# Exchange of Clinical and Omics Data According to FAIR Principles: A Review of Open Source Solutions

Philipp Pugliese[1] Christian Knell[1] Jan Christoph[1]

[1] Department of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany

Methods Inf Med 2020;59:e13–e20.

Address for correspondence Philipp Pugliese, BEng, Department of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Wetterkreuz 13, 91058 Erlangen, Germany (e-mail: phil.pugliese@fau.de).

## Abstract

**Background** Due to the ongoing increase and importance of the sustainable reusability of data, the findable, accessible, interoperable, reusable or FAIR principles were developed which are also relevant in translational research.

**Objectives** The study aims at identification of platforms by literature search that are suitable for implementation in translational research, in particular with regard to their FAIRness.

**Methods** The collected information is summarized and compared.

**Results** Platforms have been identified which are suitable for linking with other translational platforms with regard to documentation, long-term archiving, and processing as well as for FAIR handling of bioinformatic data.

**Conclusion** There are already platforms in the translational environment that take FAIR principles into account and thus improve translational research. Due to the specialization of the research platforms and the fact that FAIR are only principles and not standards, the platforms have to be examined in individual cases to see whether and how they can be integrated with other platforms.

**Keywords**

- FAIR
- biomedical research
- bioinformatic
- medical research
- ISA

## Introduction

Due to the rapid decline in costs, molecular biological data are increasingly collected and digitally analyzed in often complex manner, so that paper-based laboratory books, are reaching their limits to make these data and analyses reproducible. For this reason, the topics of documentation and long-term archiving are more and more moving into the focus of bioinformatics.[1]

In recent years in particular, both open source and proprietary platforms have been developed that offer a solution to these problems. Especially university institutions like university hospitals are therefore interested in a scientifically correct way to solve these issues. Therefore, the focus of this work is on open source solutions, as these are usually less expensive and most likely to be considered in terms of open science.

Translational research platforms such as cBioPortal[2] and tranSMART,[3] which offer sophisticated data models for the integration and analyses of clinical and omics data, form the basis of our considerations. These platforms allow the user to perform analyses via graphical web interfaces but usually do not provide support for documentation and long-term archiving. So far, this has to be achieved by using paper-based laboratory books.

This is complicated by the fact that there is currently no generally accepted international standard on how to structure data in this respect. But Wilkinson et al presented principles, the so-called FAIR (Findable, Accessible, Interoperable, Reusable) principles, which the data should follow to be sustainably reusable.

FAIR is an acronym for findable, accessible, interoperable, and reusable. These principles emphasize the ability of computer systems to find (meta-)data, access it, interact, or reuse it without or with minimal human intervention. This is becoming progressively necessary as bioinformatics data become increasingly comprehensive and complex. As a

**received**
July 31, 2019
**accepted after revision**
April 29, 2020

DOI https://doi.org/10.1055/s-0040-1712968.
ISSN 0026-1270.

result, computational support is needed.[5] In the following study, the four principles, which Wilkinson et al named, are briefly described:

*Findable*: (Meta-)data should be easy for both humans and computers to find, that is, machine-readable.

*Accessible*: (Meta-)data should be easily accessible by humans as well as machines using standard communication protocols and should be archivable in the long run. Authentication and/or authorization is not excluded.

*Interoperable*: (Meta-)data should be exchangeable, interpretable, and able to be combined with other datasets in a human- and machine-readable way in a (semi-)automated manner.

*Reusable*: (Meta-)data should be described so well that they can be reused in future research or that the results are reproducible. Furthermore, proper citation must be made possible.

Holub et al[6] extended the principles from FAIR to FAIR-Health. These include additional quality aspects regarding reproducibility, incentives for enrichment of datasets, and approaches for privacy-oriented work with human material and its data.

