objects.

Several online Q&A sites can be researched in order to better understand the expanding and profile building of a user. Because the maintenance of such platforms is dependent on the contributions of its users, those sites' creators pay attention to them and look for methods to reward them through gamification methods. Github is among the most widely used social networking platforms because of its features and usefulness. Users can use Github to follow different users, star, fork, commit or create projects, open issues, and make pull requests. The user's potential influence is transformed into real effect through those activities. Stack Overflow gives developers a social purpose on the platform by having them submit questions or read and respond to them. By creating distinctive participation features and a reputation system, which is based on social approval, the Stack Overflow platform managed to keep its users engaged and awards them with popularity points and badges. Those virtual badges, despite having no stated value, they serve as social status indicators. Different awards have different prestige values, and they can differentiate community members. Many studies proved that reputation points and community activity patterns can be strong indicators of a system's long-term worth and a user's evaluation of contribution and expertise on a platform.

3 Methodological Approaches

Although the big data revolution has drawn attention to several bioinformatics databases and tools, using them successfully frequently necessitates specialized knowledge. Many organizations lack the necessary bioinformatics competence, and they commonly discover that the software's documentation is insufficient while their local counterparts can be overworked or uninitiated with certain programs. Such issues frequently result in data analysis bottlenecks that impede the advancement of biological research. This section presents different platforms that share similarity to the OpenBio project and explores the proposed solutions they incorporated in their systems.

3.1 The Galaxy Platform

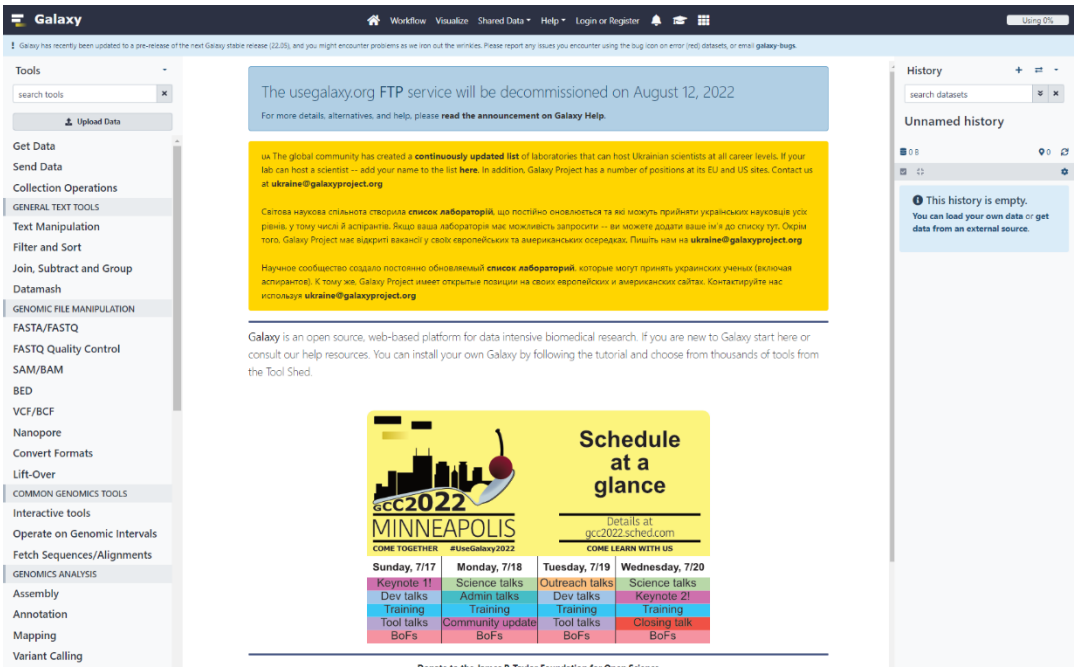

Figure 3.1: The Galaxy platform

Galaxy [195] is a web-based scientific research platform that scientists use to study huge biomedical datasets like those found in genomes, proteomics, metabolomics, and imaging. Galaxy, which was founded in 2005, continues to focus on three fundamental issues in data-driven biomedical science including the accessibility of analyses to all scientists, ensuring analyses are entirely repeatable, and making it easy to communicate analyses so they can be reused and

extended. The Galaxy team and the open-source community behind Galaxy have made significant changes to Galaxy's fundamental framework, user interface, tools, and training materials. Galaxy can now analyze hundreds or thousands of datasets due to improvements to its framework and user interface. The Galaxy ToolShed currently has over 7500 tools [196]. Its community has spearheaded a project to provide several high-quality tutorials on popular genomic analysis. Galaxy's developers and user communities are continuously growing and play an important role in the company's development. The number of Galaxy public servers, the developers who contribute to the Galaxy framework and tools, and users of the main Galaxy server have all grown considerably.

The Galaxy software ecosystem is made up of several different parts, including an integrated repository of tools for a variety of biomedical studies, a web application that allows data analysis through using integrated tools via a web interface, a large number of customized installations of the web application, a training network that offers tutorials and provides workshops on using Galaxy for various studies, and a diverse and inclusive community of users [197].

Through its community hub, the Galaxy provides updates and news. The Galaxy Community offers support for many active regional communities that write programs for the software framework depending on their requirements, manage Galaxy instances for their users, and host localized meetings and training courses. Although each user can rate a research object, they cannot open a discussion, raise questions, offer solutions, or mention possible issues that object may have. Furthermore, the Galaxy project lacks the ability to evaluate users based on their contribution to the platform. Moreover, it cannot semantically connect users with their accomplishments.

3.2 The Bio.tools Website

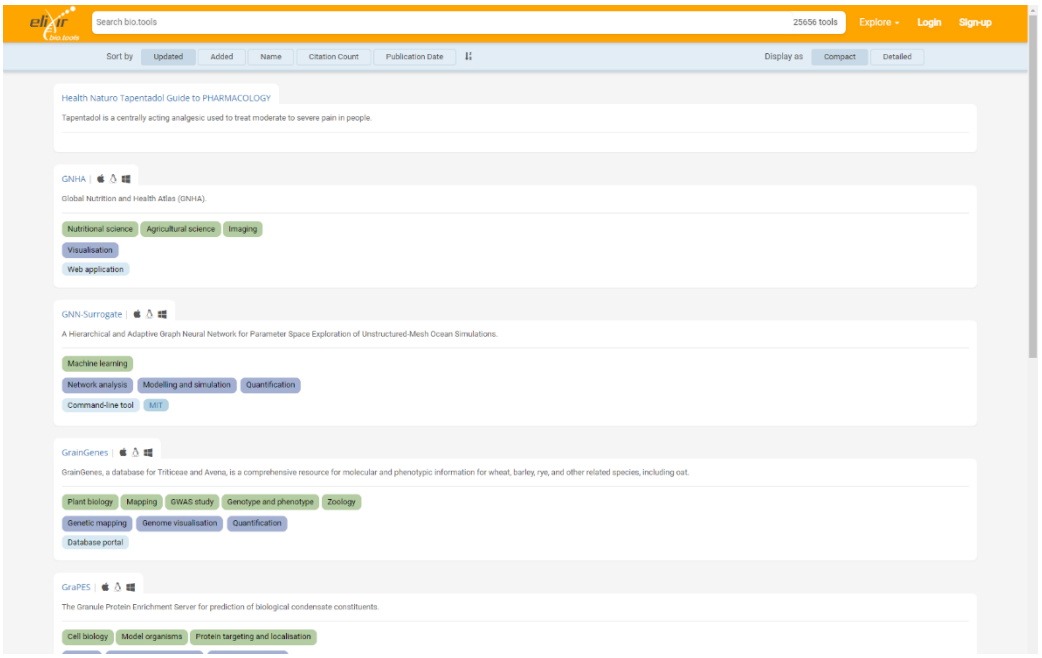

Figure 3.2: The bio.tools web site

Bio.tools [198], a web site offered by the European Infrastructure for Biological Information ELIXIR [199], can help with the discovery of bioinformatics resources, such as software programs, web applications, and database sites. Users can find and review materials using its graphical user interface. Providers of bioinformatics resources could use bio.tools to increase the visibility of their services. Although it may lack the power for a user to evaluate research objects, they provide the ability to identify and connect a user with their software and publications.

Over the last years, bio.tools has progressively grown to contain great quantity of citations and sources [200]. Globally, all life science areas are within the scope of application software of any kind. From straightforward command-line tools and Web apps to databases, workflows, and integrated workbenches, everything falls under this category. Most articles discuss open source or publicly available tools with simple features that may be combined into useful workflows. An exclusive tool identifier, which is a carefully checked, URL-safe variant of the specified tool name, is given to each access. The tool IDs offer a practical way to cite and track software when accompanied by a version label given by a developer, given the complete lack of a standard publication. The IDs can be used in permanent bio.tools URLs that resolve to informational Tool Cards. Only the basic information is required by bio.tools (name, brief description, and webpage),

but it supports extensive descriptions of 50 key scientific, technological, and administrative aspects [14]. Resources must adhere to strict semantic and syntactic requirements that are outlined in the documented schema known as biotoolsSchema [201]. For the user's convenience, controlled vocabularies are widely used and give information that is clear, dependable, and hence comparable. In the EDAM ontology, for instance, tools may be labeled with subjects, procedures, input and output types of data, and compatible formats [202]. Wherever feasible, standard identifiers are utilized, such as DOIs for publications, and descriptive content, such as manuals or citation guidelines, are referred by URL. As a result, the difficulty of bioinformatics software is simplified to groups of easily comprehendible functional units that are placed in a scientific and technological context and include data that facilitates access and usage. Therefore, end-users can benefit from the platform's collection and standardization of data in a plethora of different ways.

3.3 The Datasets2Tools Platform

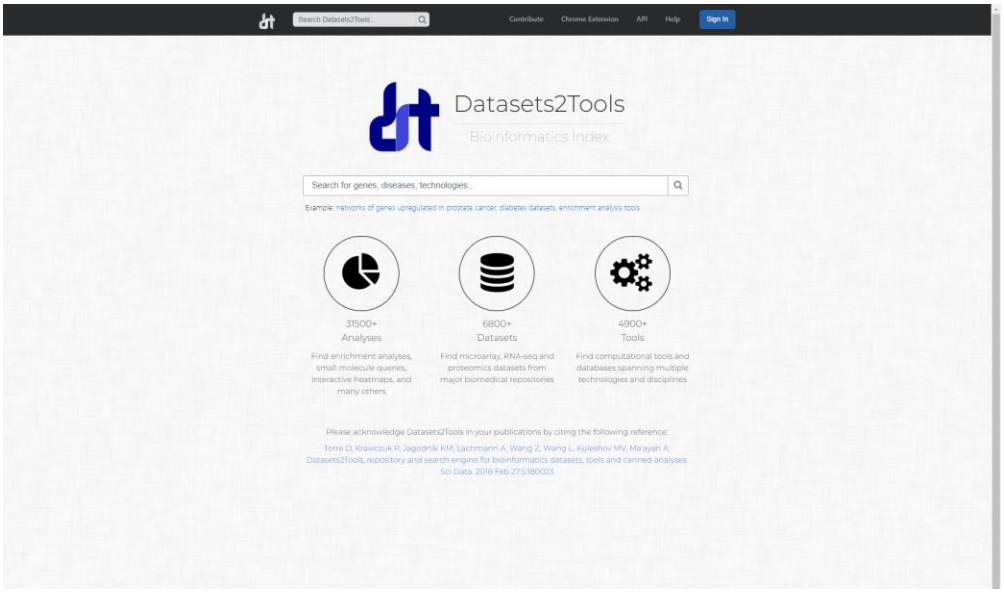

Figure 3.3: The Datasets2Tools platform

Datasets2Tools [203] is a detailed index and platform for the exploration and assessment of a preliminary collection of numerous canned bioinformatics analyses performed to datasets by carefully chosen tools that give durable canned analysis URLs. Several software tools have been indexed by Datasets2Tools and published in prominent bioinformatics journals. Users can quickly locate pertinent datasets, tools, and digital objects with Datasets2Tools. Additionally,

Datasets2Tools is provided as an Application Programming Interface (API) and a Google Chrome browser plugin.

The ability to assess datasets, tools, and canned analyses according to their adherence to the discoverable, accessible, interoperable, and reproducible (FAIR) principles is one special feature of Datasets2Tools. Datasets2Tools has three different sorts of digital items: datasets, tools, and canned analyses. Users can respond to nine Yes or No questions on how findable, accessible, interoperable, and reusable these three types of digital objects are. Evaluations are kept in a database and shown next to each dataset, tool, or canned analysis as an indicator. All listed bioinformatics tools in Datasets2Tools come from publications, hence next to each tool's card are listed its Altmetric Attention Score, PlumX rating, and PubMed citations. This metrics display can help users to easily recognize and rate bioinformatics tools.

An online interface allows users to view the data included in the Datasets2Tools database. A search engine enables unrestricted queries, as well as keyword and metadata-based filtering. Several indicators, including the date of upload, the publication's Altmetric Attention Score, and the digital object's compliance with FAIR assessment, can be used to rate the digital objects returned by search results. The opportunity to quickly identify the digital items most pertinent to the user's interests is made possible by the combination of search query, metadata filtering, and sorting. Users may acquire details about the list of retrieved digital items by going to the landing pages that correspond to them after doing a search. There are landing pages for every dataset, tool, and canned analysis. The digital object's name, description, identification number, external links, and metadata, as well as connections to pertinent publications and related metrics, are all summarized on these landing pages. References to the most relevant digital objects are also provided by landing sites. Natural Language Processing (NLP) is used to determine such suggestions for similar digital objects. Links to more relevant digital objects that are connected to the specific object through canned analyses are also displayed on landing pages for digital objects. Access to datasets that have been examined by a tool are included on its landing pages, along with links to all its canned studies. Similar to this, dataset landing pages provide links to both the canned analyses that correlate with each tool that was used to examine the dataset. Finally, landing pages for canned analyses provide access to the datasets and software used to create them.

As mentioned above, landing pages make it simple for users to access evaluation forms so they may rate digital objects based on how well they adhere to the FAIR principles. After the user

submits their rating, the database stores the results, which are then combined with all other users' ratings and shown as a grid icon on the respective landing sites [15]. Users may prioritize and choose services depending on their total FAIRness score using the search ranking algorithm, which also incorporates the FAIRness evaluation data.

Datasets2Tools is a prototype for a platform that allows for the effective management of three different categories of biomedical digital objects, however it focuses primarily on indexing a new category of digital item known as the canned analysis. Tens of thousands of previously conducted bioinformatics studies may be found, accessed, integrated, and reproduced thanks to the Datasets2Tools architecture, which is built in accordance with the FAIR principles [15]. Developers looking to market their computational tools may find Datasets2Tools a useful platform since it enables users to add their own canned analyses to the database. Bioinformatics developers may be able to connect with additional users who are not familiar with their tools by sharing analyses produced by the computational techniques they create. Additionally, researchers may find Datasets2Tools to be a useful platform for promoting their data analysis methods and experimental findings.

Finally, Datasets2Tools presents a technique for assessing digital objects in accordance with the FAIR principles. By responding to a brief questionnaire, Datasets2Tools users may assess the FAIRness of research objects. The database stores evaluation findings, which are then made accessible on the landing pages of each research object via a FAIRness indicator. Users can rank resources by their adherence to the FAIRness principles using Datasets2Tools' ranking of search results according to FAIRness ratings. Additionally, FAIRness evaluations may be automated by quickly checking if a digital item in the Datasets2Tools database complies with the FAIR standards. The work done to formulate and execute Datasets2Tools is assisting in achieving the goal of the NIH Data Commons [204].

3.4 The BioStar Forum

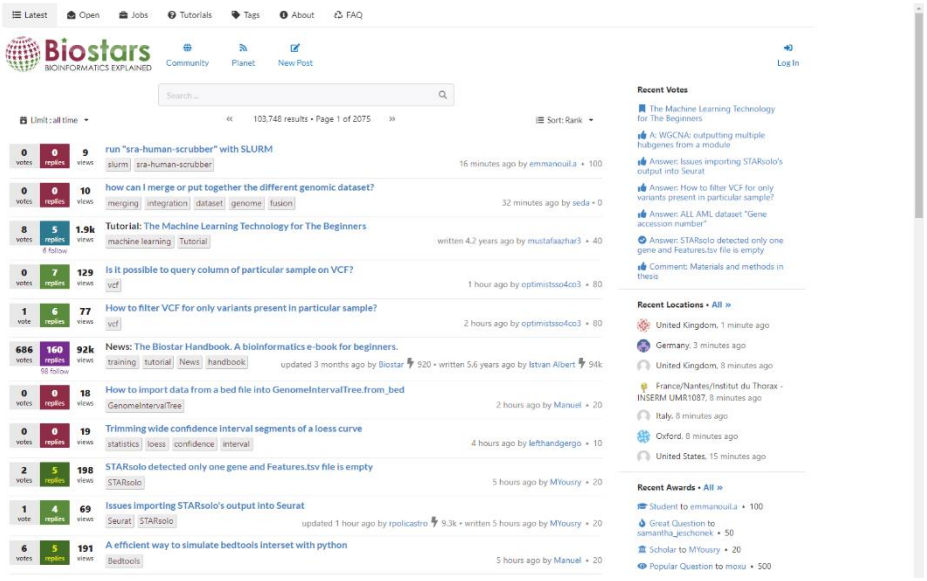

Figure 3.4: The BioStar forum

BioStar [205] is an online forum, derived from the Stack Exchange platform, that hosts discussions between specialists and others looking for answers to computational biology issues. The main advantages of BioStar include its significant and engaged community of knowledgeable users, quick response times, the way in which questions and answers are arranged so that discussion stays on the subject at hand, as well as the way in which questions and answers are ranked so that their usefulness can be determined. These rankings, which are based on community feedback, also apply a reputation score for each user, that keeps knowledgeable contributors interested.

By enabling researchers to ask questions and provide answers to bioinformatics-related issues, BioStar fosters a strong bioinformatics community. Based on the Stack Exchange technology, which enables users to ask or answer questions for a specific problem and aims to form a concise discussion limited to a single question, BioStar was developed by associates of the Bioinformatics Consulting Center in late 2009. Members of the site evaluate the questions and answers, and anybody can update them in a way that is similar to a wiki. The programming and informatics groups on the Stack Overflow website have utilized this technology with notable success [114]. Since its launch, BioStar has numerous active registered users and has gathered an immense knowledge base questions in the bioinformatics and programming fields.

The BioStar online interface offers a setting where beginner users can ask questions while interacting with experienced users who are effective to assist with the answers. The process for submitting a new technical question to the forum is quick and easy, and individuals with greater experience frequently advise rewording queries that are poorly written. To motivate users to provide accurate, relevant, and helpful responses, BioStar uses a system of user ratings, badges, and privileges. As a result, the authors of the original inquiry as well as those who submit a response are rewarded for their work. In addition, compared to a typical forum, keyword tags and a search option make it incredibly simple to identify similar topics or explore a certain topic.

Another significant function of BioStar is the awarding of reputation points based on user votes for good questions, correct or acceptable responses, and relevant comments. Users increase their involvement to improve their reputation, and the competition raises the overall site's response rate and quality [114]. The responses that have received the highest support from voters for each question rise to the top. BioStar has profited from the fact that bioinformaticians seem to be more familiar to this discussion style as early visitors have also advertised the website on various internet networks.

In conclusion, due to these features functioning altogether, BioStar has developed into a vibrant, dynamic, and quickly expanding online forum for bioinformaticians and computational biologists. It has quickly gathered a substantial and diverse group of users, serves as a point of reference for other smaller groups that have developed around other social sites, and has had measurable influence on how some research projects were conducted.

3.5 Findings

Many databases and tools have been developed by a wide range of providers, including small businesses and big service organizations, to serve the dynamic fields of biology, biotechnology, and medicine. Researchers must deal with biological data that is inherently complicated and has been incorporated into myriad of data formats for examination using a wide variety of techniques and software, installations, and interfaces. It is difficult to judge the extent and compatibility of emerging resources in the context of global offers since developments are frequently made on the spot and there is no clear source of consolidated information. For instance, software may lack a codified definition of its scientific and technological purpose, and the lack of

permanent, specific tool IDs makes it difficult to cite sources accurately and ensure that analyses can be repeated. The task of bioinformatician, that create useful workflows for scientific discovery, is difficult since there are considerable obstacles to finding and connecting the appropriate tools among a wide range of options.

To solve the problems of research objects and users' evaluation, the aforementioned platforms used suitable metrics and rewarding systems that could help with the engagement of their users. To solve the profile building issue, they provided the ability to identify and semantically connect users with their software and publications. Constantly changing computational and technical requirements has caused the biomedical research to become increasingly data intensive. Many biomedical researchers need to be able to acquire and utilize the appropriate datasets and methodologies despite the increasing biological data. Therefore, the adoption of evaluation methods can overcome the substantial obstacles for replication, distribution, and widespread reuse.