These principles serve to ensure data and workflow provenance by leading to more transparency and security. Data provenance describes the digital object itself (e.g., a data record) and strengthens its integrity by documenting who made which changes and how. Workflow provenance, on the other hand, strengthens the integrity of the processing of these objects by documenting which settings were made and how. This is of particular benefit to bioinformaticians and researchers, as a key component in biomedical research is transparency in the reporting of studies to provide reproducible results.[7]

## Objectives

Based on translational research platforms such as cBioPortal and tranSMART, which enable the integration and analysis of data but do not offer tools for FAIR data management, the aim of this work was to identify and compare platforms that close this gap in a structured way. With reference to the data and workflow provenance of these translational research platforms, the software solution should follow the FAIR principles, ideally the FAIR-Health principles.

A comparison of platforms that connect to these translational research tools to offer solutions for their documentation, long-term archiving, and processing of clinical and omics data according to FAIR principles does not yet exist to the best of our knowledge. In the following study, the procedure and results of the literature search for identifying such open source platforms for good scientific practice are described.

## Methods

We have used PubMed[8] to examine the scientific literature and identified 1.111 articles that potentially describe software solutions (PubMed queries are available in ►**Appendix 1**). The first query searched for articles that deal with data management or translational research in general. The second query then filtered out those for systems biology. The next query was to find the FAIR principles mentioned in the remaining articles. Finally, the results were filtered by agreement, published within the last 5 years and published in English. Then, we manually searched the literature to identify software solutions that could both (1) document and (long-term) archive medical/omics data and (2) process this data. A corresponding search with Google Scholar[9] completed the search.

At first, the identified articles were searched for compilations of potential software solutions, but nothing of that kind was found. However, a comparison was found by Wruck et al[10] in which platforms for large-scale systems biology were presented. This comparison, together with the platforms already examined by Wilkinson et al, was merged with the results of our keyword search.

After collating the potentially relevant platforms and their literature sources, those that did not seem to fit into the context of system biology or translational research platforms were filtered out.

For the creation of the matrices we initially oriented ourselves on Canuel et al,[11] since our aim is to link translational research platforms such as cBioPortal and tranSMART with a FAIR data management system and the study has already created an overview of the former. Therefore, we took over the axis community, system requirements, and support. The community axis grouped information about availability status and references. The system requirements axis provides information about the required operating system. The support axis supplies more detailed information about the help provided.

From Wruck et al the features of systems biological data management were (almost) completely taken over. This created the axis of requirements for a data management system for systems biology. This axis shows how data are collected, integrated, and processed. The data corresponding to the matrices have also been adopted. However, they have been checked and updated for topicality.

It was then examined whether the identified platforms were designated under the two FAIR initiatives FAIRDOM[12] or FAIRsharing.[13] The fact that these initiatives mention the identified platforms is an indication of the FAIRness of the platform in question. The FAIRDOM association references several research platforms for laboratory research projects in life sciences. The aim is to support sustainable research data management in these areas. FAIRsharing is a platform that contains a variety of curated descriptions of standards, databases, and data policies. The goal is to support reproducible and reusable scientific research. Thus, the axis of the FAIR indication was created. The matrix was then filled in as far as possible.

Finally, the development team or the responsible persons of the platforms were contacted to confirm the found information. We received feedback from each development team. The information we have found in literature which differs from the statements of the development team, is marked with "ª" in ►**Tables 1** and **2** (the information of the development teams was entered).

**Table 1** Details of the main features of the two identified platforms for workflow provenance

| Software | Galaxy | GenePattern |
|---|---|---|
| FAIR indication | | |
| FAIR reference | FAIRDOM | FAIRsharing |
| Community | | |
| Reference | Afgan et al 2018[a] | Reich et al 2006 |
| DOI or PMID | 29790989[a] | 16642009 |
| License | Academic Free License[a] | BSD |
| User mailing list or support | Yes | Yes |
| URL | https://galaxyproject.org/ | http://genepattern.org |
| System requirements | | |
| Operating system | Linux/MacOS | Linux/MacOS |
| DB managing system | PostgreSQL | HSQLDB, Oracle, MySQL[a] |
| Main programming language | Python, JavaScript[a] | Python/Java/R |
| Support | | |
| Installation procedures | Yes | Yes |
| Configuration documentation | Yes | Yes |
| User documentation | Yes | Yes |
| Regular maintenance | Yes | Yes |
| Miscellaneous | | |
| GitHub availability | Yes | Yes |
| Demo server availability | – | – |
| Requirements for data management in system biology | | |
| Compliance to standards | SBML | MAGE-TAB |
| Metadata model | SBML-related | MIAME-related |
| Automation of data collection | Unknown | Unknown |
| Upload to public repositories | ArrayExpress | ArrayExpress |
| Extensibility | Via XML files | Java extensions |
| Integration with other DMS | Unknown | Unknown |

Abbreviations: DB, database; DMS, data management system; DOI, digital object identifier; FAIR, findable, accessible, interoperable, reusable; PMID, PubMed-ID; SBML, systems biology markup language; URL, uniform resource locator.
[a]Data provided by the development team that differ from the data found in the literature.

## Results

During this literature search it became clear that Data and Workflow Provenance each requires its own platform. Therefore, the results are divided into two areas and compared in matrices with the above-mentioned properties. In the first part, therefore, the platforms that are used for processing are named. Processing means further use of the data, for example, for integration into work processes. In the second part, the platforms that are used for administration are named which means that the platform supports the functions of documentation and long-term archiving. In the following study, the identified platforms are therefore presented in a nonexhaustive enumeration.

### Platforms for Workflow Provenance

In biomedical disciplines, a workflow is typically used to perform complex data processing tasks. Workflow provenance refers to the recording of all derivations that led to the final output of the workflow. Thus, all steps of the workflow creation should be made transparent and reproducible. In addition to traceability, workflow provenance can also help to avoid redundant work; because the input of the individual steps is documented, they do not have to be re-entered each time.[14] The following platforms seem to be best suited for workflow provenance in processing.

#### Galaxy (2010)

It is a web-based open-source platform for computational biomedical research, which can also be used by researchers with no programming experience. It was developed by the Nekrutenko Laboratory at the Center for Comparative Genomics and Bioinformatics at Penn State University, the Taylor Laboratory at Johns Hopkins University, and the Goecks Laboratory at Oregon Health & Science University, along with contributions from the community. It allows analysis workflows to be executed on the basis of the researchers' data, shared across teams or other researchers to repeat the

**Table 2** Details of the main features of the eight identified platforms for data provenance.docx

| Software | BASE | Ensembl | ISA-Tools | LabKey Server | openBIS | SABIO-RK | SEEK |
|---|---|---|---|---|---|---|---|
| FAIR indication | | | | | | | |
| FAIR reference | Unknown | FAIRsharing | FAIRsharing | Unknown | FAIRDOM | FAIRDOM | FAIRDOM |
| Note | – | – | Original format: ISA-Tab | – | – | Non-Commercial Purpose License | – |
| Community | | | | | | | |
| Reference | Häkkinen et al. 2016[a] | Cunningham et al. 2019[a] | Rocca-Serra et al. 2010 | Nelson et al. 2011 | Bauch et al. 2011 | Wittig et al. 2012 | Wolstencroft et al. 2015 |
| DOI or PMID | https://doi.org/10.1101/038976[a] | 30407521[a] | 20679334 | 21385461 | 22151573 | 22102587 | 26160520 |
| License | GNU | Apache License 2.0 | CPAL | Apache License 2.0 | Apache License 2.0 | Non-Commercial License[a] | BSD |
| User mailing list or support | Yes | Yes | Yes | Yes[a] | Yes | Yes | Yes |
| URL | http://base.thep.l u.se | http://www.ense mbl.org/index.ht ml | https://isa-tools.org | https://labkey.org | https://labnotebo ok.ch | http://sabio.hits.org/ | https://seek4scie nce.org |
| System requirements | | | | | | | |
| Operating system | Platform independent | Linux[a] | Platform independent | Platform independent | Linux | Linux[a] | Linux |
| DB managing System | MySQL, PostgreSQL | MySQL | –[a] | MySQL, PostgreSQL[a] | PostgreSQL | PostgreSQL[a] | PostgreSQL[a] |
| Main programming language | Java | Perl | Python, Java[a] | Java, JavaScript | Java | Java, Grails[a] | Ruby on Rails |
| Support | | | | | | | |
| Installation | Yes | Yes | Yes | Yes | Yes | –[a] | Yes |
| Configuration | Yes | Yes | Yes | Yes | Yes | –[a] | Yes[a] |
| User documentation | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Regular maintenance | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| Miscellaneous | | | | | | | |
| GitHub availability | – | Yes | Yes | Yes | Only in GitLab[a] | – | Yes |
| Demo server availability | Yes | – | – | Yes | Yes | –[a] | Yes |
| Requirements for data management in system biology | | | | | | | |
| Compliance to standards | unknown | Track hubs | MIAME[a] | MIAME | mzXml (MIBBI projected) | SBML, BioPAX[a] | MIBBI |
| Metadata model | MIAME-related | Track hubs | ISATab-related | RDF | Generic/controlled vocabularies | SBML-related | JERM |
| Automation of data collection | batch import | unknown | batch import | batch import | dropboxes | no automation | harvesters |
| Upload to public repositories | ArrayExpress, GEO, CIBEX | –[a] | ArrayExpress | – | not yet, but in planning | –[a] | JWS-online |
| Extensibility | Java plugins / extentions | Perl modules | mostly java classes | via client API | Java extensions | java extensions | ruby classes |
| Integration with other DMS | –[a] | java servlets | reference integration | non-trivial | via interfaces to the framework | –[a] | good, planned |