4 Implementation

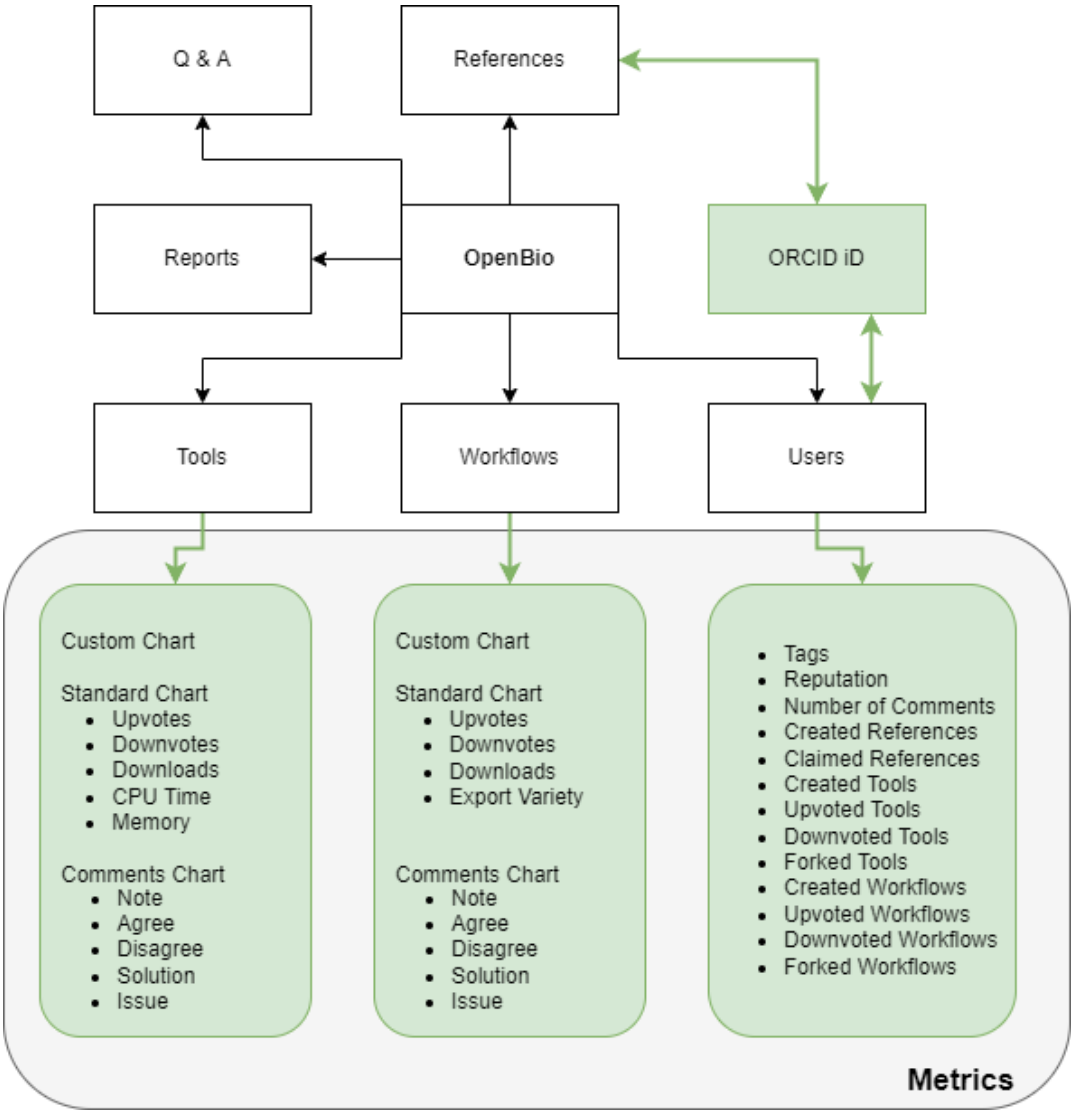

Figure 3.1: The OpenBio system with the added components

This thesis aimed to find solutions to some vital problems in the domain of bioinformatics. It tries to find a solution for the misattribution issue of bioinformatics tools, where tools are used without being properly cited in academic papers. This acts as a counter incentive for authors when it comes to submitting tools in open science repositories. Thus, it allows users to be properly cited for the work that they submit in open science environments that deviate from the typical "academic journal" setting such as tool repositories, and open Workflow Management Systems. Next, it contributed to the tool, data and workflow repository, the OpenBio platform, by adding a rich set

of both automatically acquired and user-provided quality metrics. Those metrics are displayed in an intuitive user interface, allowing users to quickly explore them in order to make the most suitable and informed decision regarding the optimal component that suits best their analysis. Finally, it provided clear and objective indications of the expertise and the scientific activity of users, acting as an incentive for them to be more open and active by ranking them according to their quality and quantity of contributions. All those components were added to the open scientific environment OpenBio. The overview of the OpenBio platform and the added components (green colored elements) are presented in Figure 3.1.

4.1 The OpenBio Platform

The OpenBio platform [17] is an open scientific online environment where researchers may build, modify, share, re-combine, export, run, evaluate, and contribute on maintaining tools, data, and workflows [16]. Its goal is to provide data stewardship capabilities that will facilitate the reuse of content outside of its original context. Any code that can be executed on a local computer can be used as research object code. This includes simple scripts, source code, binaries, Docker files, or any other containerization format. Without the need to learn any Domain Specific Language (DSL), all data can be loaded into research objects using a user-friendly interface. Importing a tool into OpenBio.eu involves the same processes as importing it into an operating system like Linux; a user simply must include the installation instructions that they would have to put in an execution environment [16].

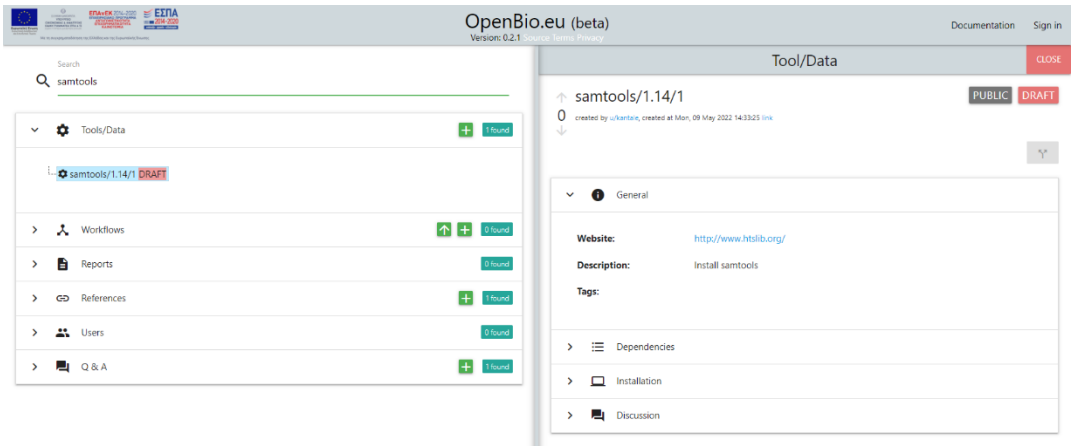

Figure 4.2: The OpenBio platform

4.1.1 Fair Principles of Research Objects in OpenBio

Even though the FAIR rules are focused on data, they can easily be applied to tools and workflows as well. By establishing suitable formats, data, for instance, may define biomedical entities (i.e., gene expression). Similar to this, describing tools and workflows may be represented by data structures. All research objects that participate in scientific analysis should thus follow the FAIR standards. This last attribute necessitates the fulfillment of another key criteria, namely, that the digital source must be embeddable. As a result, the FAIR principles are conceptualized as FAIR-E (where "E" stands for embeddable), which adds an embeddable dimension.

The included research objects in OpenBio are thought of as integrated constructions that include both the code needed for installation (or download in the context of a dataset research object) and the semantic specification needed to connect them to other research objects. The method, when combined with semantic web technologies, enables the usage of linked-data structures as linkages across FAIR research objects while also masking users from the specifics of the representation, enabling a more abstract level of engagement with the system [206]. Since issues are often articulated in terms of the job to be undertaken, and software is typically characterized in terms of functionality, FAIR-E bridges the fundamental contradiction between the problem to be addressed and its computational solution in this context [207]. To describe it more plainly, the design philosophy that guides OpenBio views the software that manipulates the data, as well as the data itself, as neighbor ideas, if not same concepts at least in the context of their semantic meaning.

4.1.2 The OpenBio Execution System

The execution of research objects, on a system a user has access to, is a vital component of any repository. Although there are several repositories that preserve and provide access to tools, data, and workflows in the biomedical field, such as OSF [208] and bio.tools [14], only OpenBio offers the ability to execute these integrated items. A new tool, set of data, or workflow may be imported into OpenBio.eu by giving clear instructions on how to install, validate, and execute a research object.

OpenBio uses the programming language BASH [209] to install and run objects. A programming language is required because precise instructions on how to install a tool, retrieve

data, or carry out a workflow is mandatory to install and run objects. BASH is the existing language of the majority of Unix-like operating systems, therefore, by hosting code in BASH, it is immediately executable in as many environments as needed. However, this does not imply that other languages are not supported when code is hosted in BASH. It was chosen because it makes it possible to combine several programming languages, programs, and scripts into a single script. OpenBio serves as a repository for BASH scripts supplied by users. These scripts are independent, so other people who want to use them don't need to download or install any other programs. Additionally, OpenBio provides a method for connecting these user-provided programs. The same situation can be applied when various tools and data are brought together through a user interface to create workflows. OpenBio serves as a repository for BASH scripts in this way, ensuring that scripts are compatible when integrated together in a workflow.

All objects on OpenBio.eu are of the same type, including tools, data, and workflows. Typically, there is a distinction between data, tools, and workflows in workflow management systems and open science environments. For each, users usually must specify unique attributes, or store them in structures and tables. However, this does not apply in the OpenBio platform, because in the context of a workflow management system there are no distinctions between tools, data, and workflows from a semantic perspective. Data can be considered meaningless without the existence of other data, whereas tools and workflows have dependencies. Therefore, OpenBio considers that tools, workflows, and data are all the same type of research objects.

4.1.3 Research Objects in OpenBio

The OpenBio platform is structured according to four primary components (Figure 4.3): the research object repository, which features workflow execution reports, references, and (possibly) inquiry answering discussions in addition to free-text search operations to find the tool, data, or workflow object most pertinent to the given task; a tool data, workflow definition or description component; the primary component of the platform, responsible for workflow creation, tool, data, or workflow object installation with its validation, and their dependencies description where a Graphical User Interface is provided to drag-and-drop the necessary research objects; and the module for defining environments and carrying out workflows themselves, together with a resource manager to keep track of the execution process.

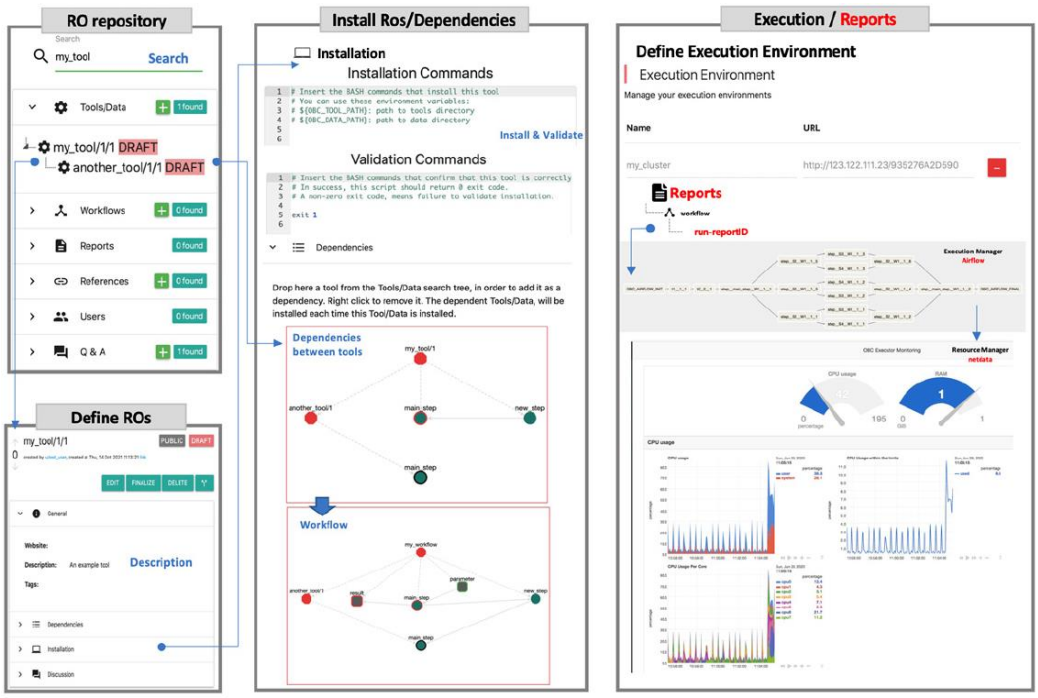

Figure 4.3: The four primary components of OpenBio

Users of OpenBio can establish a research object by just importing the installation scripts. The utility may be installed using these instructions on any machine with a BASH shell terminal. Many variations of the same research object are possible. Research objects with the same name and version can be imported by many users. A research object can also be complemented by rich-format text in markdown in addition to the BASH commands. Users of OpenBio may easily create dependencies between research objects by dragging and dropping objects. There are two stages for research objects: “Draft” and “Finalized”. Future revisions to objects in the “Draft” stage are possible. Users have the option to “Finalize” an imported item after they are certain that it doesn't require any more adjustments. An object that has been finalized is no longer editable and it is an unchangeable object that may be used in a pipeline that is repeatable and secure for the future. Users may also “Fork” either a draft or finalized version of an item in any scenario. Software development has influenced the idea of forking, where users can make an exact copy of a source code and make whatever changes they choose.

Additionally, commenting and evaluation is an important component that is lacking from workflow management systems and biomedical text annotation tools. It can be difficult to determine the current opinion regarding a research object or whether there is a strong view about it, especially for students and beginners to the subject. Additionally, people are unable to convey and publish their own thoughts using the tools that are now available. This has an impact on the existing competition in the field, restricts the development of current research objects, and hinders the collaborative emergence of innovative ideas that can advance research and emphasize innovative paths. By enabling ratings and comments on each research object, OpenBio can satisfy this demand. Moreover, it has a Q&A collaborative feature where any user may post a query, ask for help, or just leave a general comment. A collaborative component that facilitates the establishment and development of a semantically rich discussion graph regarding any research object found on the OpenBio platform underpins these functions [210]. The nodes on this discourse graph can be of various kinds, such as issues where a question may be posed by a user for discussion, solutions where a suggestion may be made by a user to address an issue in question, positions in favor where an argument may be in support of a suggested solution, positions against where an argument may aim to disprove or contest a solution, and notes where a comment may be offered to include additional information without changing how the discussion is evaluated. The collaboration feature also includes cutting-edge search capabilities, a variety of choices to enhance graph display, and the capability to score graph nodes by assigning likes or dislikes. In general, the OpenBio platform's collaboration feature enables users to establish a highly participatory process in which they can quickly choose which research objects should be taken into account, pinpoint and discuss their advantages and disadvantages, and manage the intricacy of biomedical workflows.

4.1.4 Workflows in OpenBio

A core principle in computation biology and bioinformatics is the concept of a workflow, where a workflow, strictly speaking, is a collection of interconnected computing stages. Common workflow management systems used in bioinformatics include Nextflow [211], and Galaxy [3]. Although the building of workflows using a flow-centric approach has a lengthy tradition in bioinformatics, it was inspired by strict industrial design systems and is thus not entirely

appropriate for the adaptability and flexibility of contemporary bioinformatics research. Additionally, the flow-centric design of workflows necessitates superior IT capabilities. Users of OpenBio may create workflows by simply entering the instructions that carry out each step. Once more, these are the exact commands that would be entered into a BASH terminal. The primary distinction is that they may explicitly call steps from other steps rather than implicitly establishing step order via a dependency resolution method. Users with little programming knowledge are familiar with the "function calling function" paradigm, which is followed by this abstraction.

Moreover, OpenBio offers a Docker container that connects to its web server and serves as a customized execution environment. Users are allowed to have an unlimited number of execution environments; each time a Workflow is executed, a unique object (report) is created that stores the results and logs. Users can share these with others.

4.1.5 The OpenBio Development Frameworks

Upon its creation, the OpenBio platform uses three main development frameworks: the AngularJS framework for the front-end development of the platform, the Django web framework for the back-end development of the platform, and the Materialize framework for the construction of the user interface components that exist in the platform.

Angular [212] is an open-source JavaScript framework that enables the development of RICH Internet Apps and offers scalable infrastructure and productivity to the most important applications. A framework for client-side Model View Controller architectures is offered by Angular, which makes it easier to design and test any single-page apps. It is an essential front-end development framework that makes the development process easier by establishing a comprehensible and open environment.

Django [213] is a high-level Python web framework that enables the quick creation of safe and dependable websites. Django, which was created by established programmers, handles a lot of the pain associated with web development, allowing to concentrate on development of the app without external problems. It is open source and free, has a strong community, excellent documentation, and a variety of free and paid support options. Among its many benefits is its fast development completion, its inclusion of many implemented development tasks, its security

algorithms, and its scalability and versatility. The database that is connected to Django on the OpenBio platform is the PostgreSQL [214], which is the default configuration database of Django.

Materialize [215] is a modern responsive front-end framework based on Material Design [216] and is developed using HTML, CSS, and JavaScript. By adhering to contemporary web design concepts like browser adaptability, device freedom, and progressive degradation, the Materialize components assist in creating visually appealing, consistently designed, and functioning online pages and web apps. It facilitates the development of attractive, responsive, and fast websites.

4.2 Implementation of the Connection with ORCID

According to the systematic literature review findings, regarding the issues of misattribution and lack of acknowledgment of any form of attribution, ORCID seemed to be an appropriate solution to those problems. Therefore, to find an answer to the first research question of how an effective solution to the problem of misattribution can be found, a part of implementing this thesis was to connect the OpenBio platform with ORCID. This could offer the user the ability to connect their account with ORCID and therefore allow the seamless connection of any activity in OpenBio with any other activity in a service that also uses ORCID as an authentication mechanism.

4.2.1 Overview of ORCID

One of the most important research fields in scientometrics is the examination of the scholarly pursuits of specific academics [217]–[219]. For a greater analysis of data sharing activities, a targeted examination of each researcher's actions is very relevant [220]. In order to properly analyze the actions of different researchers, researcher name disambiguation has been a lengthy issue within the discipline of scientometrics [46], [221]. In more recent years, algorithmic solutions have been found, though each one has some drawbacks [222]. Released in 2012, ORCID can give academics personalized alphanumeric identities with the aim of giving writers in the academic and scientific communities authoritative control. Authors can set themselves apart from others with similar or identical names by using an ORCID identification. It represents a more

fundamental answer to the name ambiguity problem. The ORCID identification was created to provide the scientific community with a special registry to maintain their information and records, either manually or by integrating automatically with other sources of data, given the difficulties associated with identifying individual researchers' outputs [84], [223]. Multiple research workflows are integrated into ORCID for the goal of making it simple to link data about various research outcomes. Even though scientific and conference papers are two of the most popular academic activity outputs, ORCID-registered researchers have access to a wider range of output options.

Opposed to other systems, the goal of ORCID is to offer a register of distinct persistent IDs for researchers. It collaborates with the community to enable “collection and connection points” for identifiers in systems that researchers frequently use, including systems for submitting manuscripts or applying for grants. It also enables automatic updates to a researcher's record when their innovative and academic works are posted [224]. Researchers determine what may be associated to their ORCID identity, as well as the privacy settings for their account. Registration is separate from membership, allowing for career-long usage of the identification regardless of changes in field, location, name, or affiliation. The only information found in ORCID records is relationships between identifiers and references to the source objects.

It primarily advertises to institutions and scholars. Researchers can use ORCID's distinctive IDs, and organizations can help their members maintain their profiles. To help them distinguish writers from one another more easily, several publishers have started to ask or demand authors to provide an ORCID identification when submitting in their journals recently. Additionally, ORCID identifiers are indexed in important citation databases including Scopus [50], Web of Science [225], and PubMed [21].

Researchers may register for free ORCID identities very quickly and easily by simply filling out a brief registration form. An ORCID id is connected to the researcher, not their employer or industry (Figure 4.4). Researchers can therefore use ORCID identifiers throughout their whole careers. It can also be useful to have all of the works posted in one place so that people can view the research of different researchers in different disciplines and in different formats. Researchers may readily share their whole lists of works by connecting to their ORCID profiles, for example on a CV or personal website. Due to the identifier's individuality, it also enables the tracking of

works prior to and following a name change, which is quite advantageous for individuals who modify their names throughout their employment [226].

Figure 4.4: The ORCID iD usage

The vast array of works that academics may add to their profiles with ORCID IDs, including as papers, newsletters, conference posters, presentations, videos, datasets, and many more, is one of the main benefits of this system. Researchers may easily enter each of these using the straightforward template that is given in the profile. Researchers can also keep track of and exhibit grants and other financial awards in addition to their published work. Moreover, the issue of incorrect author attribution caused by shared or ambiguous author names is eliminated by assigning distinct ORCID iDs to scientific journals, datasets, and other research products [227]. Additionally, by connecting ORCID profiles to other systems as well as enabling automatic information transfer between systems, reporting workload and time is reduced. Furthermore, as ORCID iDs are allocated to specific researchers, researchers can keep the same iD over the course of their careers even if their academic affiliation changes. The ORCID registry clearly benefits both researchers and libraries [228]. When a researcher uses ORCID iDs to collect citations,

calculate their h-index, or write a small bio for government funding organizations, it can be easier to identify their works (Figure 4.5).

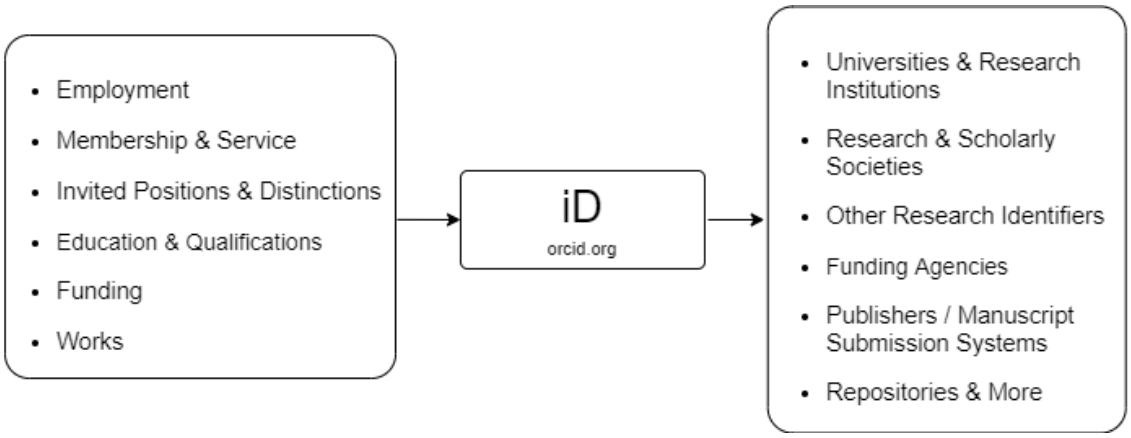

Figure 4.5: ORCID importing and exporting information

Being a completely transparent organization is important to ORCID. They offer a publicly accessible public API (applications programming interface) that anybody may utilize [229]. Anyone may create apps that obtain an authenticated ORCID identification for a user; obtain the public ORCID record of a user in machine-readable form; do a computer-generated search of the ORCID registry's open data; and enable users to login using their ORCID login credentials and passwords on non-ORCID applications. All its presentations and materials are available online under the Creative Commons Zero license [230], which has no restrictions on use. Its software code is also published on a public repository with an open-source MIT license [231].

It has long been difficult to distinguish between writers with similar names or to follow down researchers who have altered their identities in the scientific and publishing worlds. Scientists, journals, and organizations may trace a person's whole scholarly career with the use of ORCID identifiers. Having an ORCID identification is becoming increasingly desirable for anyone looking to publish because of publishers demanding them and reference databases indexing them. Getting an ORCID identification is worthwhile an individual researcher's cost and effort to sign up for and manage a free account, whether an institution pays in a membership. A thorough list of academic activity reported by researchers may be found at ORCID; it may not always be widely adopted by academics [232], [233], but it is the sole source that explicitly links specific researchers to their research results.

4.2.2 The OAuth 2.0 Standard

OAuth 2.0 [234] is a standard that enables an application or website to access resources maintained by other web applications on a user's behalf. OAuth 2.0 stands for "Open Authorization." It became the industry's accepted standard for online authorization in 2012, taking the place of OAuth 1.0. Without ever disclosing the user's credentials, OAuth 2.0 offers consented access and limits the activities that the client app can carry out on resources on the user's behalf. OAuth 2's primary platform is the web, but the standard also explains how to manage delegated access for additional client types, such as web apps, server-side web apps, native/mobile apps, linked devices, etc.