Abbreviations: API, application programming interface; BASE, BioArray Software Environment; BSD, Berkeley Software Distribution; CPAL, Common Public Attribution License; DMS, data management system; DOI, digital object identifier; GEO, Gene Expression Omnibus; FAIR, findable, accessible, interoperable, reusable; FAIRDOM, Findable, Accessible, Interoperable, Reusable, Data, Operation, Models; gGmbH, gemeinnützige Gesellschaft mit beschränkter Haftung (eng.: not-for-profit limited liability company); GNU, proper name (GNU's not Unix); ISA, Investigation, Study, Assay; MIBBI, minimum information for biological and biomedical investigations; JERM, Just Enough Results Model; JWS, Java Web Service; MAGE-TAB, MicroArray Gene Expression Tabular; MIAME, Minimum Information About a Microarray Experiment; MySQL, My Structured Query Language; PMID, PubMed-Identificator; RDF, Resource Description Framework; SABIO-RK, system for the analysis of biochemical pathways—reaction kinetics; SBML, systems biology markup language; SysMO, Systems Biology for Micro-Organisms.

[a]Data provided by the development team that differ from the data found in the literature.

same analysis. It thus supports reproducibility and simplifies the exchange of data and results. Furthermore, the user does not need to compile and install a software.[15–18]

### GenePattern (2006)

It is a web-based open-source software package for computer-aided bioinformatics that is developed and distributed at the Broad Institute. It offers a variety of calculation methods for the analysis of genomic data. It enables researchers to analyze data and visualize their results. Reproducibility is ensured because the analysis methods used, the sequence of method applications, and the parameter settings are recorded. This ensures the data and workflow provenance. The extensible architecture makes it easy to add additional analysis and visualization modules so researchers can regularly access new methods.[19–21]

## Platforms for Data Provenance

In contrast to workflow provenance, data provenance provides a more detailed view of the integrity of the data. The focus is on the description of the digital object itself (where it comes from, how it has reached its current state, etc.). In the field of bioinformatics, data provenance can help to create trust in a system, e.g. to ensure verifiability.[22] The following platforms seem to be best suited for data prevention in document management and long-term archiving.