OAuth 2.0 is a protocol for authorization, not for authentication. As a result, its main purpose is to enable access to a range of resources, such as external APIs or user data. Access Tokens are used with OAuth 2.0. An Access Token is an information piece that symbolizes the end-authority user's to access resources. There is no set structure for Access Tokens in OAuth 2.0. However, the JSON Web Token (JWT) format [235] is frequently utilized in some situations. This allows token issuers to embed data directly into the token. Access Tokens may also have a deadline due to security concerns.

The OAuth2.0 authorization framework's main definition includes the concept of roles. The fundamental parts of an OAuth 2.0 system include:

- The Resource Owner, where a person or machine is the legal owner of and has access to the restricted resources.
- The Client who is known as the system that needs access to the resources that are secured, and the Client must possess the necessary Access Token in order to access resources
- The Authorization Server which accepts requests from Clients for Access Tokens and provides them upon successful resource owner authentication and consent. The Authorization endpoint, which manages the user's interactive authentication and consent, and the Token endpoint, which is used in a machine-to-machine connection, are the two endpoints that the Authorization Server exposes.
- The Resource Server that handles access requests from the Client and safeguards the user's resources. The necessary resources are then returned to the Client after accepting and validating an Access Token from them.

The idea of scopes is crucial to OAuth 2.0. They are used to precisely define the grounds for granting access to resources. The Resource Server determines whether resources are relevant to acceptable scope values.

After the Resource Owner has given permission for access, the OAuth 2 Authorization server could not immediately return an Access Token. An Authorization Code may instead be returned and subsequently traded for an Access Token for more security. Along with the Access Token, the Authorization server could also send a Refresh Token. Refresh Tokens, in contrast to Access Tokens, often have lengthy expiration dates and may be traded for fresh Access Tokens as the former do. Due to these characteristics, clients must securely store Refresh Tokens.

The Client must first get its own credentials from the Authorization Server, a client id and client secret, in order to uniquely identify and verify itself when seeking an Access Token before OAuth 2.0 may be utilized. Access requests are made via OAuth 2.0 by the Client, such as a desktop application, smart TV app, website, mobile app, etc. The overall flow of the token request, transaction, and answer starts when the with the client id and secret as means of identification, the client submits an authorization request to the authorization server together with the scopes and an endpoint URI (redirect URI) to which the access token or the authorization code should be sent. Then, the Client's identity is verified by the Authorization server, which also confirms that the requested scopes are legal. Then, to allow access, the resource owner communicates with the authorization server. Depending on the grant type, the authorization server redirects back to the client using either an authorization code or an access token. The refund may also include a Refresh Token. Finally, the Client asks the Resource server for permission to the resource using the Access Token. Grants in OAuth 2.0 are the series of actions a Client must do to get access authorization. The authorization framework offers a variety of grant types to handle various circumstances that include the Authorization Code grant, the Implicit grant, the Authorization Code Grant with Proof Key for Code Exchange (PKCE), the Resource Owner Credentials Grant Type, the Client Credentials Grant Type, the Device Authorization Flow, and the Refresh Token Grant.

4.2.3 Solving Misattribution with ORCID

To overcome the problem of misattribution, a part of this thesis was to connect the platform of OpenBio with ORCID. This can offer the user the ability to connect their account with ORCID

and therefore allow the seamless connection of any activity in OpenBio with any other activity in a service that also uses ORCID as an authentication mechanism.

4.2.3.1 Implementation of the Association of ORCID with OpenBio

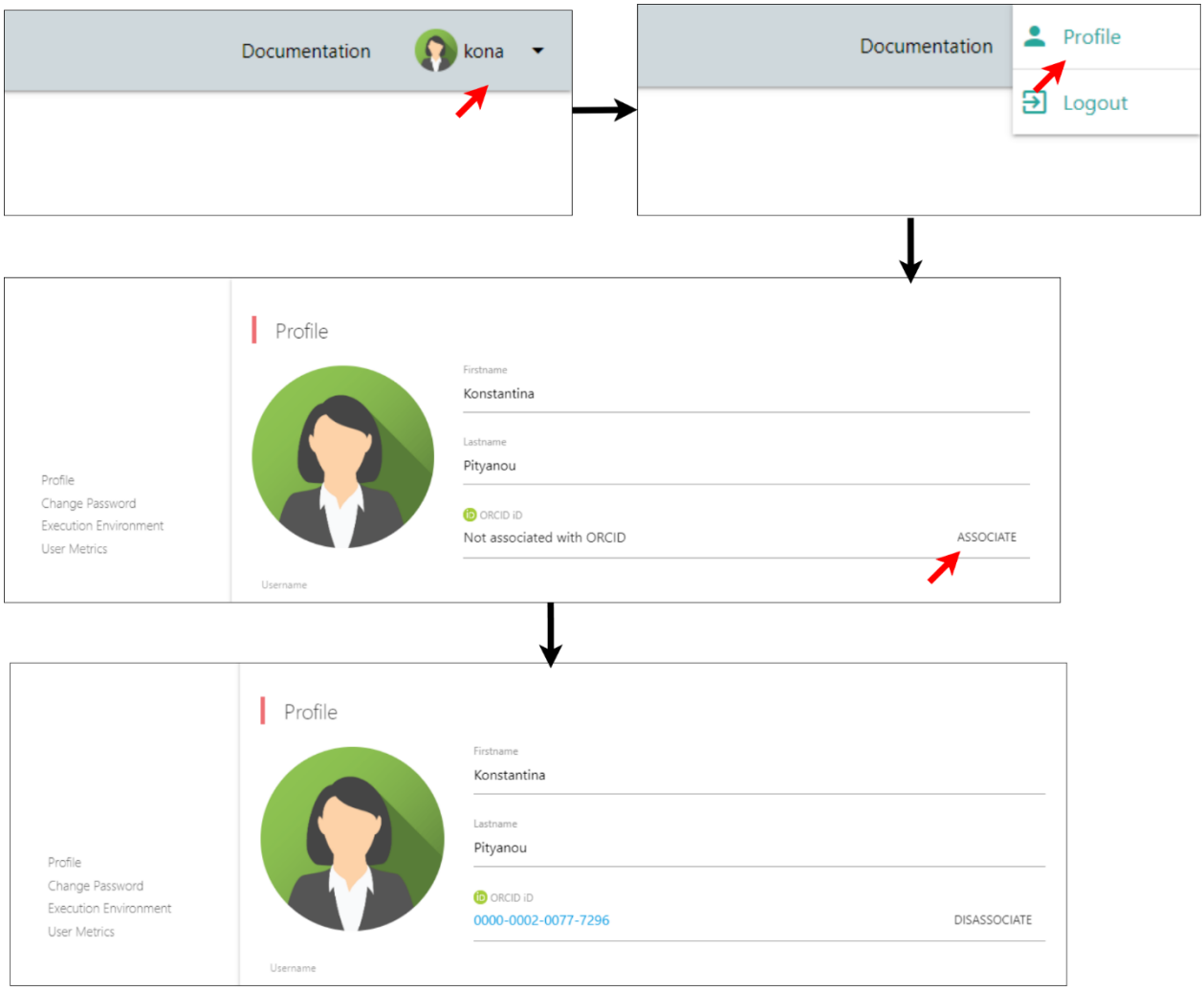

Figure 4.6: Association of user with ORCID

The user can associate his account in OpenBio with ORCID by navigating to his "Profile" and clicking the button "Associate" that appears next to the "ORCID iD" input field, under the "Profile" section. After clicking the "Associate" button, the user is sent to ORCID, where ORCID asks the user to login to their account. By accepting to grant permission to the OpenBio platform, the user is redirected back to the OpenBio platform, and the association is complete. The "ORCID

iD" input field, after the successful association, it changes its text from "Not associated with ORCID" to the user's ORCID iD (Figure 4.6).

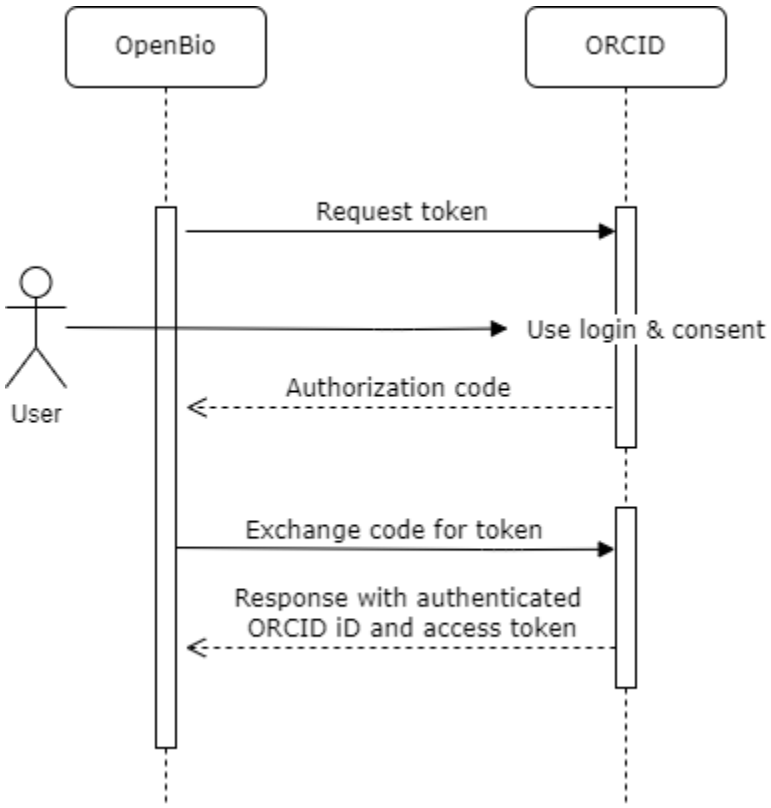

Figure 4.7: Sequence diagram of the association of OpenBio with ORCID

The implementation of the code regarding the connection with ORCID starts after the user clicks the "Associate" button. The "3 legged OAuth" method is used by ORCID integrations to authenticate users and ask for permission to connect with their information. Using the Public API, any integration can request read rights. The user is redirected to ORCID in order to login using their credentials. The redicted URL is a customized authorization URL that includes the user's details and the scopes that indicate the precise parts of their record the OpenBio want to access. After accepting the permissions, the user returns back to OpenBio with an authorization code. The OpenBio exchanges the authoriziation code for an access token. Their ORCID iD and access token that is valid for the specified scopes are obtained in JSON format [236]. Python's social auth [237] was used as the social authentication and authorization mechanism for the OpenBio project that supports OAuth 2 protocol. After the association is complete, the user's ORCID iD is stored in the database for future use. The sequence diagram of this process is presented in (Figure 4.7).

4.2.3.2 Implementation of the Credits Claim in OpenBio with ORCID

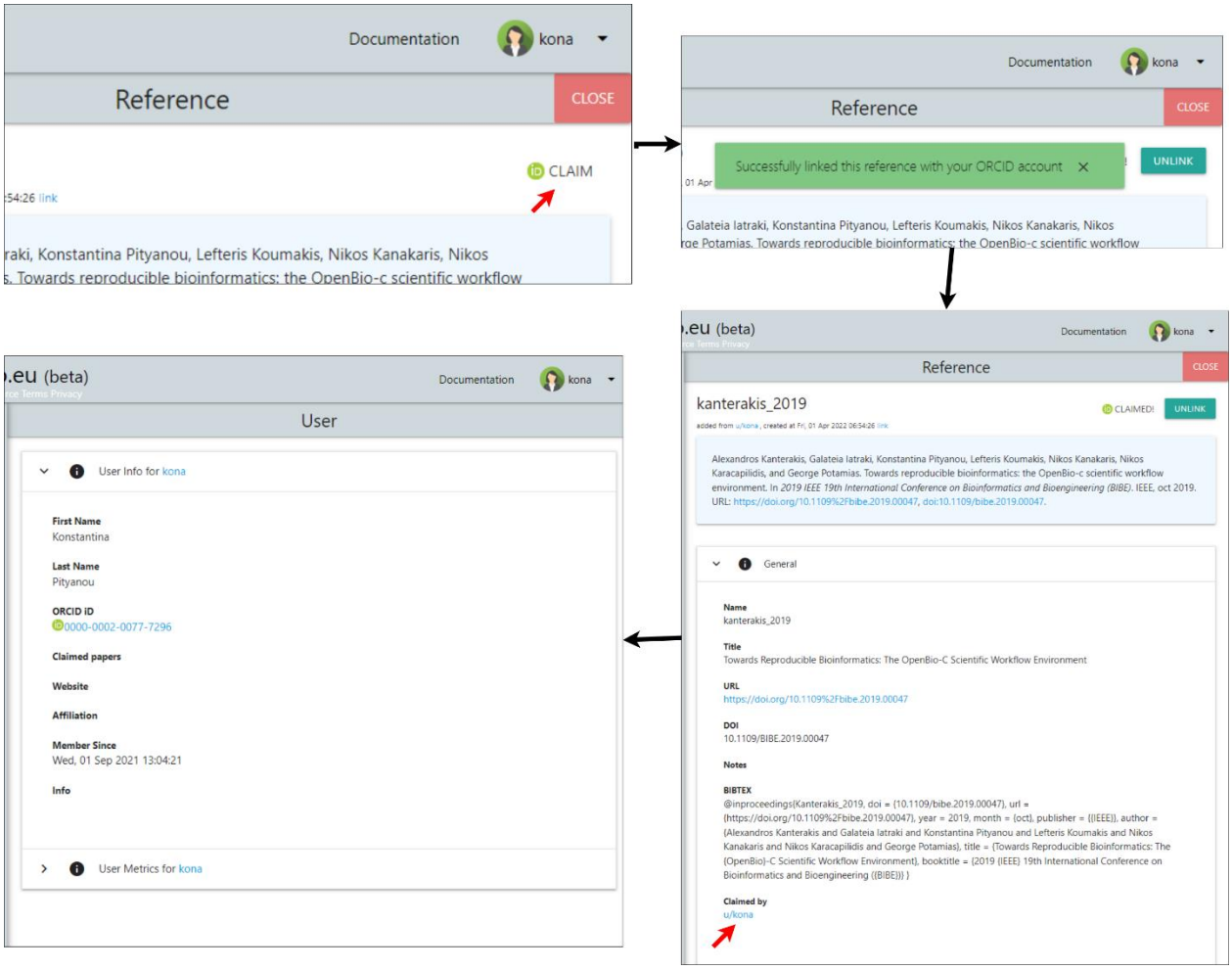

Figure 4.8: User claim reference with ORCID

Once the user has associated their account with ORCID, they can claim credits for the publications in which they are authors or coauthors. Regardless of whether they have uploaded the publication on OpenBio or not, the user can request their contribution rights. In order to claim a publication, the user can click the “Claim” button that is located on the top right of each opened reference. If the users ORCID account includes the requested DOI, then the publication is

successfully claimed. A success message notifies the user that the paper is claimed, and their username appears as an author with a redirection link to their OpenBio profile (Figure 4.8).

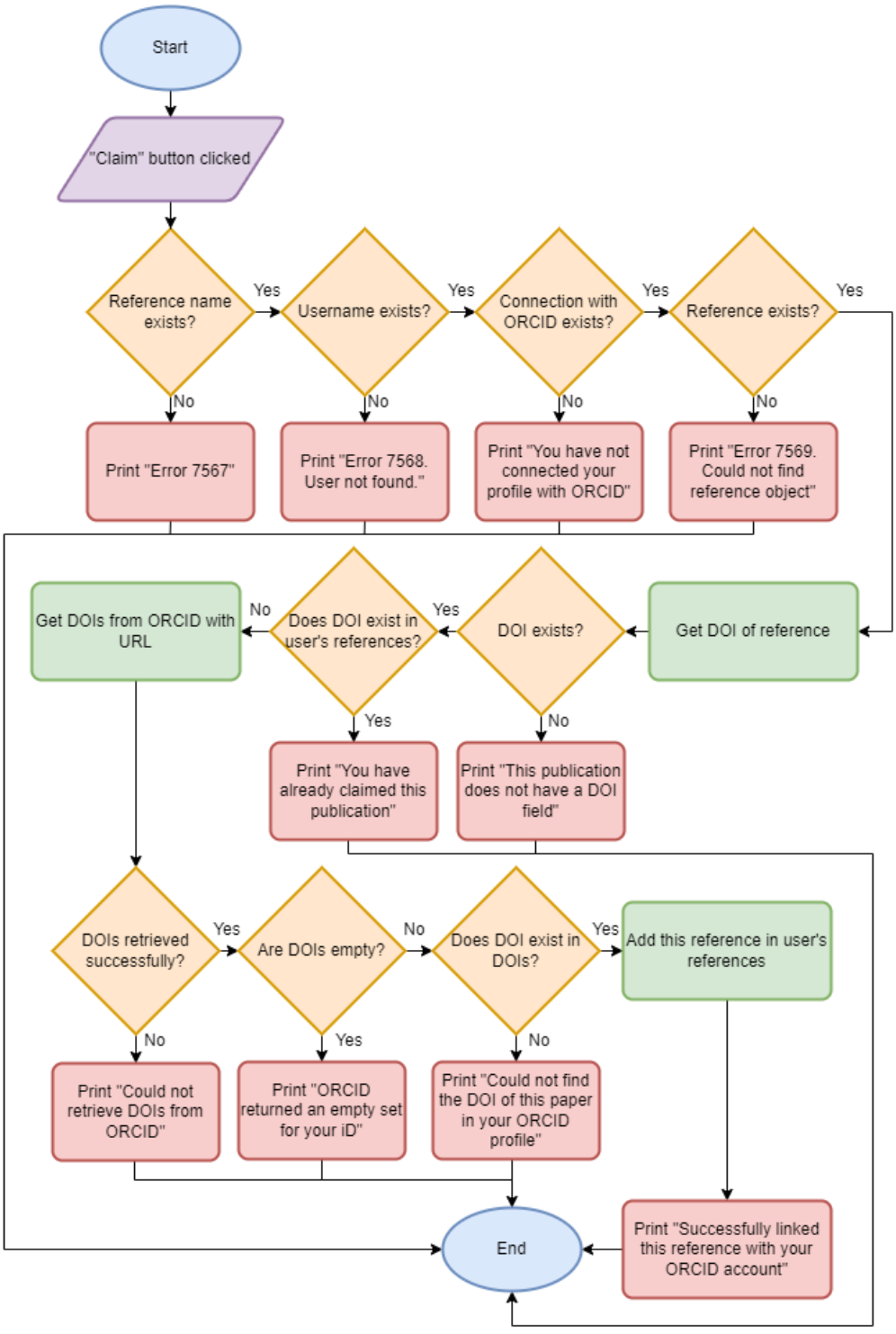

Figure 4.9: Flowchart of the credits claim process

The implementation of the code regarding the claim of a user's reference starts when the user clicks the "Claim" button. Using Django, the system first checks that the reference name, the user, and the connection with ORCID exists. It then retrieves the reference's DOI and compares it with the list of DOIs that exist in the user's ORCID profile. If the DOI is found, the user successfully claims that reference. The list of DOIs on the user's ORCID profile are accessed through a URL that includes the user's ORCID iD. The URL returns the requested information in JSON format [236], which is then modified to extract all the available DOIs for comparison. The flowchart implementation of this process is presented in (Figure 4.9).

4.3 Implementation of the Evaluation Metrics on Research Objects

In biomedical research, the integrated analysis of large and increasingly complex volumes of data is challenging. For the purpose of extracting and analyzing the biological insights from this enormous quantity of data, it is critical to have effective and practical sets of research objects, such as tools and workflows. Given the significance of bioinformatics in several biological and biomedical research, great effort must be placed into creating workflow systems and tools that are maintainable as well as reproducible computational analyses. A scientist should be given the freedom to choose the research object that best fits their needs. Evaluation metrics can help in keeping track of the scientific system's progress toward openness at all levels, and in measuring performance, to recognize and reward better methods of functioning in groups and individuals. All these measures can aid in the direction of qualitative analyses, enhancing the understanding of the importance of research findings and study objects.

To answer the second research question on the method a research object can be evaluated, this thesis integrated a set of metrics on each tool and workflow that exist on the OpenBio platform. The implementation of these metrics is to assist users on evaluating research objects and examine on whether they could fit their needs.

4.3.1 Data Visualization

Visual representations give data contexts that engage viewers' minds and disclose information that is typically hidden in data tables. Using visual elements to represent the actual

substance of data is typically far more natural. Data may be used to create tales through charts and maps in a compelling, understandable, and efficient way. They make it simpler to compare data, offer insights, and expose trends, correlations, causation, and other characteristics hidden in the numbers. They may also compress a lot of information into a short amount of space.

Figure 4.10: Data visualization charts

Statistical and informational charts, graphs, data maps, and other visual displays of quantitative information are examples of data visualization (Figure 4.10). However, it can also include any type of visual representation of data, including mathematical graphics, path networks (subway systems, roadways, and circuit design diagrams), musical and sound representations, timelines, geographic information systems, and any other visual artifact used to code data [238].

Data values can be utilized in visualizations based on three types that include the quantitative value that can be counted or measured, such as a number, a length, an area, or an angle; the ordinal value that may be ranked or contrasted, such as words, area, angle, length, or color saturation; and nominal that can be a category, such as a name.

It is possible to categorize visualizations, making it simpler to select the chart type. Most maps and charts fall into one of these groups: time-series that plots a single variable over a period of time, such as a line chart showing a trend; temporal/linear that are categories placed in a time-line, such as a sequence of activities; spatial/planar/volumetric that are categories distributed in a spatial map, such as an illustration that might be a cartogram or choropleth that displays data scattered across a map; comparison that exists within a specific time period, or categories related to amounts that are compared and rated, such as a bar chart that compares values; part-to-whole

that are categorical subdivisions as ratio to a whole, such as a pie chart displaying percentages for the slices; and correlation that compares two or more variables, such as a bubble chart comparing three variables or a scatterplot comparing two variables [239].

Edward Tufte [240] outlined a few criteria that may be used to assess the validity and reliability of visualizations. These include the “data-ink ratio”, which is the percentage of ink (or pixels) used to display the data; the “chart junk” which is the visual junk that is often distracting and unrelated to the data being displayed; and the “lie factor”, which is a measurement of the integrity of a visualization and is being used to identify charts that are inaccurately depicting lengths and proportions.

By eliminating unneeded lines and labels from charts, one may increase the data-ink ratio. Interactive Web visualizations should be quite minimalistic even if the lines can occasionally be crucial for context. Information about demand can be expressed with tooltips and other engaging resources. Human perception of graphics has a significant impact on communication, which may be influenced or enhanced through optical illusions. There are no charts with a zero “lie factor”, but a good selection can greatly reduce it. A poor decision raises the likelihood of a falsehood and can lead to viewers' having incorrect impressions. Quantitative information is best represented by position and length. After that, area, volume, curvature, angles, and lastly shadows, brightness, and color. Data in a bar chart is seen more accurately compared to the same data in a pie chart because lengths and locations are simpler to understand and compared than angles and areas [241].

Charts draw the viewer's attention while highlighting and revealing facts. They can help viewers evaluate and reason about data in many situations by deconstructing complicated data sets to encourage discovery and comprehension. They can, however, also exaggerate, deceive, and even lie. It's crucial to create aesthetically appealing charts, but designers should also learn how to strike a balance between function and form. Both art and science are involved in data visualization. Not everything needs to be explained in a chart. It's not always necessary to be exact. It may be intended for a particular audience, which would give the essential background information to comprehend and decode it.

4.3.2 The Chart.js Library

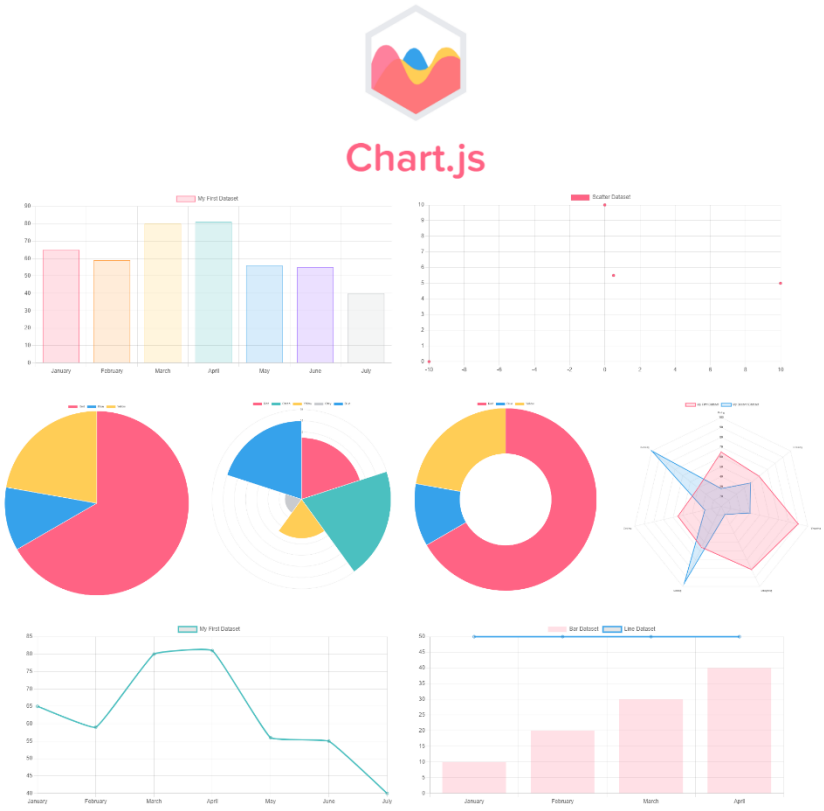

Figure 4.11: Chart.js sample charts

For the visual representation of the charts that were implemented in this thesis, the Chart.js library was used. Chart.js [242] is open source and free JavaScript charting library for designers and developers. It is maintained by a strong developer community on GitHub [243]. Following D3.js [244], it is the second most popular data visualization package on GitHub in terms of stars. Chart.js is a simple and small sized library that is available under the MIT license. It is based on JavaScript, but uses additional Web standards like HTML, CSS, DOM, and Canvas. Charts are automatically drawn in Canvas and provide complete control over canvas dimensions, pixel ratios, and settings. Furthermore, charts on chart.js are responsive and chart are redrawn on window resize supporting scale granularity. Chart,js supports eight basic chart types that include bar (horizontal and vertical), line or area (including stacked), radar, polar area, scatter, bubble, pie, and doughnut chart.

4.3.3 Implementation of the Metrics Charts

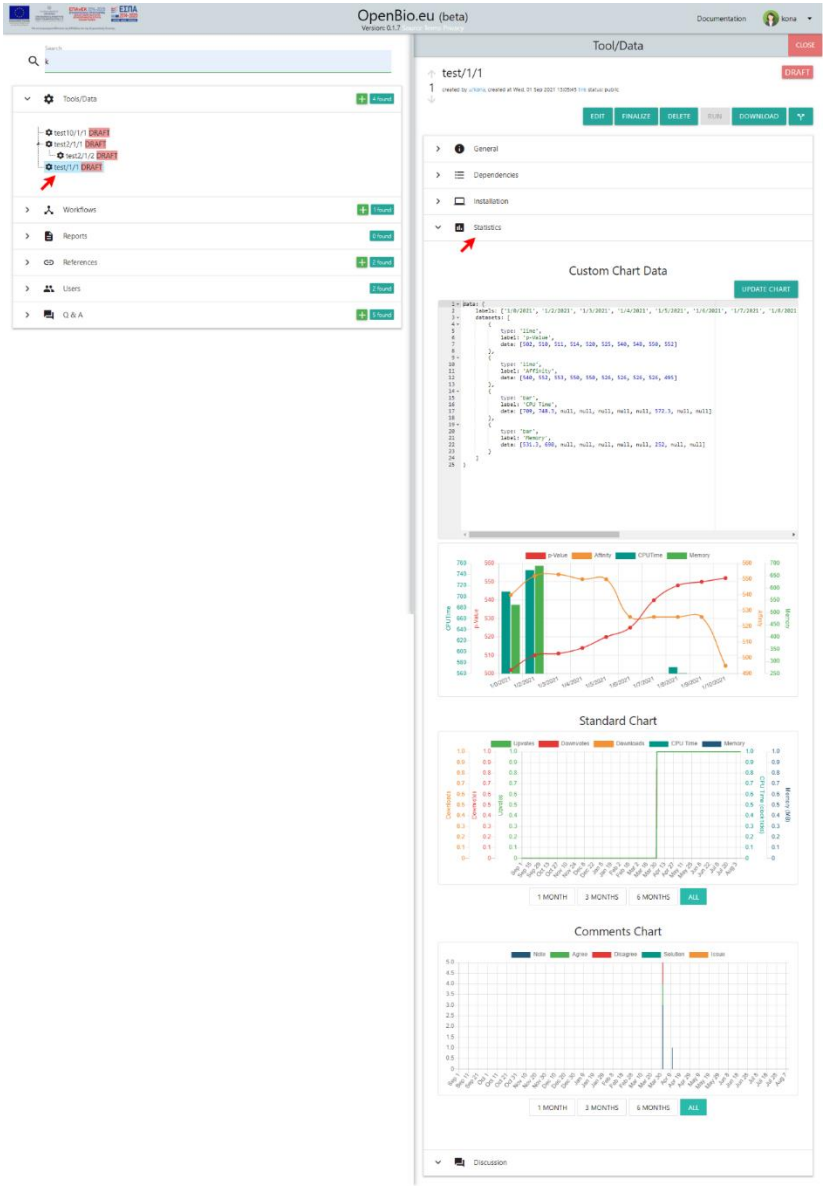

Figure 4.12: The Statistics tab in OpenBio

Each tool and workflow that exist on the OpenBio platform has a set of metrics that can assist a user in their evaluation. For the user to view those metrics they must click on the tool or workflow that they want to investigate and then open the tab that exists under the name “Statistics”. Inside that tab they can find all the available metrics for their evaluation (Figure 4.12). Those metrics are categorized in three different charts: the Custom chart, the Standard chart, and the

Comments chart. Each chart represents a different category of metrics, and they can all collectively be used to evaluate the effectiveness and usefulness of the selected research object.

4.3.3.1 The Custom Chart

Figure 4.13: The Custom chart in Statistics

The custom chart offers the user the ability to create their own chart with custom data metrics. By default, the editor contains a multi-axis chart with sample values. The user can modify, add, or remove an axis and its values. The supported types of charts in the custom chart include the linear and bar chart. After the user has modified the chart data in the editor, they can click the

button “Update Chart” to see their results in the chart below (Figure 4.13). The implementation of the custom chart functions in the same way for both tools and workflows. The user can hide or show any of the values that appear on the custom chart by clicking on the respective label on top of the chart. Tooltips are also included; hovering over the chart, a tooltip appears displaying all the values (y-axis) and each value’s number at that specific date (x-axis).

The text editor that was used in this implementation is the Ace code editor [245]. Ace is a JavaScript-based embeddable code editor. It matches the capabilities and efficiency of native editors like TextMate, Sublime, and Vim. Any web page and JavaScript application can quickly embed it. Ace is the replacement for the Mozilla Skywriter (Bespin) project and is being followed up as the main editor for the Cloud9 IDE.

The implementation of the code, regarding the custom chart, starts once the user clicks the “Update chart” button. The function that’s connected to the “Update button” gets called and the data from the ace text editor are retrieved. The custom chart changes its values according to the formatted data and gets updated so the user can view the results.

4.3.3.2 The Standard Chart

Table 4.1: Standard available metrics for tools and workflows

Metrics	Tool	Workflow
Upvotes	✓	✓
Downvotes	✓	✓
Downloads	✓	✓
CPU Time	✓	
Memory	✓	
Export Variety		✓

The standard chart contains the standard metrics that a user can use to evaluate whether that specific tool or workflow is credible and can meet their needs. The findings from the systematic literature review helped to formulate the necessary values that should be viewed by the

user in order to make their assessment. Therefore, in the tool statistics tab, the standard chart contains metric values for the number of upvotes, downvotes, and downloads of the selected tool, as well its CPU time and memory allocation upon its execution (Figure 4.14). In the workflow statistics tab, the standard chart contains metric values for the number of upvotes, downvotes, and downloads of the selected workflow, as well its export variety (Figure 4.15). The available metrics are represented in Table 4.1.

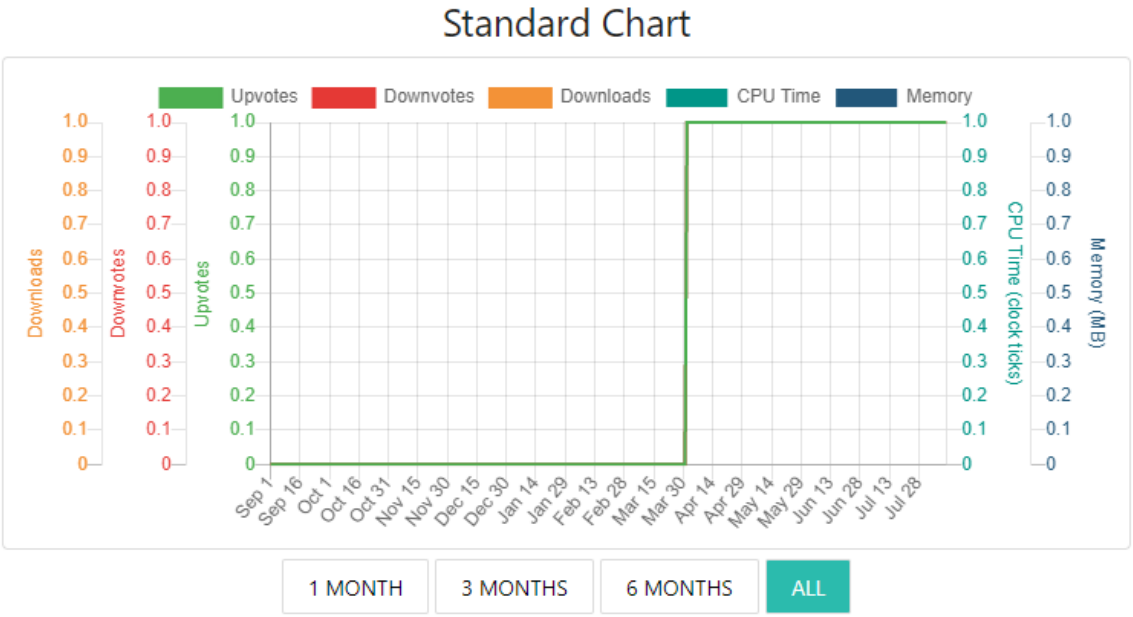

Figure 4.14: Tool Standard chart

Regarding the chart's functions, the user can change the time range of the x-axis on the chart and see the results of the past month, past 3 months, past 6 months, or all the values since the research object's creation date. The time can change by selecting one of the buttons: "1 month", "3 months", "6 months", or "all" that exist right below the standard chart. Furthermore, the user can hide or show any of the values that appear on the standard chart by clicking on the respective label on top of the chart. Moreover, tooltips are also included; hovering over the chart, a tooltip appears displaying all the values (y-axis) and each value's number at that specific date (x-axis).

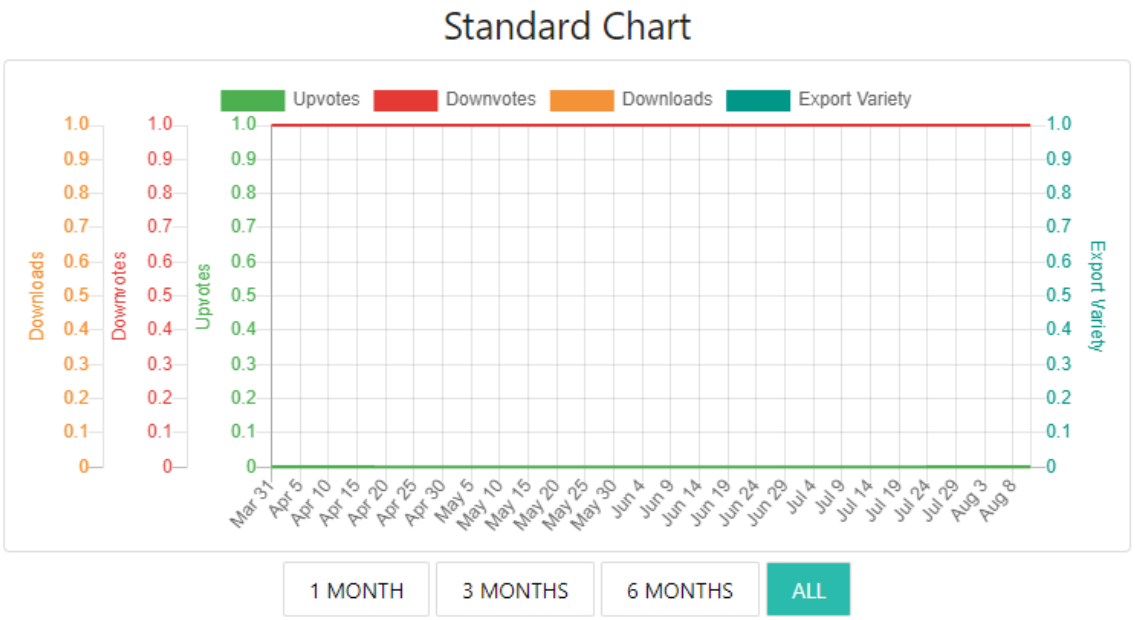

Figure 4.15: Workflow Standard chart

The standard chart is defined as a canvas object in a HTML file. A JavaScript [246] file named “statistics.js” also exists and contains all the necessary functions that are needed for the charts to be loaded to the webpage. After the user clicks on a tool or workflow, a function in the JavaScript file gets called to load the standard chart with the research object’s necessary metric values. The research object’s metric values are retrieved from the database using the Django framework. For each tool, the retrieved values include the date the tool was created, its upvotes, downvotes, downloads, CPU time, and memory allocation with the timestamps that each of those values occurred. For each workflow, the retrieved values include the date the workflow was created, its upvotes, downvotes, downloads, and export variety with the timestamps that each of those values occurred. Those values are transferred, with the help of the Angular Framework [212], to the “statistics.js” JavaScript file so the corresponding function can load them into the canvas and present the chart to the user.

4.3.3.3 The Comments Chart

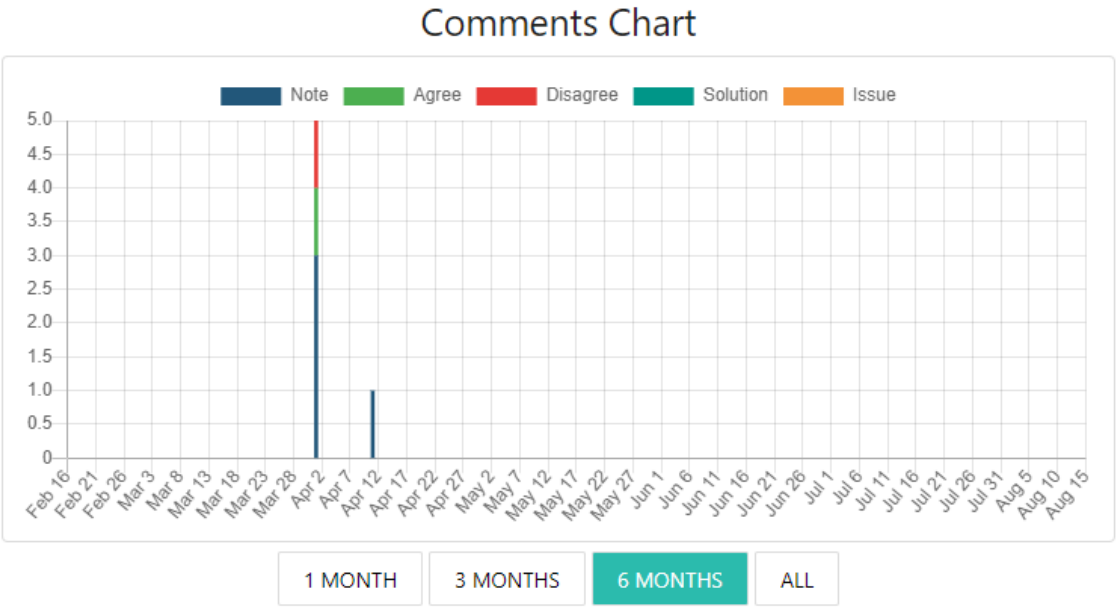

Figure 4.16: The Comments chart in Statistics

The comments chart represents an overview of the discussion tab that each research object has (Figure 4.16). The discussion tab, on each tool and workflow, offers users the ability to post questions or remarks regarding that research object and open a discussion panel where other users are able to comment and reply to each other. Upon each discussion, the user can label their comment as a way to present their purpose of posting them. The available labels are “Note”, “Agree”, “Disagree”, “Solution”, and “Issue”. Those labels can be monitored, and an overall evaluation of the research object can be extracted from them. Therefore, the comments chart captures the labels of those comments and the dates where each of those occurred and presents them to the user.

Regarding the chart’s functions, the user can change the time range of the x-axis on the chart and see the results of the past month, past 3 months, past 6 months, or all the values since the research object’s creation date. The time can change by selecting one of the buttons: “1 month”, “3 months”, “6 months”, or “all” that exist right below the comments chart. Furthermore, the user can hide or show any of the values that appear on the comments chart by clicking on the respective label on top of the chart. Moreover, tooltips are also included; hovering over the chart, a tooltip appears displaying all the values (y-axis) and each value’s number at that specific date (x-axis).

The comments chart is defined as a canvas object in the HTML file. After the user clicks on a tool or workflow, a function in the JavaScript file “statistics.js” gets called to load the comments chart with the research object’s necessary metric values. The research object’s metric values are retrieved from the database using the Django framework. For each tool or workflow, the retrieved values include the date the research object was created and its comments thread. Those values are transferred, with the help of the Angular Framework, to the “statistics.js” JavaScript file. The comments thread is formatted, and for each comment its label and creation date are extracted. The corresponding function loads those values into the canvas and present the chart to the user.

4.4 Implementation of the Profile Building and Evaluation of Users

The developers of many platforms pay attention to their users and search for ways to reward them through gamification techniques, especially when the upkeep of these platforms depends on their contributions. The findings from the systematic literature review, regarding the user profile building and evaluation, revealed that Github and Stack Overflow are among the most widely used social networking platforms because of their features and usefulness. In those platforms, the users gain reputation based on their contributions, and differentiate from other community members. Their potential influence is transformed into real effect and therefore, can be evaluated by others. Numerous studies have demonstrated the power of reputation points and community activity patterns in predicting a system's long-term value and users' perceptions of contributions and platform expertise.

To answer the third research question about finding a solution for the problem of profile building and evaluation of users, a set of metrics were calculated that represented the overall activity of the user on the OpenBio platform. Those metrics were created to assist users in evaluating other users based on their contributions.

4.4.1 Implementation of the User Metrics

The user metrics are visible in two different sections of the OpenBio platform. They are available on the public profile of the user, and they are also available on the private profile page, that each user has to edit their info.

4.4.1.1 The Public Profile User Metrics

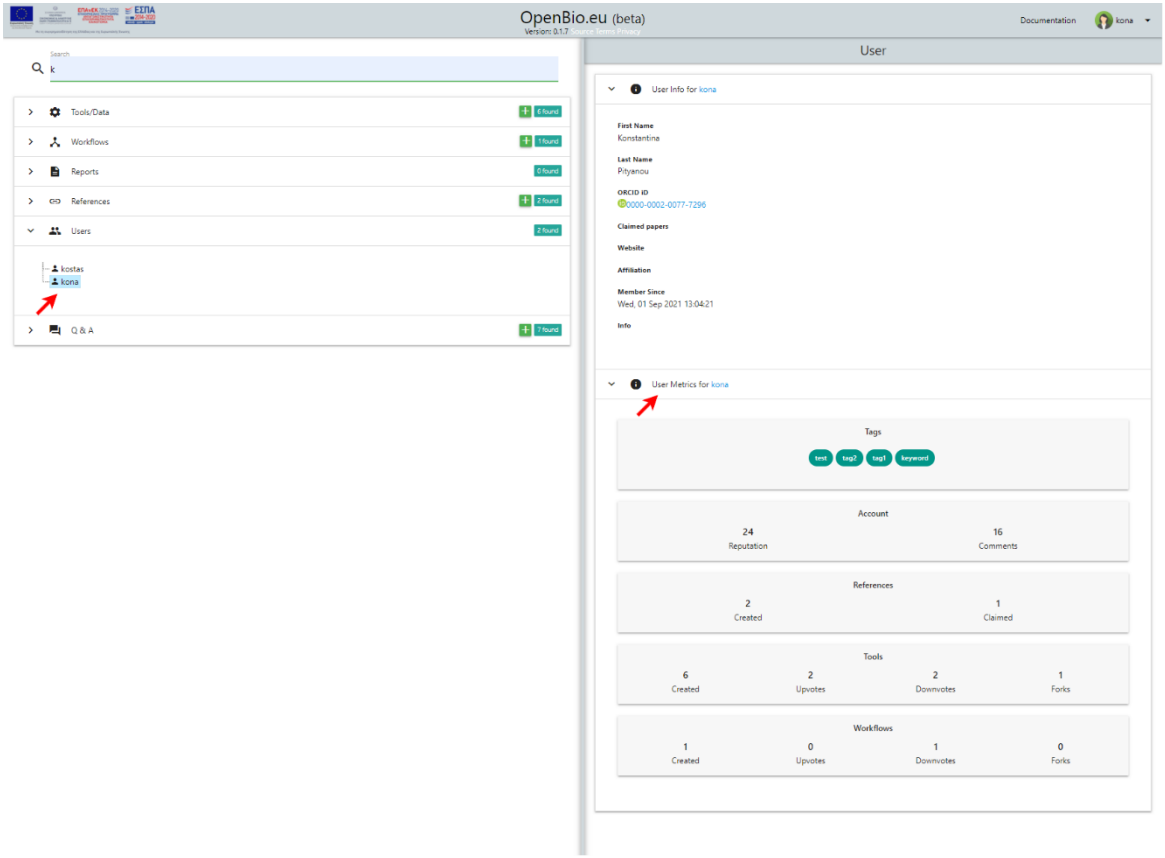

Figure 4.17: The public profile user metrics

The user can view the metrics of another user by clicking on their profile and navigate to the tab “User Metrics” (Figure 4.17). The “User Metrics” tab contains all the available metrics that can assist another user to evaluate them and see their contributions to various parts of the OpenBio platform. The metrics are separated into five different sections: the “Tags”, “Account”, “References”, “Tools”, and “Workflows”. Upon hovering on each metric, a tooltip appears that briefly explains the use of that metric.