### BASE (2005)

BioArray Software Environment (BASE) is a web-based open-source laboratory informatics application developed at Lund University. It serves as an annotatable microarray data repository and analysis application that offers researchers both information management and analysis. The system integrates biomaterial information, raw images, and extracted data. The plug-in architecture provides data transformation, data display, and analysis modules. Individual elements can be shared with other users within the database.[23–26]

### Ensembl (1999)

It was developed as open-source software by the European Molecular Biology Laboratories-European Bioinformatics with the aim of providing creation, maintenance, and updating of reference genome annotation and comparative genomic resources. It thus offers a bioinformatic framework for the organization of biological sequences of genomes. The functionalities range from sequence analysis to data storage and visualization.[27–29]

### ISA-Tools (2010)

The open-source framework, developed and distributed by Investigation, Study, and Assay (ISA)-Tools, consists of five platform-independent components. ISA is an acronym for "Investigation," "Study" and "Assay" and describes a data model. These components are each desktop applications and function both as an independent and a unified system. The entire software package consists of the ISA-Tab (structuring and communication of metadata), ISAcreator (creation, editing, and import of experimental metadata sets), ISA configurator (editing input fields in ISAcreator), ISA validator (checking requirements and links of the finished data in ISA-Tab), and ISA converter (export to public repositories). The main objective is the administration and storage of experimental metadata from the fields of life sciences, environmental sciences, and biomedical sciences. It is based on the ISA framework of the same name for data models and serialization, so that the resulting data and findings are reproducible and reusable.[30,31]

### LabKey Server (2006)

The web-based open source software was developed under the name Computational Proteomics Analysis System at the Fred Hutchinson Cancer Research Center. Since 2011 this solution is distributed by LabKey and available under LabKey Server. Based on the data entered, projects are created in which the participating researchers can collaborate. The data can be integrated either from different databases or from XLSM files. The objective is the integration, analysis, and sharing of biomedical data, especially for very large amounts of data.[32,33]

### openBIS (2007)

The web-based open source software open Biology Information System (openBIS) is developed and distributed at ETH Zurich. But it is also offered through the FAIRDOM initiative. It consists of an information system which extracts (meta-)data from the measuring devices used and uploads them to the system via drop boxes (directories). After the import, the (meta-)data can be further processed and analyzed. The main objective is to support biological research data workflows from source to publication. This makes the data provenance more secure and allows researchers to access, edit, and share (meta-)data.[34,35]

### SABIO-RK (2006)

System for the Analysis of Biochemical Pathways-Reaction Kinetics (SABIO-RK) is a web-based database and stores comprehensive information about biochemical reactions and their kinetic properties. It is distributed by Hits gGmbH. It is free of charge except for commercial use. Unlike other databases, it is reaction-oriented and contains quantitative information about reaction dynamics. The data are manually entered and commented by the researchers. The consistency of the data is checked automatically. The database also supports data export, e.g., for further modeling tools, including annotation using the Systems Biology Markup Language.[36,37]

### SEEK (2009)

This web-based open-source solution consists of several tools and supports as cataloguing and community platform in the administration, exchange, and research of (meta-)data and models in systems biology. Originally developed for SysMO, a pan-European consortium for the study of molecular processes in microorganisms, it is now distributed by the FAIRDOM initiative. The access driven platform enables researchers to collaborate on projects on a daily basis and to publish data and models. The plug-in architecture allows the linking of experiments, their protocols, data, and models. This preserves associations between datasets, individuals, and organizations.[38,39]

## Conclusion

The variety of possible platforms for documentation, long-term archiving, and processing of clinical and omics data makes the search for a suitable platform a challenge; in particular because the solutions can be either too general or too specific.

Therefore, this work should give a structured overview of the platforms which basically exist, to make them comparable. In addition, it was also shown which platforms seem to take the FAIR principles into account. Thus, a first preselection for a suitable platform can be made. Within the scope of this literature research, it can therefore not be ruled out that further potentially suitable platforms may exist. In addition, the focus of the search was on open source platforms, commercial products were not considered. Furthermore, the ongoing development of the platforms cannot guarantee that the information provided here is up to date or how long this will remain so.

Based on this literature research, it was not possible to determine to what extent the platforms mentioned were actually implementing the FAIR principles. It was only possible to identify indications of compliance with the principles. The same also applies to the principles extended to FAIR-Health. The fulfilment of the principles by the identified platforms must therefore be assessed on a case-by-case basis, as well as their suitability for the individual needs of researchers and organizations.

A further result of this literature research was the finding that a distinction is made between the fulfilment of provenance between data and work processes. Therefore, it seems reasonable to select at least two platforms to ensure the provenance from the time of documentation to processing and (long-term) archiving.