The “Tags” section refers to the keywords of the user’s tools and workflows. The user enters tags upon creating or editing a tool or workflow. The tags’ purpose is to navigate other users with the help of keywords to better understand the purpose and goal of a specific tool or workflow. Showing the tags that the user has enters in their tools and workflows can be useful because it shows their main interests and expertise on different fields.

The “Account” section contains information regarding the contribution of the user on the platform. It includes the “Reputation” and “Comments” metrics, where the reputation refers to the level of expertise the user has based on their evaluated contributions to the platform, and the comments refer to the number of comments the user has posted on the discussion tabs of tools and workflows, or on the “Q&A” section. Reputation is very important and plays a major role on the evaluation of the user as the systematic literature review results revealed. It can assist users to evaluate others and decide whether to count on their opinions and contributions. The number of comments can offer users the ability to evaluate others on their level of contribution and how often that occurs.

The “References” section contains information regarding the references that exist on the OpenBio platform. It includes the number of the references the user has created, and the number of references the user has claimed with their ORCID account. These metrics can be useful because they can assist users to analyze the level of contribution of other users and the claimed references can show their level of expertise in the research field.

The “Tools” section contains information regarding the user’s created tools. It includes the number of tools the user has created, the number of upvotes and downvotes those tools have received, and the number of forks that they have. This section offers others the ability to examine the contribution of tools of the user on the platform. The number of upvotes and downvotes the user’s tools have received can reveal their expertise on the tools that they have created, and the number of forks shows the effectiveness and usefulness of those tools.

The “Workflows” section contains information regarding the user’s created workflows. It includes the number of workflows the user has created, the number of upvotes and downvotes those workflows have received, and the number of forks that they have. This section offers others the ability to examine the contribution of workflows of the user on the platform. The number of upvotes and downvotes the user’s workflows have received can reveal their expertise on the

workflows that they have created, and the number of forks shows the effectiveness and usefulness of those workflows.

4.4.1.2 The Private Profile User Metrics

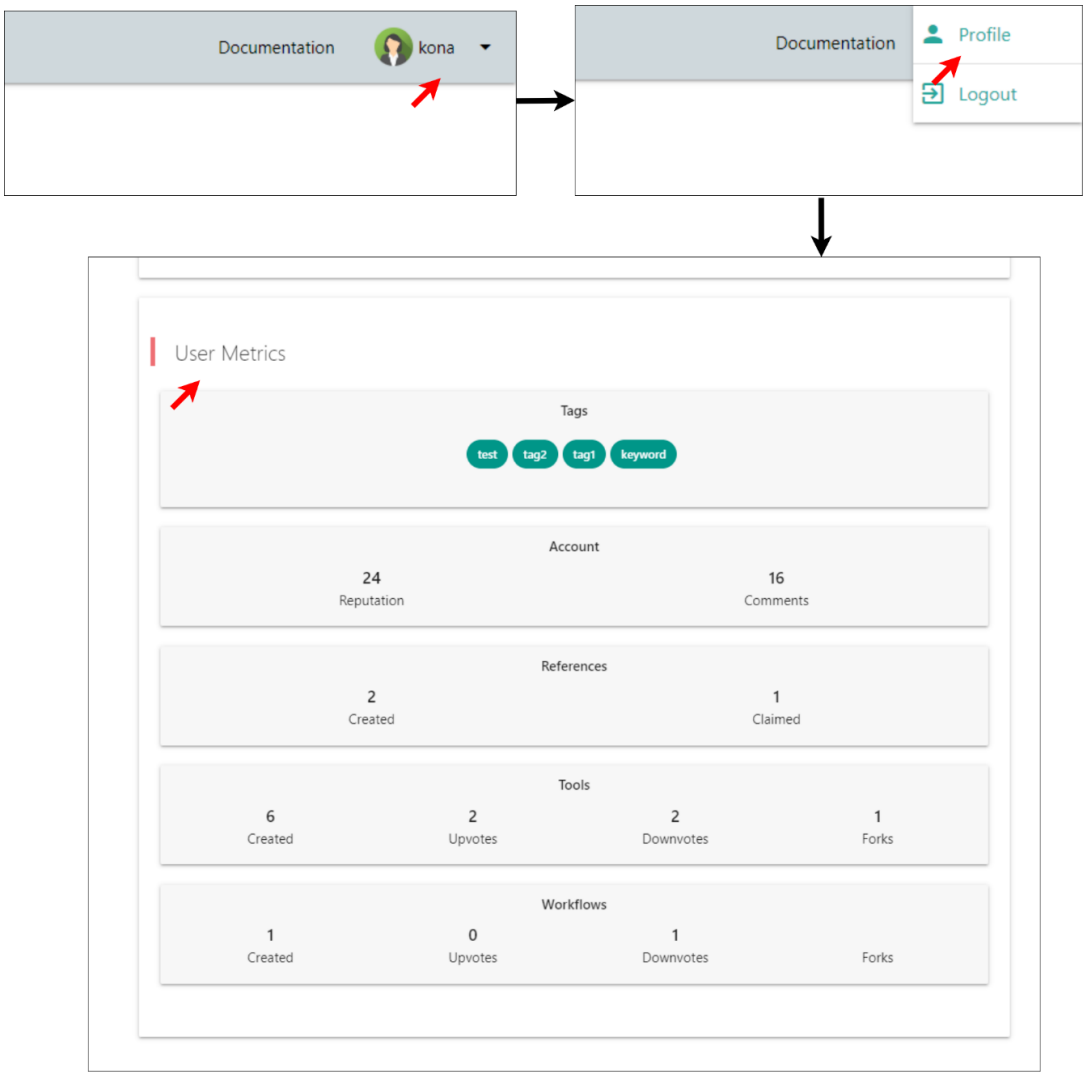

Figure 4.18: The private profile user metrics

The user can view their own metrics by clicking on their profile and navigate to the tab "User Metrics" (Figure 4.18). The "User Metrics" tab contains all the available metrics that can assist them to evaluate their contributions to various parts of the OpenBio platform. The tab contains the same categories and metrics that exist in the public profile user metrics.

The “Tags” section can show the user the fields in which they contribute the most on the OpenBio platform. The “Account” section is very important and can show to the user their popularity points and therefore, their influence level on the platform. The comments also allow them to see their contribution level. The “References” section helps the user to keep track of their references record and the claims that they had with connection to ORCID. The “Tools” section allows the user to examine their contribution in tools. The number of upvotes and downvotes can help them evaluate and improve themselves. The number of forks can assist them to see the effectiveness and usefulness of their created tools. The “Workflows” section allows the user to examine their contribution in workflows. The number of upvotes and downvotes can help them evaluate and improve themselves. The number of forks can assist them to see the effectiveness and usefulness of their created workflows.

4.4.1.3 The Code Implementation in User Metrics

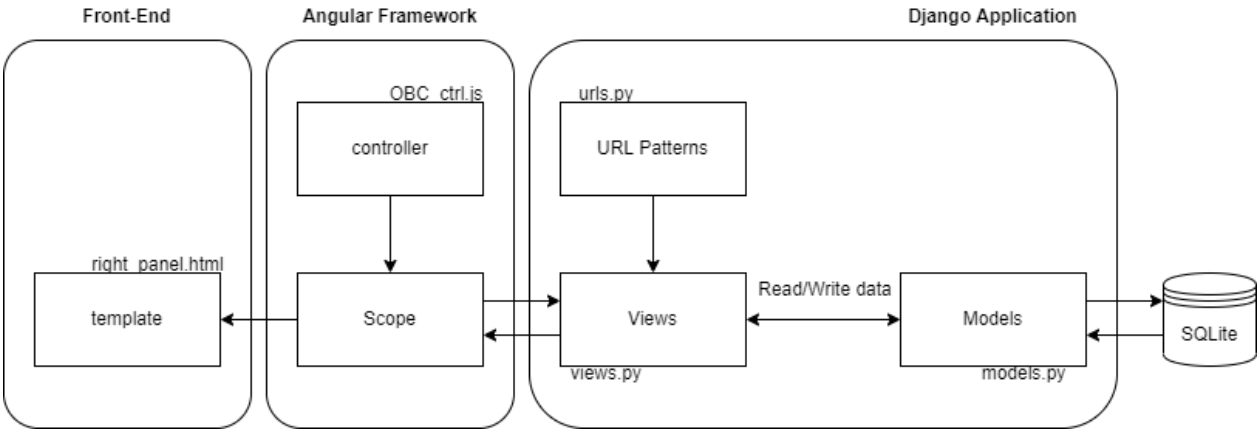

Figure 4.19: The user metrics system architecture

The implementation of the code regarding the user metrics functions in the same way for both the public and private profile of the user. The necessary information that is needed for the metrics is retrieved from the database using Django. The variables that are accessed refer to the classes of tools, workflows, references, users, comments, upvotes, and downvotes. In the “views.py” Python file the retrieved data are modified and assigned to new variables. Those variables are retrieved from the OBC controller JavaScript file using Angular. Those retrieved variables are assigned to variables under the Angular scope and get bind to the front-end HTML

template that refers to the user metrics on the webpage. Any change automatically updates the corresponding variable, and the results are presented to the user. The architecture of the front-end, Angular, and Django connection is present in Figure 4.19.

Table 4.2: Points analysis in the reputation formula

	Tool voted up	Workflow voted up	Comment voted up	Tool voted down	Workflow voted down	Comment voted down
Points	+10	+10	+10	-2	-2	-2

The reputation of each user is formed based on their overall activity and level of contribution to the OpenBio platform. The formula from which the reputation is calculated was inspired by the Stack Overflow reputation system [247]. Therefore, the reputation of each user is affected by the evaluation of their created tools, workflows, and comments. Table 4.2 presents the given points for each evaluation.

4.5 The User Experience Questionnaire Evaluation

Ensuring that a product or service has a good user experience is essential for producing successful products or services. Because they have distinct requirements or varied capabilities or skills to utilize the product, various users or groups of users may evaluate the same product's user experience somewhat differently. Therefore, using validated questionnaires to do such assessments is an effective and affordable strategy.

In ISO 9241-210 [248], the term “user experience” is described as a person's perceptions and responses that emerge from the usage or expected use of a product, system, or service. Throughout this approach, user experience is viewed as a comprehensive notion that encompasses all different kinds of emotional, cognitive, and physical responses to the actual or even just hypothetical use of a product. However, in the evaluation of a product’s quality, this definition, which is rather wide and abstract, is of little use.

4.5.1 Overview

The User Experience Questionnaire (UEQ) [249] is a questionnaire that can measure user experience quickly and directly. The UEQ takes into account hedonic and pragmatic quality elements [250]. The UEQ's initial German version was developed in 2005. In order to guarantee the produced scales' applicability, a data analytical technique was applied, indicating the scales were developed from data pertaining to a larger pool of objects. Each scale outlines a certain facet of an interactive product's quality.

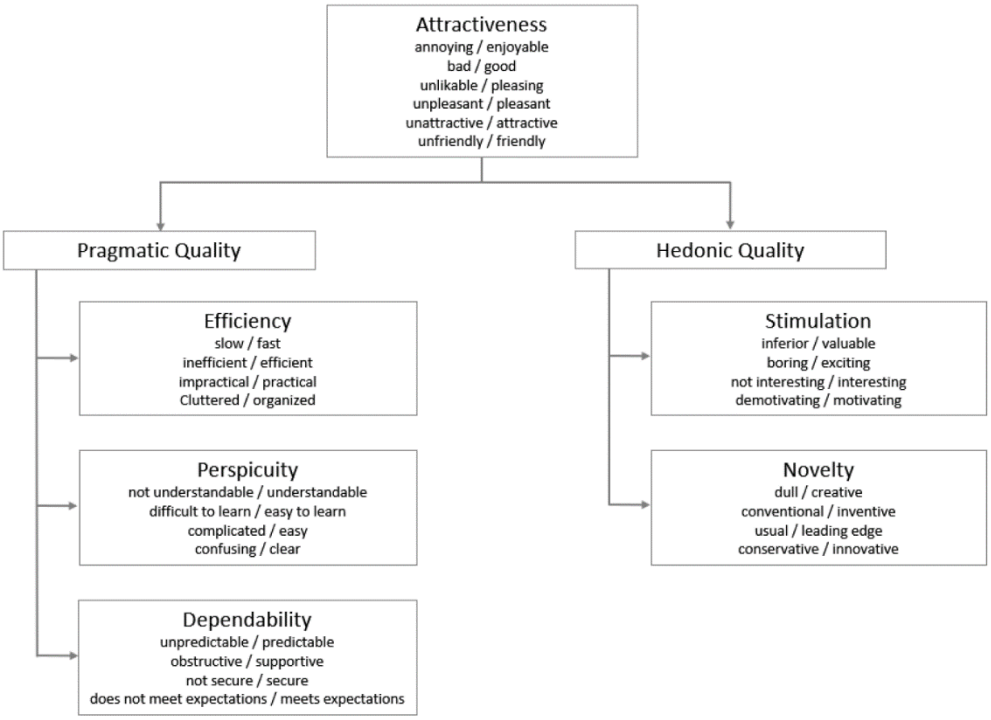

Figure 4.20: Scale structure of the UEQ

The UEQ initially included several user experience-related possible elements. Following extensive research, the final version of the questionnaire was created, and it had six scales with a total of 26 items: attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. The proposed scale structure of the UEQ in its final form is shown in Figure 4.20.

	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

Figure 4.21: The User Experience Questionnaire

The items take the shape of a semantic difference, with two terms with opposing meanings standing in for each item. The terms are randomly arranged for each item, so that for a scale, half the items begin with the positive phrase and the other half with the negative phrase. To lessen the regression coefficients bias for these kinds of items, the UEQ utilizes a seven-stage scale. The final version of the 26 evaluation items to be answered is presented in Figure 4.21.

4.5.2 The UEQ Evaluation Process

The User Experience Questionnaire was used in order to evaluate the user experience of the implemented components in OpenBio and validate whether those components could be applied as solutions to answer the research questions that were formed as part of this thesis. Because the OpenBio platform is designed for bioinformaticians, inclusion criteria were set for the selection of the participants. More specifically, in order to be deemed adequate to participate in the evaluation, the participant had to have a good knowledge of computer and bioinformatics skills. The inclusion criteria did not consider the age, ethnicity, or gender as important factors for the participation. The application of those criteria resulted in the gathering of 8 participants, ages between 26 and 30 and fair knowledge of bioinformatic skills.

After the participants were introduced to OpenBio, its purpose, and basic functions, they were asked to navigate through the platform and focus on the added components. After they were satisfied with the browsing on the OpenBio platform, they were instructed to fill out the User Experience Questionnaire.

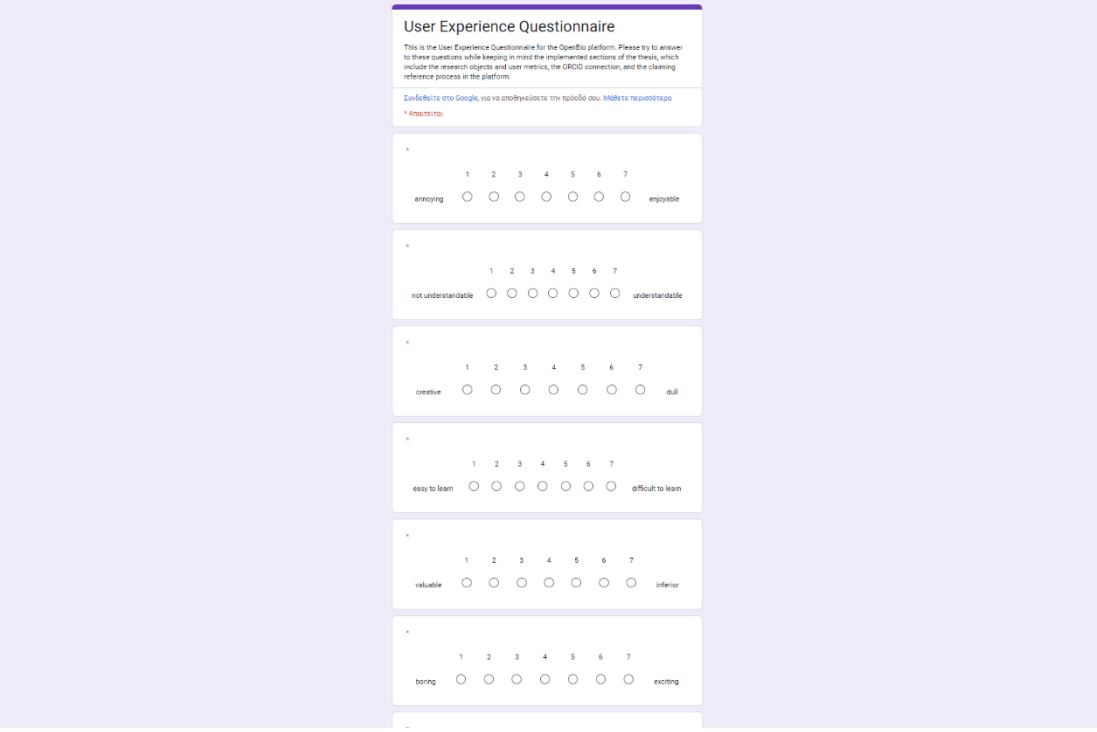

Figure 4.22: The UEQ in Google Forms

The introduction to OpenBio and description of the evaluation process was explained to the participants through Skype. The navigation on the platform was done by the participants on the OpenBio platform that is up and running under the URL: https://www.openbio.eu/platform/. The UEQ was filled out with the help of Google Forms. More specifically, the 26 pairs of contrasting attributes were added in a newly created Google Form and was given to the participants through a link. The form was anonymous, and all the 26 questions were considered mandatory to be answered. Figure 4.22 presents the adaption of the UEQ in a Google Form.

4.5.3 The UEQ Data

Table 4.3: The UEQ answers

Items																									
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
5	6	1	2	1	7	7	6	1	1	6	1	6	7	7	7	1	1	1	7	2	7	1	1	1	6
7	7	3	1	2	6	6	7	1	2	7	1	7	7	5	6	1	2	2	7	1	7	1	2	1	5
4	3	3	5	2	4	5	7	2	5	4	4	3	6	4	5	4	4	5	3	5	3	2	3	4	4
3	3	4	3	5	3	3	5	3	4	3	5	4	4	3	3	5	5	5	3	5	3	5	5	5	3
6	6	3	3	1	7	7	7	1	2	5	1	7	7	5	7	2	1	2	6	1	7	1	2	1	5
5	5	3	3	3	3	3	5	3	5	5	4	4	4	3	4	3	4	5	3	4	4	3	4	4	3
5	6	4	3	3	5	5	5	2	3	5	3	6	5	4	5	4	4	5	5	4	5	3	3	3	5
5	4	2	2	2	6	5	6	3	3	5	2	5	5	5	5	3	3	2	5	3	6	3	2	2	5

The UEQ answers were obtained after all 8 participants had completed the questionnaire. The recorded answers for each question is presented in Table 4.3. Each row represents the participant, and each column represents the question (1-26). Each cell represents the answer (1-7), where the participant gave for that question.

Table 4.4: Transformed value per answer

	-3	-2	-1	0	+1	+2	+3	
complicated	○	○	○	○	○	○	○	easy

As explained, each item has the appearance of a semantic difference, with two terms with opposing meanings in each side. Scales for the items range from -3 to +3. Therefore, the most negative response is represented by -3, the neutral response by 0, and the most positive response

by +3 (Table 4.4).

Table 4.5: Transformed values per item

Items																								
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
1	2	3	2	3	3	3	2	3	3	2	3	2	3	3	3	3	3	3	3	2	3	3	3	3	2
3	3	1	3	2	2	2	3	3	2	3	3	3	3	1	2	3	2	2	3	3	3	3	2	3	1
0	-1	1	-1	2	0	1	3	2	-1	0	0	-1	2	0	1	0	0	-1	-1	-1	-1	2	1	0	0
-1	-1	0	1	-1	-1	-1	1	1	0	-1	-1	0	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	2	1	1	3	3	3	3	3	2	1	3	3	3	1	3	2	3	2	2	3	3	3	2	3	1
1	1	1	1	1	-1	-1	1	1	-1	1	0	0	0	-1	0	1	0	-1	-1	0	0	1	0	0	-1
1	2	0	1	1	1	1	1	2	1	1	1	2	1	0	1	0	0	-1	1	0	1	1	1	1	1
1	0	2	2	2	2	1	2	1	1	1	2	1	1	1	1	1	1	2	1	1	2	1	2	2	1

Table 4.6: Scale means per participant

Attractiveness	Perspicuity	Efficiency	Dependability	Stimulation	Novelty
2.67	2.00	3.00	2.50	3.00	2.75
2.67	3.00	3.00	2.75	2.00	1.25
0.67	-1.00	0.50	0.50	0.75	0.00
-0.83	-0.25	-0.50	-0.50	-1.00	-0.50
2.67	2.25	2.75	2.00	3.00	1.25
0.17	0.50	0.25	0.50	-0.25	-0.50
1.00	1.25	1.25	0.25	0.75	0.50
1.50	1.00	1.25	1.50	1.50	1.25

Table 4.5 presents the transformed values per item, according to the given answers in Table 4.3. The +3 represents the most positive and -3 the most negative value. From those transformed values, the scale means per participant were calculated and categorized in the 6 scales of the UEQ. Table 4.6 presents the scale means per participant.

4.5.4 The UEQ Evaluation Results

Data analysis was applied to the UEQ data in order to estimate the effectiveness of user experience in the implementation of this thesis. Although the UEQ does not produce an overall

score for the user experience, the calculation of certain values can allow the detection of outliers in the evaluation, and therefore, use these big deviations to follow a probable result.

Table 4.7: Calculated values per item

Item	Mean	Variance	Std. Dev.	No.	Left	Right	Scale	
1	1.0	1.4	1.2	8	annoying	enjoyable	Attractiveness	
2	1.0	2.3	1.5	8	not understandable	understandable	Perspicuity	
3	1.1	1.0	1.0	8	creative	dull	Novelty	
4	1.3	1.4	1.2	8	easy to learn	difficult to learn	Perspicuity	
5	1.6	1.7	1.3	8	valuable	inferior	Stimulation	
6	1.1	2.7	1.6	8	boring	exciting	Stimulation	
7	1.1	2.4	1.6	8	not interesting	interesting	Stimulation	
8	2.0	0.9	0.9	8	unpredictable	predictable	Dependability	
9	2.0	0.9	0.9	8	fast	slow	Efficiency	
10	0.9	2.1	1.5	8	inventive	conventional	Novelty	
11	1.0	1.4	1.2	8	obstructive	supportive	Dependability	
12	1.4	2.6	1.6	8	good	bad	Attractiveness	
13	1.3	2.2	1.5	8	complicated	easy	Perspicuity	
14	1.6	1.7	1.3	8	unlikable	pleasing	Attractiveness	
15	0.5	1.7	1.3	8	usual	leading edge	Novelty	
16	1.3	1.9	1.4	8	unpleasant	pleasant	Attractiveness	
17	1.1	2.1	1.5	8	secure	not secure	Dependability	
18	1.0	2.3	1.5	8	motivating	demotivating	Stimulation	
19	0.6	3.1	1.8	8	meets expectations	does not meet expectations	Dependability	
20	0.9	3.0	1.7	8	inefficient	efficient	Efficiency	
21	0.9	2.7	1.6	8	clear	confusing	Perspicuity	
22	1.3	3.1	1.8	8	impractical	practical	Efficiency	
23	1.6	2.0	1.4	8	organized	cluttered	Efficiency	
24	1.3	1.6	1.3	8	attractive	unattractive	Attractiveness	
25	1.4	2.6	1.6	8	friendly	unfriendly	Attractiveness	
26	0.5	1.1	1.1	8	conservative	innovative	Novelty	

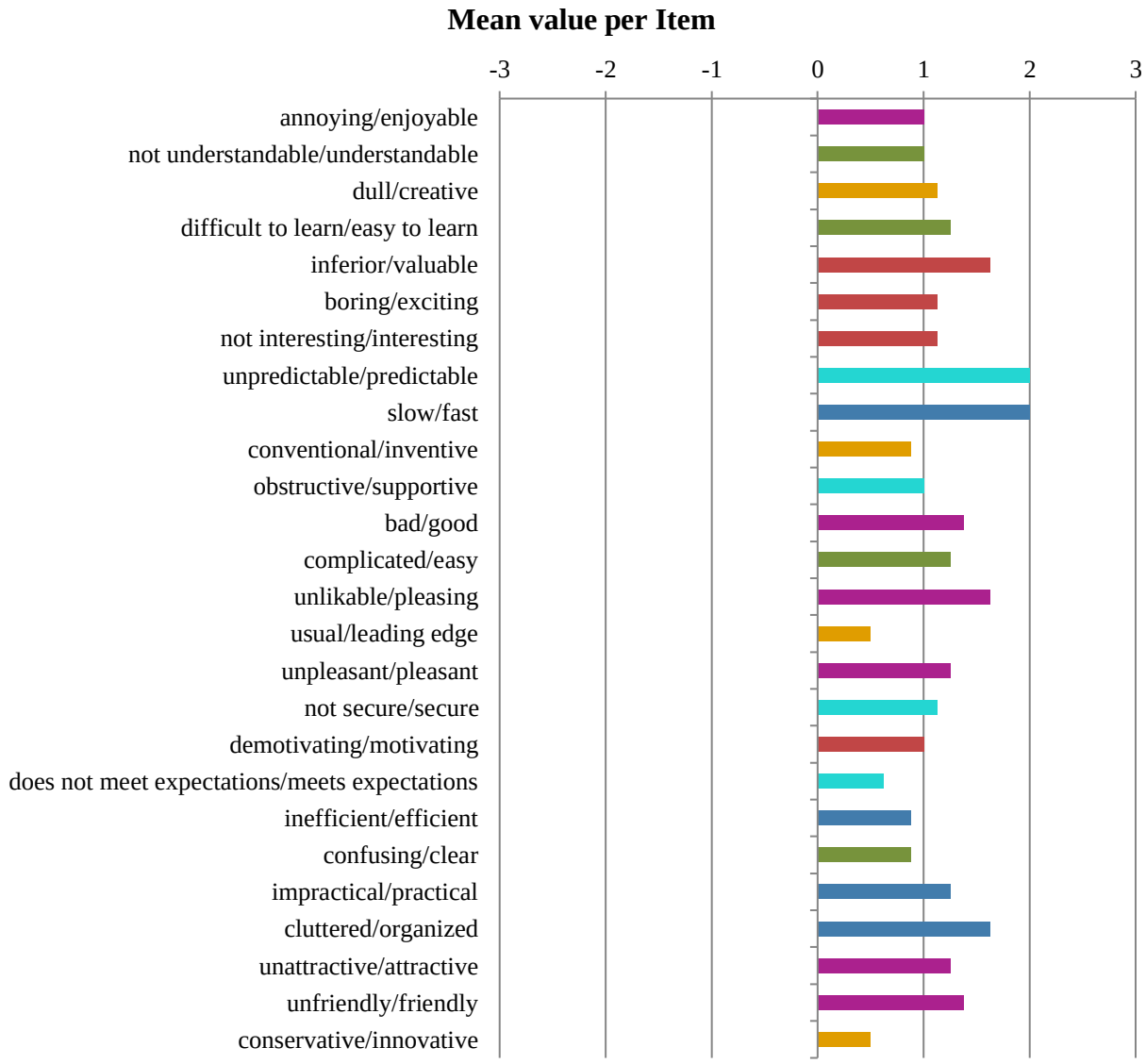

Figure 4.23: Comparison of mean values per item

Table 4.7 presents, for each item in the UEQ, its mean value, variance, standard deviation, number of inputs, its left and right term, and its scale category. Based on the mean value, Figure 4.23 presents the comparison of mean values per item. Although the range of scales is between -3 and +3, in real-life applications it is extremely unlikely to observe values above +2 or below -2. Therefore, values < -0.8 represent a negative evaluation, values between -0.8 and 0.8 represent a neutral evaluation, and values > 0.8 represent a positive evaluation. From Table 4.7 and Figure 4.23, the overall user experience evaluation proves to be positive; most items have mean values >

0.8, with the exception of “leading edge”, “meets expectations”, and “innovative”, where their means, although > 0, result in a neutral evaluation.

Table 4.8: Mean and Variance values per scale

UEQ Scales (Mean and Variance)		
Attractiveness	1.313	1.71
Perspicuity	1.094	1.77
Efficiency	1.438	1.82
Dependability	1.188	1.37
Stimulation	1.219	2.08
Novelty	0.750	1.21

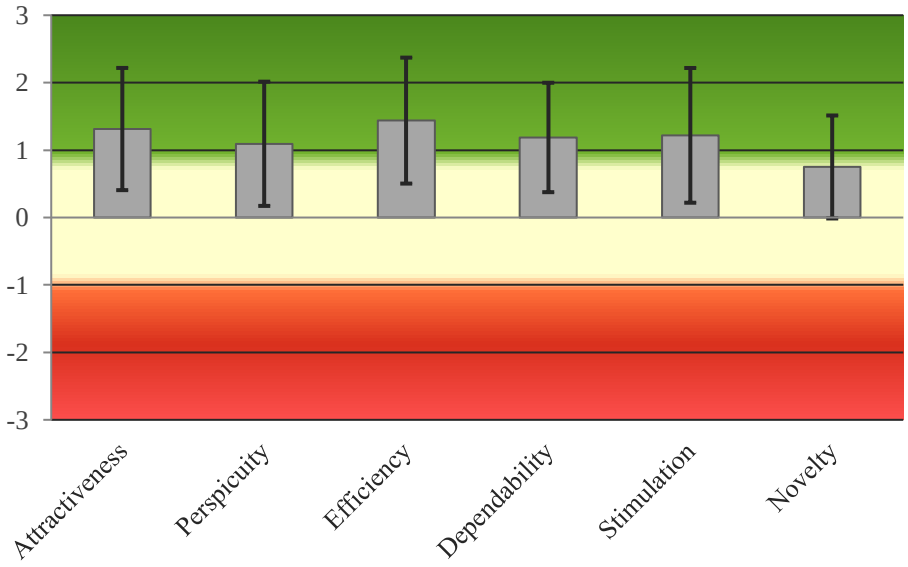

Figure 4.24: Comparison of mean values per scale

Table 4.8 presents the calculated mean and variance values per scale; most scales seem to have an overall positive evaluation, with exception of the scale “Novelty”, which produces a neutral result. Figure 4.24 presents the comparison between the mean values per scale; the “Efficiency” scale seems to have the most positive results, and the “Attractiveness” scale follows shortly after.

Table 4.9: Mean values of Attractiveness, Pragmatic, and Hedonic Quality

Pragmatic and Hedonic Quality	
Attractiveness	1.31
Pragmatic Quality	1.24
Hedonic Quality	0.98

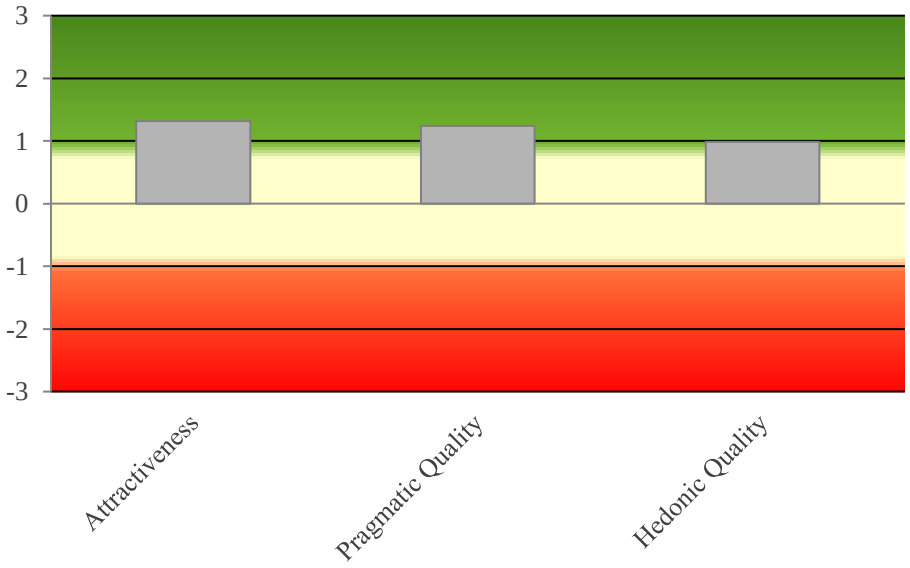

Figure 4.25: Comparison of mean values of Attractiveness, Pragmatic, and Hedonic Quality

The scales of the UEQ can be grouped into “Pragmatic Quality” (Perspicuity, Efficiency, Dependability) and “Hedonic Quality” (Stimulation, Originality). Pragmatic quality describes the task related quality aspects, while Hedonic quality describes the non-task related quality aspects. Table 4.9 presents the calculated mean values for the groups of “Attractiveness”, “Pragmatic Quality”, and “Hedonic Quality”. Based on those values, the three different groups in the UEQ seem to produce a positive evaluation. Based on the comparison between the groups, as presented in Figure 4.25, the most positive evaluation belongs to the “Attractiveness”, followed by the “Pragmatic Quality”, and last the “Hedonic Quality”.

Table 4.10: Confidence interval of 5% per item mean

Confidence interval (p=0.05) per item						
Item	**Mean**	**Std. Dev.**	**N**	**Confidence**	**Confidence interval**	
1	1.000	1.195	8	0.828	0.172	1.828
2	1.000	1.512	8	1.048	-0.048	2.048
3	1.125	0.991	8	0.687	0.438	1.812
4	1.250	1.165	8	0.807	0.443	2.057
5	1.625	1.302	8	0.903	0.722	2.528
6	1.125	1.642	8	1.138	-0.013	2.263
7	1.125	1.553	8	1.076	0.049	2.201
8	2.000	0.926	8	0.642	1.358	2.642
9	2.000	0.926	8	0.642	1.358	2.642
10	0.875	1.458	8	1.010	-0.135	1.885
11	1.000	1.195	8	0.828	0.172	1.828
12	1.375	1.598	8	1.107	0.268	2.482
13	1.250	1.488	8	1.031	0.219	2.281
14	1.625	1.302	8	0.903	0.722	2.528
15	0.500	1.309	8	0.907	-0.407	1.407
16	1.250	1.389	8	0.962	0.288	2.212
17	1.125	1.458	8	1.010	0.115	2.135
18	1.000	1.512	8	1.048	-0.048	2.048
19	0.625	1.768	8	1.225	-0.600	1.850
20	0.875	1.727	8	1.197	-0.322	2.072
21	0.875	1.642	8	1.138	-0.263	2.013
22	1.250	1.753	8	1.214	0.036	2.464
23	1.625	1.408	8	0.976	0.649	2.601
24	1.250	1.282	8	0.888	0.362	2.138
25	1.375	1.598	8	1.107	0.268	2.482
26	0.500	1.069	8	0.741	-0.241	1.241

Table 4.11 Confidence interval of 5% per scale mean

Confidence intervals (p=0.05) per scale						
Scale	**Mean**	**Std. Dev.**	**N**	**Confidence**	**Confidence interval**	
Attractiveness	1.313	1.308	8	0.906	0.406	2.219
Perspicuity	1.094	1.329	8	0.921	0.173	2.015
Efficiency	1.438	1.348	8	0.934	0.503	2.372
Dependability	1.188	1.171	8	0.811	0.376	1.999
Stimulation	1.219	1.442	8	0.999	0.220	2.218

Novelty	0.750	1.102	8	0.764	-0.014	1.514

In order to measure the precision of the estimation of the scale mean per item and per scale, the confidence interval was also calculated. Regarding the confidence interval value, the smaller it is, the higher is the precision of the estimation and the presented results can be more trustfull. The width of the of the confidence interval depends on the number of available data and on how consistency the participants judged the evaluated product. Table 4.10 and Table 4.11 present the 5% confidence intervals for the scale means and the means of the single items. Due to low number of participants, the margin error proves that the evaluation by a different group of participants, can possibly lead to different results. Therefore, it should be noted that in order to be more precise and certain of the result's output, the evaluation could benefit with the inclusion of more participants.

Table 4.12: Guttman's lambda-2 Coefficient per scale

Attractiveness	
Lambda1	0.80823905

Items	**Covar²**
1, 12	1.89
1, 14	0.77
1, 16	1.00
1, 24	0.77
1, 25	1.89
12, 14	2.30
12, 16	3.17
12, 24	2.74
12, 25	4.99
14, 16	2.16
14, 24	1.49
14, 25	2.30
16, 24	2.07
16, 25	3.17
24, 25	2.74
Sum	33.44
Lambda2	**0.95**

Perspicuity	
Lambda1	0.69740998

Items	**Covar²**
2, 4	1.31
2, 13	4.00
2, 21	4.00
4, 13	1.47
4 , 21	1.75
13, 21	4.75
Sum	17.27
Lambda2	**0.94**

Efficiency	
Lambda1	0.69410319

Items	**Covar²**
9, 20	1.27
9, 22	1.00
9, 23	1.00
20, 22	6.41
20, 23	2.49
22, 23	2.54
Sum	14.70
Lambda2	**0.91**

Dependability	
Lambda1	0.6563518

Items	**Covar²**
8, 11	0.14
8, 17	0.39
8, 19	0.56
11, 17	1.89
11, 19	1.56
17, 19	3.69
Sum	8.24
Lambda2	**0.87**

Stimulation

Novelty

Lambda1	0.7267848

Items	Covar^2
5, 6	2.39
5, 7	2.39
5, 18	2.25
6, 7	4.45
6, 18	4.00
7, 18	3.52
Sum	19.00
Lambda2	**0.94**

Lambda1	0.6930147

Items	Covar^2
3, 10	0.41
3, 15	0.88
3, 26	0.32
10, 15	2.07
10, 26	1.41
15, 26	1.27
Sum	6.35
Lambda2	**0.90**

In order to estimate the reliability of the UEQ answers, the Guttman's Lambda-2 Coefficient was used. In general, the items in the UEQ that belong to the same scale should show a high correlation. The Lambda2 statistic can show what variance is due to true scores. Table 4.12 presents the calculation of the Lambda2 for each scale; the calculations show that the lambda values are > 0.8, and therefore the scales have a high percentage that their answers are due to true scores.

5 Discussion

This thesis aimed to find solutions to some vital problems in the domain of bioinformatics by contributing the results to the OpenBio platform. The implementation included the development of components as a mean to answer the three research questions that were formed as part of this thesis.

The first research question was: “How can we effectively solve the problem of misattribution?”. Since misattribution is an issue that the literature review has widely addressed, this research question was formulated in order to examine an appropriate solution to this problem. Researchers have underlined the numerous problems that can result from not acknowledging an author’s contribution to scientific fields, and especially in bioinformatics. In an effort to try and overcome the problem of misattribution, many methods have been developed including the manual labeling, or the development of corresponding algorithms. Although every approach can be beneficiary in the appropriate context, the ORCID authentication seem to be an effective method to authority control across different repositories and digital libraries. Consequently, the implementation of this thesis considered to examine whether the connection of OpenBio with ORCID can effectively minimize misattribution. The results from the UEQ indicate positive outcomes in the user experience of considering ORCID as an effective method to address this problem. ORCID can identify the identities of researchers from a wide range of backgrounds, overcoming the weaknesses of existing authentication sources, and therefore may be viewed as an efficient method to overcome the problem of misattribution.

The second research question was: “How can we approach the evaluation, in terms of effectiveness and efficiency, of a tool or workflow that exists in a repository?”. Literature review indicates the importance of efficient and usable sets of research objects, such as tools and workflows, that can help discover and analyze the biological insights available from the massive amounts of complicated sets of data. Given the significance of bioinformatics in several biological and biomedical research, great effort should be given into creating workflow systems and tools that are maintainable and reproducible. A scientist should be given the freedom to choose the research tool that best fits their needs. A set of standards for evaluating various bioinformatics workflow management tools has been presented in various studies. Therefore, in order to approach a method, under which a tool or workflow can be evaluated, a set of various metrics were developed for each tool and workflow that exists on the OpenBio platform. Those metrics aimed

to assist users to evaluate the effectiveness and efficiency of a tool or workflow, and make an informed decision of whether that research object could meet their requirements. The UEQ results support that the implemented components have a positive user experience, and the "Efficiency" section is produced the most positive results. By taking into account those results, the assumption that the presentation of the implemented metrics can help in the evaluation of a research object can be made.

The third research question was: "How can we evaluate the expertise and contribution of users and assist them in the problem of profile building?". The analysis and review of literature regarding the methods, in which the profile building and evaluation of users can be achieved, support the creation of distinctive participation features and reputation systems. Different set of metrics can indicate the level of expertise and contribution of a user on a platform. The literature supports that Q&A platforms can be a guidance for the embedment of such metrics, since the maintenance of those platforms is depended on the contributions of their users. Thus, in order to examine a method in which the evaluation of users can be effectively achieved and the problem of profile building can be eliminated, a set of metrics were developed and added on the OpenBio platform. The UEQ results suggest that the implementation of those metrics produce a positive user experience, which can indicate that the developed components are effective. Therefore, the calculated metrics that exist for every user on the OpenBio platform can possibly assist users in their profile building, and help other users to evaluate their expertise and level of contribution.

Despite the positive results regarding the user experience, it should be noted that the confidence interval calculations revealed that additional participants could be required to convey a more stable outcome. The Lambda-2 coefficient per scale, however, demonstrated that the responses provided in the UEQ had a high proportion of being accurate and reliable., and therefore, can be accounted as true.

6 Conclusion and Future Work

The main purpose of this Master Thesis was to create a solution to some vital problems in the domain of bioinformatics, by extending the repository of OpenBio, which includes tools and data that continuously tests and measures the included components. Three research questions were formed that would search for solutions about the problem of misattribution, the evaluation of research objects, and the problem of profile building and evaluation of users. Respectively, three components, aiming to solve those issues, were developed and embedded in the OpenBio platform.

The connection of OpenBio with ORCID was implemented with the help of the ORCID API and the Auth2.0 standard. The ORCID authentication assisted users to the linkage between the OpenBio platform and the wide database of ORCID which also includes their list of publications. With the implementation of this linkage, users were also able to claim their publications that exist on the OpenBio platform and semantically connect it to them. This connection assisted to the solution of misattribution, by linking research data to publications and software, in an open and semantically consistent manner, giving users the ability to take credit for their contributions to the scientific field.

For the purpose of this thesis a set of various metrics were also developed for the evaluation of tools and workflows that exist on the platform. For each tool, the users are presented with a chart that includes information on its standard metrics that include its number of upvotes, downvotes, and downloads, its CPU time, and memory allocation. For each workflow, its standard metrics include its number of upvotes, downvotes, downloads, and export variety. Each tool and workflow also include a chart that describes the overall activity of its discussion panel; the comment chart include the labels under which a comment was posted and its date. Furthermore, each tool and workflow include a custom chart with a text editor in which the user can create and visualize their own metrics with the ability to include multiple lines and bar charts in one single chart with multiple axis.

Another component that was added to the OpenBio platform, as part of this thesis, was the implementation of user metrics. The developed component offered users the ability to see and evaluate other users by the presentation of various metrics that showed their overall contribution and level of contribution on the platform. The user metrics include their used tags for tools and workflows; their overall reputation; the number of their posted comments; the numbers of their created and claimed references; the number of their created tools and the upvotes, downvotes, and

forks those have; and the number of their created workflows and the upvotes, downvotes, and forks those have.

In order to evaluate the implementation of those added components, the User Experience Questionnaire was used. The questionnaire was filled out by 8 participants after they had tested those components on the OpenBio platform. The overall results showed that the user experience regarding the added components produced positive results. The calculation of the confidence interval showed that more participants could be needed in order to express a more stable result. However, the Lambda-2 coefficient per scale showed that the answers that were given in the UEQ had high percentages to be true and valid.

Nevertheless, there are still a lot of research avenues that can be explored, such as the research and inclusion of more evaluation metrics on the research objects, or the creation of award badges on users as reward methods. Another future consideration could be the implementation of a usability evaluation that would include a large number of participants in order to produce more stable results.

7 References

[1] S. Hoon *et al.*, “Biopipe: A flexible framework for protocol-based bioinformatics analysis,” *Genome Res.*, vol. 13, no. 8, 2003, doi: 10.1101/gr.1363103.

[2] T. Oinn *et al.*, “Taverna: A tool for the composition and enactment of bioinformatics workflows,” *Bioinformatics*, vol. 20, no. 17, 2004, doi: 10.1093/bioinformatics/bth361.

[3] J. Goecks *et al.*, “Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences,” *Genome Biol.*, vol. 11, no. 8, 2010, doi: 10.1186/gb-2010-11-8-r86.

[4] F. Halbritter, H. J. Vaidya, and S. R. Tomlinson, “GeneProf: Analysis of high-throughput sequencing experiments,” *Nature Methods*, vol. 9, no. 1. 2012. doi: 10.1038/nmeth.1809.

[5] S. P. Shah *et al.*, “Pegasys: Software for executing and integrating analyses of biological sequences,” *BMC Bioinformatics*, vol. 5, 2004, doi: 10.1186/1471-2105-5-40.

[6] L. Goodstadt, “Ruffus: A lightweight python library for computational pipelines,” *Bioinformatics*, vol. 26, no. 21, 2010, doi: 10.1093/bioinformatics/btq524.

[7] M. Tanaka and O. Tatebe, “Pwrake: A parallel and distributed flexible workflow management tool for wide-area data intensive computing,” 2010. doi: 10.1145/1851476.1851529.

[8] K. Taura *et al.*, “Design and implementation of GXP make - A workflow system based on make,” in *Future Generation Computer Systems*, 2013, vol. 29, no. 2. doi: 10.1016/j.future.2011.05.026.

[9] S. P. Sadedin, B. Pope, and A. Oshlack, “Bpipe: A tool for running and managing bioinformatics pipelines,” *Bioinformatics*, vol. 28, no. 11, 2012, doi: 10.1093/bioinformatics/bts167.

[10] E. K. Samota and R. P. Davey, “Knowledge and Attitudes Among Life Scientists Toward Reproducibility Within Journal Articles: A Research Survey,” *Front. Res. Metrics Anal.*, vol. 6, Jun. 2021, doi: 10.3389/frma.2021.678554.

[11] “APT (software) - Wikipedia.” https://en.wikipedia.org/wiki/APT_(software) (accessed Aug. 17, 2022).

[12] “PyPI · The Python Package Index.” https://pypi.org/ (accessed Aug. 17, 2022).

[13] “npm.” https://www.npmjs.com/ (accessed Aug. 17, 2022).

[14] J. Ison *et al.*, “The bio.tools registry of software tools and data resources for the life

sciences," *Genome Biol.*, vol. 20, no. 1, 2019, doi: 10.1186/s13059-019-1772-6.

[15] D. Torre *et al.*, "Datasets2Tools, repository and search engine for bioinformatics datasets, tools and canned analyses," *Sci. Data*, vol. 5, 2018, doi: 10.1038/sdata.2018.23.

[16] A. Kanterakis *et al.*, "Towards reproducible bioinformatics: The openbio-c scientific workflow environment," 2019. doi: 10.1109/BIBE.2019.00047.

[17] "OpenBio.eu - Open, Social, Reproducible Science." https://www.openbio.eu/platform/ (accessed Aug. 03, 2022).

[18] "GitHub: Where the world builds software · GitHub." https://github.com/ (accessed May 16, 2022).

[19] "Stack Overflow - Where Developers Learn, Share, & Build Careers", Accessed: May 16, 2022. [Online]. Available: https://stackoverflow.com/

[20] M. J. Page *et al.*, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *The BMJ*, vol. 372. 2021. doi: 10.1136/bmj.n71.

[21] "PubMed." https://pubmed.ncbi.nlm.nih.gov/ (accessed May 06, 2022).

[22] "ScienceDirect.com | Science, health and medical journals, full text articles and books." https://www.sciencedirect.com/ (accessed May 06, 2022).

[23] "ACM Digital Library." https://dl.acm.org/ (accessed May 06, 2022).

[24] "Home - Springer." https://link.springer.com/ (accessed May 06, 2022).

[25] "Google Scholar." https://scholar.google.com/ (accessed May 06, 2022).

[26] E. Xie, K. S. Reddy, and J. Liang, "Country-specific determinants of cross-border mergers and acquisitions: A comprehensive review and future research directions," *Journal of World Business*, vol. 52, no. 2. 2017. doi: 10.1016/j.jwb.2016.12.005.

[27] L. Bornmann and H. D. Daniel, "What do citation counts measure? A review of studies on citing behavior," *J. Doc.*, vol. 64, no. 1, 2008, doi: 10.1108/00220410810844150.

[28] R. N. Broadus, "An investigation of the validity of bibliographic citations," *J. Am. Soc. Inf. Sci.*, vol. 34, no. 2, 1983, doi: 10.1002/asi.4630340206.

[29] P. Eichorn and A. Yankauer, "Do authors check their references? A survey of accuracy of references in three public health journals," *Am. J. Public Health*, vol. 77, no. 8, 1987, doi: 10.2105/AJPH.77.8.1011.

[30] J. T. Evans, H. I. Nadjari, and S. A. Burchell, "Quotational and Reference Accuracy in Surgical Journals: A Continuing Peer Review Problem," *JAMA J. Am. Med. Assoc.*, vol.

263, no. 10, 1990, doi: 10.1001/jama.1990.03440100059009.

[31] M. C. Teixeira *et al.*, "Incorrect citations give unfair credit to review authors in ecology journals," *PLoS ONE*, vol. 8, no. 12. 2013. doi: 10.1371/journal.pone.0081871.

[32] "ISI Impact Factor (IIF)." https://isi-impactfactor.com/ (accessed Apr. 21, 2022).

[33] L. Bornmann and H. D. Daniel, "What do we know about the h index?," *J. Am. Soc. Inf. Sci. Technol.*, vol. 58, no. 9, 2007, doi: 10.1002/asi.20609.

[34] J. A. Kunze *et al.*, "Practices, Trends, and Recommendations in Technical Appendix Usage for Selected Data-Intensive Disciplines," 2011.

[35] K. I. Berns, E. C. Bond, and F. J. Manning, *Resource Sharing in Biomedical Research*. 1996. doi: 10.17226/5429.

[36] T. R. Cech *et al.*, "Sharing publication-related data and materials: Responsibilities of authorship in the life sciences," *Plant Physiology*, vol. 132, no. 1. 2003. doi: 10.1104/pp.900068.

[37] R. J. Shavelson and L. Towne, *Scientific research in education. Committee on scientific principles for education research*, vol. 16. 2002.

[38] P. F. Uhlir and P. Schröder, "Open Data for Global Science," *Data Sci. J.*, vol. 6, 2007, doi: 10.2481/dsj.6.od36.

[39] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsiouliklis, "Two supervised learning approaches for name disambiguation in author citations," 2004. doi: 10.1145/996350.996419.

[40] Y. Qian, Q. Zheng, T. Sakai, J. Ye, and J. Liu, "Dynamic author name disambiguation for growing digital libraries," *Inf. Retr. Boston.*, vol. 18, no. 5, 2015, doi: 10.1007/s10791-015-9261-3.

[41] A. F. Santana, M. A. Gonçalves, A. H. F. Laender, and A. A. Ferreira, "Incremental author name disambiguation by exploiting domain-specific heuristics," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 4, 2017, doi: 10.1002/asi.23726.

[42] X. Wang, J. Tang, H. Cheng, and P. S. Yu, "ADANA: Active name disambiguation," 2011. doi: 10.1109/ICDM.2011.19.

[43] I. S. Kang, P. Kim, S. Lee, H. Jung, and B. J. You, "Construction of a large-scale test set for author disambiguation," *Inf. Process. Manag.*, vol. 47, no. 3, 2011, doi: 10.1016/j.ipm.2010.10.001.

[44] M. Song, E. H. J. Kim, and H. J. Kim, “Exploring author name disambiguation on PubMed-scale,” *J. Informetr.*, vol. 9, no. 4, 2015, doi: 10.1016/j.joi.2015.08.004.

[45] R. G. Cota, A. A. Ferreira, C. Nascimento, M. A. Gonçalves, and A. H. F. Laender, “An unsupervised heuristic-based hierarchical method for name disambiguation in bibliographic citations,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 9, 2010, doi: 10.1002/asi.21363.

[46] V. I. Torvik and N. R. Smalheiser, “Author name disambiguation in MEDLINE,” *ACM Trans. Knowl. Discov. Data*, vol. 3, no. 3, 2009, doi: 10.1145/1552303.1552304.

[47] M. Levin, S. Krawczyk, S. Bethard, and D. Jurafsky, “Citation-based bootstrapping for large-scale author disambiguation,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 5, 2012, doi: 10.1002/asi.22621.

[48] J. Kim, J. Kim, and J. Owen-Smith, “Generating automatically labeled data for author name disambiguation: an iterative clustering method,” *Scientometrics*, vol. 118, no. 1, 2019, doi: 10.1007/s11192-018-2968-3.

[49] H. Kawashima and H. Tomizawa, “Accuracy evaluation of scopus author ID based on the largest funding database in Japan,” *Scientometrics*, vol. 103, no. 3, 2015, doi: 10.1007/s11192-015-1580-z.

[50] “Scopus preview - Scopus - Welcome to Scopus.” https://www.scopus.com/home.uri (accessed Apr. 21, 2022).

[51] “KAKEN - Research Projects.” https://kaken.nii.ac.jp/ (accessed Apr. 21, 2022).

[52] C. A. D’Angelo, C. Giuffrida, and G. Abramo, “A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 62, no. 2, 2011, doi: 10.1002/asi.21460.

[53] L. Reijnhoudt, R. Costas, E. Noyons, K. Börner, and A. Scharnhorst, “‘Seed + expand’: a general methodology for detecting publication oeuvres of individual researchers,” *Scientometrics*, vol. 101, no. 2, 2014, doi: 10.1007/s11192-014-1256-0.

[54] “RePORT - RePORTER.” https://reporter.nih.gov/exporter (accessed Apr. 21, 2022).

[55] “Recipients - Highly Cited | Researcher Recognition”, Accessed: Apr. 21, 2022. [Online]. Available: https://recognition.webofscience.com/awards/highly-cited/2021/

[56] K. Kim, A. Sefid, B. A. Weinberg, and C. L. Giles, “A Web Service for Author Name Disambiguation in Scholarly Databases,” 2018. doi: 10.1109/ICWS.2018.00041.

[57] M. J. Lerchenmueller and O. Sorenson, “Author disambiguation in PubMed: Evidence on the precision and recall of authority among NIH-funded scientists,” *PLoS One*, vol. 11, no. 7, 2016, doi: 10.1371/journal.pone.0158731.

[58] W. Liu *et al.*, “Author name disambiguation for PubMed,” *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 4, 2014, doi: 10.1002/asi.23063.

[59] “ORCID.” https://orcid.org/ (accessed May 03, 2022).

[60] J. Kim, “Evaluating author name disambiguation for digital libraries: a case of DBLP,” *Scientometrics*, vol. 116, no. 3, 2018, doi: 10.1007/s11192-018-2824-5.

[61] J. Kim, “Scale-free collaboration networks: An author name disambiguation perspective,” *J. Assoc. Inf. Sci. Technol.*, vol. 70, no. 7, 2019, doi: 10.1002/asi.24158.

[62] L. Francis, “More than Just a Number: the ORCID Unique Identifier for Academics and Their Research Activities,” *Ed. Bull.*, vol. 9, no. 2, pp. 42–44, Jul. 2013, doi: 10.1080/17521742.2013.870719.

[63] M. Mallery, “Scholarly Identification Systems in a Global Market: The ORCID Solution,” *Int. Inf. Libr. Rev.*, vol. 48, no. 4, 2016, doi: 10.1080/10572317.2016.1243962.

[64] W. M. J. Thomas, B. Chen, and G. Clement, “ORCID Identifiers: Planned and Potential Uses by Associations, Publishers, and Librarians,” *Ser. Libr.*, vol. 68, no. 1–4, 2015, doi: 10.1080/0361526X.2015.1017713.

[65] M. D. Wilkinson *et al.*, “Comment: The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, 2016, doi: 10.1038/sdata.2016.18.

[66] “FORCE11.” https://force11.org/ (accessed May 04, 2022).

[67] P. E. Bourne *et al.*, “Improving The Future of Research Communications and e-scholarship,” *Dagstuhl Manifestos*, vol. 1, no. 1, 2012.

[68] A. M. Smith, D. S. Katz, and K. E. Niemeyer, “Software citation principles,” *PeerJ Comput. Sci.*, vol. 2016, no. 9, 2016, doi: 10.7717/peerj-cs.86.

[69] M. Martone, “Joint Declaration of Data Citation Principles,” *Data Cit. Synth. Gr. Jt. Declar. Data Cit. Princ.*, 2014.

[70] D. S. Katz *et al.*, “Software vs. data in the context of citation,” *PeerJ Prepr.*, vol. 4, 2016.

[71] N. P. Chue Hong *et al.*, “Software Citation Checklist for Authors,” Oct. 2019, doi: 10.5281/ZENODO.3479199.

[72] N. P. Chue Hong *et al.*, “Software Citation Checklist for Developers,” Oct. 2019, doi:

10.5281/ZENODO.3482769.

[73] D. S. Katz *et al.*, “Recognizing the value of software: A software citation guide,” *F1000Research*, vol. 9, 2021, doi: 10.12688/f1000research.26932.2.

[74] G. P. Patrinos *et al.*, “Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain,” *Hum. Mutat.*, vol. 33, no. 11, 2012, doi: 10.1002/humu.22144.

[75] B. Mons *et al.*, “The value of data,” *Nat. Genet.*, vol. 43, no. 4, pp. 281–283, Apr. 2011, doi: 10.1038/ng0411-281.

[76] M. Parsons and Y. M. Socha, “Correction to: Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data (Data Science Journal, 2013),” *Data Science Journal*, vol. 20, no. 1. 2021. doi: 10.5334/dsj-2021-021.

[77] B. Giardine *et al.*, “Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach,” *Nat. Genet.*, vol. 43, no. 4, 2011, doi: 10.1038/ng.785.

[78] E. Fogarty, “Credit where credit is due,” *Science*, vol. 370, no. 6520. 2020. doi: 10.1126/science.370.6520.1130.

[79] “OpenID Foundation Website.” https://openid.net/ (accessed May 05, 2022).

[80] “VIAF.” http://viaf.org/ (accessed May 05, 2022).

[81] “ResearchGate | Find and share research.” https://www.researchgate.net/ (accessed May 05, 2022).

[82] “ISNI | Home Page.” https://isni.org/ (accessed May 05, 2022).

[83] “ISO - ISO 27729:2012 - Information and documentation — International standard name identifier (ISNI).” https://www.iso.org/standard/44292.html (accessed May 05, 2022).

[84] L. L. Haak, M. Fenner, L. Paglione, E. Pentz, and H. Ratner, “ORCID: A system to uniquely identify researchers,” *Learn. Publ.*, vol. 25, no. 4, 2012, doi: 10.1087/20120404.

[85] “You are Crossref - Crossref.” https://www.crossref.org/ (accessed May 05, 2022).

[86] “Home | Wellcome.” https://wellcome.org/ (accessed May 05, 2022).

[87] D. Butler, “Scientists: Your number is up,” *Nature*, vol. 485, no. 7400. 2012. doi: 10.1038/485564a.

[88] “ISO - ISO 9241-110:2020 - Ergonomics of human-system interaction — Part 110: Interaction principles.” https://www.iso.org/standard/75258.html (accessed May 08,

2022).

[89] H. Chen, T. Yu, and J. Y. Chen, “Semantic web meets integrative biology: A survey,” *Briefings in Bioinformatics*, vol. 14, no. 1. 2013. doi: 10.1093/bib/bbs014.

[90] T. P. Sneddon, P. Li, and S. C. Edmunds, “GigaDB: Announcing the GigaScience database,” *GigaScience*, vol. 1, no. 1. 2012. doi: 10.1186/2047-217X-1-11.

[91] S. Ghosh, Y. Matsuoka, Y. Asai, K. Y. Hsin, and H. Kitano, “Software for systems biology: From tools to integrated platforms,” *Nature Reviews Genetics*, vol. 12, no. 12. 2011. doi: 10.1038/nrg3096.

[92] S. A. Beaulah, M. A. Correll, R. E. J. Munro, and J. G. Sheldon, “Addressing informatics challenges in Translational Research with workflow technology,” *Drug Discovery Today*, vol. 13, no. 17–18. 2008. doi: 10.1016/j.drudis.2008.06.005.

[93] K. Wolstencroft *et al.*, “The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud.,” *Nucleic Acids Res.*, vol. 41, no. Web Server issue, 2013, doi: 10.1093/nar/gkt328.

[94] C. A. Goble *et al.*, “myExperiment: A repository and social network for the sharing of bioinformatics workflows,” *Nucleic Acids Res.*, vol. 38, no. SUPPL. 2, 2010, doi: 10.1093/nar/gkq429.

[95] P. Mates, E. Santos, J. Freire, and C. T. Silva, “CrowdLabs: Social analysis and visualization for the sciences,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011, vol. 6809 LNCS. doi: 10.1007/978-3-642-22351-8_38.

[96] J. Zhao *et al.*, “Why workflows break - Understanding and combating decay in Taverna workflows,” 2012. doi: 10.1109/eScience.2012.6404482.

[97] G. K. Sandve, A. Nekrutenko, J. Taylor, and E. Hovig, “Ten Simple Rules for Reproducible Computational Research,” *PLoS Computational Biology*, vol. 9, no. 10. 2013. doi: 10.1371/journal.pcbi.1003285.

[98] M. Baker, “1,500 scientists lift the lid on reproducibility,” *Nature*, vol. 533, no. 7604, 2016, doi: 10.1038/533452a.

[99] G. J. Lithgow, M. Driscoll, and P. Phillips, “A long journey to reproducible results,” *Nature*, vol. 548, 2017.

[100] D. B. Searls, “The roots of bioinformatics,” *PLoS Comput. Biol.*, vol. 6, no. 6, 2010, doi:

10.1371/journal.pcbi.1000809.

[101] S. Kanwal, F. Z. Khan, A. Lonie, and R. O. Sinnott, "Investigating reproducibility and tracking provenance - A genomic workflow case study," *BMC Bioinformatics*, vol. 18, no. 1, 2017, doi: 10.1186/s12859-017-1747-0.

[102] R. C. Gentleman *et al.*, "Bioconductor: open software development for computational biology and bioinformatics.," *Genome Biol.*, vol. 5, no. 10, 2004.

[103] A. R. Colombo, T. J. Triche Jr, and G. Ramsingh, "Arkas: Rapid reproducible RNAseq analysis," *F1000Research*, vol. 6, 2017, doi: 10.12688/f1000research.11355.2.

[104] C. Van Neste *et al.*, "Forensic massively parallel sequencing data analysis tool: Implementation of MyFLq as a standalone web- and Illumina BaseSpace®-application," *Forensic Sci. Int. Genet.*, vol. 15, 2015, doi: 10.1016/j.fsigen.2014.10.006.

[105] W. Digan *et al.*, "An architecture for genomics analysis in a clinical setting using Galaxy and Docker," *Gigascience*, vol. 6, no. 11, 2017, doi: 10.1093/gigascience/gix099.

[106] "BaseSpace Developers." https://developer.basespace.illumina.com/ (accessed May 08, 2022).

[107] E. S. Dove *et al.*, "Genomic cloud computing: Legal and ethical points to consider," *Eur. J. Hum. Genet.*, vol. 23, no. 10, 2015, doi: 10.1038/ejhg.2014.196.

[108] A. Osz, L. S. Pongor, D. Szirmai, and B. Gyorffy, "A snapshot of 3649 Web-based services published between 1994 and 2017 shows a decrease in availability after 2 years," *Brief. Bioinform.*, vol. 20, no. 3, 2017, doi: 10.1093/bib/bbx159.

[109] J. D. Wren, C. Georgescu, C. B. Giles, and J. Hennessey, "Use it or lose it: Citations predict the continued online availability of published bioinformatics resources," *Nucleic Acids Research*, vol. 45, no. 7. 2017. doi: 10.1093/nar/gkx182.

[110] J. D. Wren, "URL decay in MEDLINE - A 4-year follow-up study," *Bioinformatics*, vol. 24, no. 11, 2008, doi: 10.1093/bioinformatics/btn127.

[111] S. Veretnik, J. L. Fink, and P. E. Bourne, "Computational biology resources lack persistence and usability," *PLoS Computational Biology*, vol. 4, no. 7. 2008. doi: 10.1371/journal.pcbi.1000136.

[112] T. Thireou, G. Spyrou, and V. Atlamazoglou, "A Survey of the Availability of Primary Bioinformatics Web Resources," *Genomics, Proteomics Bioinforma.*, vol. 5, no. 1, 2007, doi: 10.1016/S1672-0229(07)60017-5.

[113] P. H. Russell, R. L. Johnson, S. Ananthan, B. Harnke, and N. E. Carlson, “A large-scale analysis of bioinformatics code on GitHub,” *PLoS One*, vol. 13, no. 10, 2018, doi: 10.1371/journal.pone.0205898.

[114] L. D. Parnell *et al.*, “BioStar: An online question & answer resource for the bioinformatics community,” *PLoS Comput. Biol.*, vol. 7, no. 10, 2011, doi: 10.1371/journal.pcbi.1002216.

[115] S. Mangul *et al.*, “Challenges and recommendations to improve the installability and archival stability of omics computational tools,” *PLoS Biol.*, vol. 17, no. 6, 2019, doi: 10.1371/journal.pbio.3000333.

[116] F. Kern, T. Fehlmann, and A. Keller, “On the lifetime of bioinformatics web services,” *Nucleic Acids Research*, vol. 48, no. 22. 2020. doi: 10.1093/nar/gkaa1125.

[117] S. J. Schultheiss, M. C. Münch, G. D. Andreeva, and G. Rätsch, “Persistence and availability of web services in computational biology,” *PLoS ONE*, vol. 6, no. 9. 2011. doi: 10.1371/journal.pone.0024914.

[118] J. Ison *et al.*, “Community curation of bioinformatics software and data resources,” *Brief. Bioinform.*, vol. 21, no. 5, 2020, doi: 10.1093/bib/bbz075.

[119] A. Goderis *et al.*, “Benchmarking workflow discovery: A case study from bioinformatics,” in *Concurrency and Computation: Practice and Experience*, 2009, vol. 21, no. 16. doi: 10.1002/cpe.1447.

[120] R. Stevens, C. Goble, P. Baker, and A. Brass, “A classification of tasks in bioinformatics,” *Bioinformatics*, vol. 17, no. 2, 2001, doi: 10.1093/bioinformatics/17.2.180.

[121] Z. Lacroix and H. Ménager, “Evaluating workflow management systems for bioinformatics,” Nov. 2005.

[122] Y. Gil *et al.*, “Examining the challenges of scientific workflows,” *Computer (Long. Beach. Calif).*, vol. 40, no. 12, 2007, doi: 10.1109/MC.2007.421.

[123] D. De Roure *et al.*, “Towards open science: The myExperiment approach,” *Concurr. Comput. Pract. Exp.*, vol. 22, no. 17, 2010, doi: 10.1002/cpe.1601.

[124] A. Whyte and G. Pryor, “Open Science in Practice: Researcher Perspectives and Participation,” *Int. J. Digit. Curation*, vol. 6, no. 1, 2011, doi: 10.2218/ijdc.v6i1.182.

[125] N. Levin, S. Leonelli, D. Weckowska, D. Castle, and J. Dupré, “How Do Scientists Define Openness? Exploring the Relationship Between Open Science Policies and Research

Practice," *Bull. Sci. Technol. Soc.*, vol. 36, no. 2, 2016, doi: 10.1177/0270467616668760.

[126] D. Hicks, P. Wouters, L. Waltman, S. De Rijcke, and I. Rafols, "Bibliometrics: The Leiden Manifesto for research metrics," *Nature*, vol. 520, no. 7548. 2015. doi: 10.1038/520429a.

[127] J. Wilsdon, *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management*. 2017. doi: 10.4135/9781473978782.

[128] M. R. Munafò *et al.*, "A manifesto for reproducible science," *Nature Human Behaviour*, vol. 1, no. 1. 2017. doi: 10.1038/s41562-016-0021.

[129] S. Collini, "Who are the spongers now ?," *London Rev. Books*, no. November, 2015.

[130] B. Martin, "What is Happening to Our Universities?," *SSRN Electron. J.*, 2016, doi: 10.2139/ssrn.2745139.

[131] R. Benedictus, F. Miedema, and M. W. J. Ferguson, "Fewer numbers, better science," *Nature*, vol. 538, no. 7626. 2016. doi: 10.1038/538453a.

[132] D. Sarewitz, "The pressure to publish pushes down quality," *Nature*, vol. 533. 2016. doi: 10.1038/533147a.

[133] J. W. Raff, "The San Francisco declaration on research assessment," *Biology Open*, vol. 2, no. 6. 2013. doi: 10.1242/bio.20135330.

[134] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 46, 2005, doi: 10.1073/pnas.0507655102.

[135] E. Garfield, "Citation analysis as a tool in journal evaluation," *Science (80-.).*, vol. 178, no. 4060, 1972, doi: 10.1126/science.178.4060.471.

[136] B. González-Pereira, V. P. Guerrero-Bote, and F. Moya-Anegón, "A new approach to the metric of journals scientific prestige: The SJR indicator," *J. Informetr.*, vol. 4, no. 3, 2010, doi: 10.1016/j.joi.2010.03.002.

[137] L. Waltman and N. J. van Eck, "Source normalized indicators of citation impact: An overview of different approaches and an empirical comparison," *Scientometrics*, vol. 96, no. 3, 2013, doi: 10.1007/s11192-012-0913-4.

[138] C. T. Bergstrom, J. D. West, and M. A. Wiseman, "The Eigenfactor^{TM} metrics," *Journal of Neuroscience*, vol. 28, no. 45. 2008. doi: 10.1523/JNEUROSCI.0003-08.2008.

[139] H. F. Moed, "Measuring contextual citation impact of scientific journals," *J. Informetr.*, vol. 4, no. 3, 2010, doi: 10.1016/j.joi.2010.01.002.

[140] C. James, L. Colledge, W. Meester, N. Azoulay, and A. Plume, “CiteScore metrics: Creating journal metrics from the Scopus citation index,” *Learn. Publ.*, vol. 32, no. 4, 2019, doi: 10.1002/leap.1246.

[141] B. M. Webster, “Principles to guide reliable and ethical research evaluation using metric-based indicators of impact,” *Perform. Meas. Metrics*, vol. 18, no. 1, 2017, doi: 10.1108/PMM-06-2016-0025.

[142] L. Zhang, R. Rousseau, and G. Sivertsen, “Science deserves to be judged by its contents, not by its wrapping: Revisiting Seglen’s work on journal impact and research evaluation,” *PLoS One*, vol. 12, no. 3, 2017, doi: 10.1371/journal.pone.0174205.

[143] F. Verleysen and R. Rousseau, “How the existence of a regional bibliographic information system can help evaluators to conform to the principles of the Leiden Manifesto,” *J. Educ. Media Libr. Sci.*, vol. 54, no. 1, 2017.

[144] Di. Chavarro, I. Ràfols, and P. Tang, “To what extent is inclusion in the Web of Science an indicator of journal ‘quality’?,” *Res. Eval.*, vol. 27, no. 2, 2018, doi: 10.1093/reseval/rvy001.

[145] H. H. Bi, “Four problems of the h-index for assessing the research productivity and impact of individual authors,” *Scientometrics*, 2022, doi: 10.1007/s11192-022-04323-8.

[146] L. Waltman and N. J. Van Eck, “The inconsistency of the h-index,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 2, 2012, doi: 10.1002/asi.21678.

[147] J. K. Vanclay, “On the robustness of the h-index,” *Journal of the American Society for Information Science and Technology*, vol. 58, no. 10. 2007. doi: 10.1002/asi.20616.

[148] H. F. Moed, “Citation analysis in research evaluation,” in *Proceedings of ISSI 2005: 10th International Conference of the International Society for Scientometrics and Informetrics*, 2005, vol. 2. doi: 10.5117/2006.019.002.007.

[149] H. Cousijn, P. Feeney, D. Lowenberg, E. Presani, and N. Simons, “Bringing citations and usage metrics together to make data count,” *Data Sci. J.*, vol. 18, no. 1, 2019, doi: 10.5334/dsj-2019-009.

[150] I. Rowlands and D. Nicholas, “The missing link: Journal usage metrics,” *Aslib Proc. New Inf. Perspect.*, vol. 59, no. 3, 2007, doi: 10.1108/00012530710752025.

[151] P. Huntington, D. Nicholas, and H. R. Jamali, “Website usage metrics: A re-assessment of session data,” *Inf. Process. Manag.*, vol. 44, no. 1, 2008, doi: 10.1016/j.ipm.2007.03.003.

[152] M. J. Kurtz and J. Bollen, “Usage bibliometrics,” *Annual Review of Information Science and Technology*, vol. 44. 2010. doi: 10.1002/aris.2010.1440440108.

[153] H. D. White, S. K. Boell, H. Yu, M. Davis, C. S. Wilson, and F. T. H. Cole, “Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 60, no. 6, 2009, doi: 10.1002/asi.21045.

[154] “Search | Mendeley.” https://www.mendeley.com/search/ (accessed May 10, 2022).

[155] I. Tahamtan and L. Bornmann, “Altmetrics and societal impact measurements: Match or mismatch? a literature review,” *Profesional de la Informacion*, vol. 29, no. 1. 2020. doi: 10.3145/epi.2020.ene.02.

[156] C. Neylon and S. Wu, “Article-level metrics and the evolution of scientific impact,” *PLoS Biology*, vol. 7, no. 11. 2009. doi: 10.1371/journal.pbio.1000242.

[157] J. Priem and B. M. Hemminger, “Scientometrics 2.0: Toward new metrics of scholarly impact on the social Web,” *First Monday*, vol. 15, no. 7, 2010, doi: 10.5210/fm.v15i7.2874.

[158] J. Priem, D. Taraborelli, P. Groth, and C. Neylon, “Altmetrics: a manifesto,” *October*. 2010.

[159] “Twitter. It’s what’s happening / Twitter.” https://twitter.com/ (accessed May 10, 2022).

[160] S. Deterding, R. Khaled, L. Nacke, and D. Dixon, “Gamification: toward a definition,” *Chi 2011*, 2011.

[161] J. Zhou, S. Wang, C. P. Bezemer, and A. E. Hassan, “Bounties on technical Q&A sites: a case study of Stack Overflow bounties,” *Empir. Softw. Eng.*, vol. 25, no. 1, 2020, doi: 10.1007/s10664-019-09744-3.

[162] K. Seaborn and D. I. Fels, “Gamification in theory and action: A survey,” *Int. J. Hum. Comput. Stud.*, vol. 74, 2015, doi: 10.1016/j.ijhcs.2014.09.006.

[163] “Bitbucket | Git solution for teams using Jira.” https://bitbucket.org/product/ (accessed May 16, 2022).

[164] R. Hebig, T. H. Quang, M. R. V. Chaudron, G. Robles, and M. A. Fernandez, “The quest for open source projects that use UML: Mining GitHub,” 2016. doi: 10.1145/2976767.2976778.

[165] R. Saxena and N. Pedanekar, “I Know What You Coded Last Summer,” 2017. doi: 10.1145/3022198.3026354.

[166] Y. Hu, S. Wang, Y. Ren, and K. K. R. Choo, “User influence analysis for Github developer social networks,” *Expert Syst. Appl.*, vol. 108, 2018, doi: 10.1016/j.eswa.2018.05.002.

[167] R. Abdalkareem, E. Shihab, and J. Rilling, “What Do Developers Use the Crowd For? A Study Using Stack Overflow,” *IEEE Softw.*, vol. 34, no. 2, 2017, doi: 10.1109/MS.2017.31.

[168] D. Van Dijk, M. Tsagkias, and M. De Rijke, “Early detection of topical expertise in community question answering,” 2015. doi: 10.1145/2766462.2767840.

[169] L. Fernando Capretz, “Bringing the human factor to software engineering,” *IEEE Software*, vol. 31, no. 2. 2014. doi: 10.1109/MS.2014.30.

[170] H. Cavusoglu, Z. Li, and K. W. Huang, “Can gamification motivate voluntary contributions? The case of StackOverflow Q&A community,” in *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*, 2015, vol. 2015-January. doi: 10.1145/2685553.2698999.

[171] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, “Discovering value from community activity on focused question answering sites: A case study of stack overflow,” 2012. doi: 10.1145/2339530.2339665.

[172] A. Begel, J. Bosch, and M. A. Storey, “Social networking meets software development: Perspectives from git hub, MSDN, stack exchange, and top coder,” *IEEE Softw.*, vol. 30, no. 1, 2013, doi: 10.1109/MS.2013.13.

[173] V. Singh, M. B. Twidale, and D. M. Nichols, “Users of open source software? How do they get help?,” 2009. doi: 10.1109/HICSS.2009.489.

[174] M. A. Storey, C. Treude, A. Van Deursen, and L. Te Cheng, “The impact of social media on software engineering practices and tools,” 2010. doi: 10.1145/1882362.1882435.

[175] B. Vasilescu, A. Serebrenik, P. Devanbu, and V. Filkov, “How social Q&Amp;A sites are changing knowledge sharing in open source software communities,” 2014. doi: 10.1145/2531602.2531659.

[176] C. Parnin, C. Treude, L. Grammel, and M.-A. Storey, “Crowd documentation: Exploring the coverage and the dynamics of API discussions on Stack Overflow,” *Georg. Tech Tech. Rep.*, 2012.

[177] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, “Design lessons

from the fastest Q&A site in the west," 2011. doi: 10.1145/1978942.1979366.

[178] S. Deterding, K. O'Hara, M. Sicart, D. Dixon, and L. Nacke, "Gamification: Using game design elements in non-gaming contexts," 2011. doi: 10.1145/1979742.1979575.

[179] A. Capiluppi, A. Serebrenik, and L. Singer, "Assessing technical candidates on the social web," *IEEE Softw.*, vol. 30, no. 1, 2013, doi: 10.1109/MS.2012.169.

[180] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, "Social coding in GitHub: Transparency and collaboration in an open software repository," 2012. doi: 10.1145/2145204.2145396.

[181] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, "Leveraging transparency," *IEEE Softw.*, vol. 30, no. 1, 2013, doi: 10.1109/MS.2012.172.

[182] J. Antin and E. F. Churchill, "Badges in social media: A social psychological perspective," *Chi 2011*, 2011.

[183] Y. Chen, F. M. Harper, J. Konstan, and S. X. Li, "Social comparisons and contributions to online communities: A field experiment on MovieLens," *Am. Econ. Rev.*, vol. 100, no. 4, 2010, doi: 10.1257/aer.100.4.1358.

[184] J. Frith, "Turning life into a game: Foursquare, gamification, and personal mobility," *Mob. Media Commun.*, vol. 1, no. 2, 2013, doi: 10.1177/2050157912474811.

[185] S. Jain, Y. Chen, and D. C. Parkes, "Designing incentives for online question-and-answer forums," *Games Econ. Behav.*, vol. 86, 2014, doi: 10.1016/j.geb.2012.11.003.

[186] O. Nov, O. Arazy, and D. Anderson, "Scientists@Home: What drives the quantity and quality of online citizen science participation?," *PLoS One*, vol. 9, no. 4, 2014, doi: 10.1371/journal.pone.0090375.

[187] N. Immorlica, G. Stoddard, and V. Syrgkanis, "Social status and badge design," 2015. doi: 10.1145/2736277.2741664.

[188] J. A. Roberts, I. H. Hann, and S. A. Slaughter, "Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the Apache projects," *Manage. Sci.*, vol. 52, no. 7, 2006, doi: 10.1287/mnsc.1060.0554.

[189] G. Wang, K. Gill, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Wisdom in the social crowd: An analysis of Quora," 2013.

[190] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Steering user behavior with badges," 2013. doi: 10.1145/2488388.2488398.

[191] T. Mutter and D. Kundisch, "Behavioral mechanisms prompted by badges: The goal-

gradient hypothesis," 2014.

[192] T. Kusmierczyk and M. Gomez-Rodriguez, "On the causal effect of badges," 2018. doi: 10.1145/3178876.3186147.

[193] "Scientific Data." https://www.nature.com/sdata/ (accessed Sep. 04, 2022).

[194] "JMIR Data." https://data.jmir.org/ (accessed Sep. 04, 2022).

[195] "Galaxy." https://usegalaxy.org/ (accessed Jul. 29, 2022).

[196] V. Jalili *et al.*, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2020 update," *Nucleic Acids Res.*, vol. 48, no. W1, 2021, doi: 10.1093/NAR/GKAA434.

[197] E. Afgan *et al.*, "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update," *Nucleic Acids Res.*, vol. 46, no. W1, 2018, doi: 10.1093/nar/gky379.

[198] "bio.tools · Bioinformatics Tools and Services Discovery Portal." https://bio.tools/ (accessed Jul. 29, 2022).

[199] "ELIXIR | A distributed infrastructure for life-science information." https://elixir-europe.org/ (accessed Jul. 29, 2022).

[200] J. Ison *et al.*, "Tools and data services registry: A community effort to document bioinformatics resources," *Nucleic Acids Res.*, vol. 44, no. D1, 2016, doi: 10.1093/nar/gkv1116.

[201] J. Ison *et al.*, "BiotoolsSchema: A formalized schema for bioinformatics software description," *Gigascience*, vol. 10, no. 1, 2021, doi: 10.1093/gigascience/giaa157.

[202] J. Ison *et al.*, "EDAM: An ontology of bioinformatics operations, types of data and identifiers, topics and formats," *Bioinformatics*, vol. 29, no. 10, 2013, doi: 10.1093/bioinformatics/btt113.

[203] "Datasets2Tools | Bioinformatics Index." https://maayanlab.cloud/datasets2tools (accessed Jul. 30, 2022).

[204] K. M. Jagodnik *et al.*, "Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop," in *Journal of Biomedical Informatics*, 2017, vol. 71. doi: 10.1016/j.jbi.2017.05.006.

[205] "Bioinformatics Answers." https://www.biostars.org/ (accessed Jul. 30, 2022).

[206] M. D. Wilkinson *et al.*, "Interoperability and FAIRness through a novel combination of

Web technologies," *PeerJ Comput. Sci.*, vol. 2017, no. 4, 2017, doi: 10.7717/peerj-cs.110.

[207] S. Henninger, "Using Iterative Refinement to Find Reusable Software," *IEEE Softw.*, vol. 11, no. 5, 1994, doi: 10.1109/52.311059.

[208] E. D. Foster and A. Deardorff, "Open Science Framework (OSF)," *J. Med. Libr. Assoc.*, vol. 105, no. 2, 2017, doi: 10.5195/jmla.2017.88.

[209] B. Fox, "Bash software," *GNU Project*, 2020.

[210] A. Kanterakis, N. Karacapilidis, L. Koumakis, and G. Potamias, "On the development of an open and collaborative bioinformatics reSearch environment," in *Procedia Computer Science*, 2018, vol. 126. doi: 10.1016/j.procS.2018.08.043.

[211] P. DI Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, and C. Notredame, "Nextflow enables reproducible computational workflows," *Nature Biotechnology*, vol. 35, no. 4. 2017. doi: 10.1038/nbt.3820.

[212] "Angular." https://angular.io/ (accessed Aug. 14, 2022).

[213] "The web framework for perfectionists with deadlines | Django." https://www.djangoproject.com/ (accessed Aug. 17, 2022).

[214] "PostgreSQL: The world's most advanced open source database." https://www.postgresql.org/ (accessed Aug. 30, 2022).

[215] "Documentation - Materialize." https://materializecss.com/ (accessed Aug. 17, 2022).

[216] "Design - Material Design." https://material.io/design (accessed Aug. 17, 2022).

[217] G. Abramo, C. A. D'Angelo, and M. Solazzi, "The relationship between scientists' research performance and the degree of internationalization of their research," *Scientometrics*, vol. 86, no. 3, 2011, doi: 10.1007/s11192-010-0284-7.

[218] R. Costas, T. N. Van Leeuwen, and M. Bordons, "A bibliometric classificatory approach for the study and assessment of research performance at the individual level: The effects of age on productivity and impact," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 8, 2010, doi: 10.1002/asi.21348.

[219] L. Wildgaard, J. W. Schneider, and B. Larsen, "A review of the characteristics of 108 author-level bibliometric indicators," *Scientometrics*, vol. 101, no. 1, 2014, doi: 10.1007/s11192-014-1423-3.

[220] P. Mongeon, N. Robinson-Garcia, W. Jeng, and R. Costas, "Incorporating data sharing to the reward system of science: Linking DataCite records to authors in the Web of Science,"

Aslib J. Inf. Manag., vol. 69, no. 5, 2017, doi: 10.1108/AJIM-01-2017-0024.

[221] E. Caron and N. J. van Eck, "Large scale author name disambiguation using rule-based scoring and clustering," *Proc. Sci. Technol. Indic. Conf. 2014 Leiden*, no. September, 2014.

[222] A. Tekles and L. Bornmann, "Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches," in *17th International Conference on Scientometrics and Informetrics, ISSI 2019 - Proceedings*, 2019, vol. 2.

[223] J. Brown, T. Demeranville, and A. Meadows, "Open access in context: Connecting authors, publications and workflows using ORCID identifiers," *Publications*, vol. 4, no. 4, 2016, doi: 10.3390/publications4040030.

[224] A. Meadows, "Everything you ever wanted to know about ORCID ⋯ but were afraid to ask," *Coll. Res. Libr. News*, vol. 77, no. 1, 2016, doi: 10.5860/crln.77.1.9428.

[225] "Clarivate." https://access.clarivate.com/login?app=wos&alternative=true&shibShireURL=https:%2F%2Fwww.webofknowledge.com%2F%3Fauth%3DShibboleth&shibReturnURL=https:%2F%2Fwww.webofknowledge.com%2F%3FauthCode%3Dnull%26app%3Dwos%26referrer%3Dapp%253Dwos%2526authCode%253Dnull%2526locale%253Den-US%26locale%3Den-US&referrer=app%3Dwos%26authCode%3Dnull%26locale%3Den-US&roaming=true (accessed Aug. 04, 2022).

[226] E. R. Sprague, "ORCID," *J. Med. Libr. Assoc.*, vol. 105, no. 2, Apr. 2017, doi: 10.5195/jmla.2017.89.

[227] K. G. Akers, A. Sarkozy, W. Wu, and A. Slyman, "ORCID Author Identifiers: A Primer for Librarians," *Med. Ref. Serv. Q.*, vol. 35, no. 2, 2016, doi: 10.1080/02763869.2016.1152139.

[228] M. Foley and D. Kochalko, "Open Researcher and Contributor Identification (ORCID)," 2012. doi: 10.5703/1288284314850.

[229] "Public API - ORCID." https://info.orcid.org/documentation/features/public-api/ (accessed Aug. 04, 2022).

[230] "When we share, everyone wins - Creative Commons." https://creativecommons.org/ (accessed Aug. 04, 2022).

[231] “GitHub - ORCID/ORCID-Source: ORCID Open Source Project.” https://github.com/ORCID/ORCID-Source (accessed Aug. 04, 2022).

[232] C. Boudry and M. Durand-Barthez, “Use of author identifier services (ORCID, ResearcherID) and academic social networks (Academia.edu, ResearchGate) by the researchers of the University of Caen Normandy (France): A case study,” *PLoS One*, vol. 15, no. 9 September, 2020, doi: 10.1371/journal.pone.0238583.

[233] M. Choraś and D. Jaroszewska-Choraś, “The scrutinizing look on the impending proliferation of mandatory ORCID use from the perspective of data protection, privacy and freedom of science,” *Interdiscip. Sci. Rev.*, vol. 45, no. 4, 2020, doi: 10.1080/03080188.2020.1780773.

[234] “OAuth 2.0 — OAuth.” https://oauth.net/2/ (accessed Aug. 05, 2022).

[235] “JSON Web Token Introduction - jwt.io.” https://jwt.io/introduction (accessed Aug. 05, 2022).

[236] “JSON.” https://www.json.org/json-en.html (accessed Aug. 06, 2022).

[237] “Welcome to Python Social Auth’s documentation! — Python Social Auth documentation.” https://python-social-auth.readthedocs.io/en/latest/ (accessed Aug. 06, 2022).

[238] A. Kirk, S. Timms, Æ. Rininsland, and S. Teller, *Data Visualization: Representing Information on Modern Web*. 2016.

[239] J. Dougherty and I. Ilyankou, “Hands-on data visualization,” *O’Reilly Media, Inc.* 2021.

[240] E. R. Tufte, “The visual display of quantitative information.” 1983. doi: 10.2307/530384.

[241] H. da Rocha, *Learn Chart.js: Create interactive visualizations for the Web with Chart.js 2*, 1st edition. Packt Publishing, 2019.

[242] “Chart.js | Open source HTML5 Charts for your website.” https://www.chartjs.org/ (accessed Aug. 10, 2022).

[243] “GitHub - chartjs/Chart.js: Simple HTML5 Charts using the <canvas> tag.” https://github.com/chartjs/Chart.js (accessed Aug. 10, 2022).

[244] “D3.js - Data-Driven Documents.” https://d3js.org/ (accessed Aug. 10, 2022).

[245] “Ace - The High Performance Code Editor for the Web.” https://ace.c9.io/ (accessed Aug. 10, 2022).

[246] “JavaScript.com.” https://www.javascript.com/ (accessed Aug. 14, 2022).

[247] “What is reputation? How do I earn (and lose) it? - Help Center - Stack Overflow.” https://stackoverflow.com/help/whats-reputation (accessed Aug. 19, 2022).

[248] DIN-Normenausschuss Ergonomie (NAErg) and Ergonomics Standards Committee, *DIN EN ISO 9241-210*, vol. 47. 2020.

[249] M. Schrepp, A. Hinderks, and J. Thomaschewski, “Applying the user experience questionnaire (UEQ) in different evaluation scenarios,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8517 LNCS, no. PART 1. doi: 10.1007/978-3-319-07668-3_37.

[250] B. Laugwitz, T. Held, and M. Schrepp, “Construction and evaluation of a user experience questionnaire,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 5298 LNCS. doi: 10.1007/978-3-540-89350-9_6.