### Related Work

In this review, a total of nine platforms were presented. Similarities and differences between the platforms in the two matrices were shown. In addition, indications of the compliance with the FAIR principles were identified.

Parts of the matrix created in this paper were adopted from Canuel et al. The study focused on the analysis of clinical and omics data itself, e.g., safety and interoperability functions. The ability to FAIR documentation or long-term archiving was not part of his scope. There was also a gap in terms of further processing of the data.

Wruck et al presented data management strategies for large-scale projects in systems biology. Parts of the matrix created in this paper were adopted. The focus of Wruck's work was more on the technical perspective than on a general overview of existing solutions. The FAIR principles were also not addressed because they had not yet been developed at the time of writing.

### Rating of the Identified Platforms

In order of documentation, further processing and (long-term) archiving the data models of clinical and omics data from translational research platforms such as cBioPortal and tranSMART in accordance with the FAIR principles, the identified platforms are now structured. Starting with those that we consider will meet our needs the most.

*SEEK* convinces with its functionality for linking data records, persons, and organizations within projects. This ensures the relationships between the objects and thus the data provenance. The platform also offers the possibility of daily collaboration. As part of the FAIRDOM initiative, it can be assumed that the FAIR principles are taken into account.

*Galaxy* processes and shares the uploaded data using analysis workflows. Galaxy is named by the FAIRDOM initiative. This already indicates that the processed data will be used according to FAIR principles. A further point is the user friendliness of the system toward noninformaticians. Although the field of bioinformatics requires a certain affinity to computer science, other end users may (at least partially) come from other, noninformatics disciplines such as human biology.

*openBIS* supports biological research, from the generation to the publication of research data. The permanent availability of the data within the research process ensures the data provenance. Furthermore, by participating in the FAIRDOM initiative, it can be concluded that the FAIR principles are taken into account.

*GenePattern* offers a variety of calculation options for the analysis of the uploaded data. The data provenance is additionally saved because parameter settings are recorded in addition to the method used. The software package apparently also takes the FAIR principles into account.

*ISA-Tools* convince with its relatively high flexibility: only the desired modules can be used without having to give up their full functionality. For example, the use of ISAcreator can be dispensed with if the input fields provided in ISA-Tab are sufficient for the metadata. On the other hand, you have to implement ISAcreator to edit the input fields. A further important criterion also seems to fulfil ISA-Tool: the compliance with the FAIR principles.

*Ensembl* also takes the FAIR principles into account, making it a potentially usable resource. However, specialization in genome sequences seems to limit its application to this discipline. Whether and to what extent this solution can be applied to other fields of bioinformatics must be examined in each individual case.

*LabKey* organizes datasets in projects, so researchers can potentially collaborate better than if the datasets were "loosely" distributed. Of particular interest is the functionality of importing data from different databases and files. Within this literature search, however, no clues could be found as to whether and to what extent the FAIR principles were taken into account. This point would have to be checked in a further investigation.

*BASE* is a software particularly developed for laboratory applications. The focus of this application seems to be limited to sequencing and microarray experiments. A potential use of this solution would therefore initially be limited to this area only. Furthermore, no evidence could be found to document participation in a FAIR initiative or consideration of the principles.

*SABIO-RK* provides a reaction-oriented database for biochemical reactions. Thus, the potential field of application is also strongly limited to this discipline. This database

considers the FAIR principles, at least there are indications for it in the literature.

## Future Work

The development of the FAIR principles creates a regulatory framework that facilitates the sustainable reusability of such data. This ensures the provenance of both the data and the workflow. Initiatives have been formed for their spreading and are already being considered by several platforms. Together with the further development of the principles, this underlines the importance of this subject.

However, it has also been shown that efforts have been made and are being made to improve the existing provenance in bioinformatics. Given the variety of platforms and the ever-increasing availability of clinical and omics data, a FAIR solution is essential for the consistent and sustainable reuse of this data.

In the next step the identified platforms, probably SEEK/openBIS for documentation and archiving and Galaxy for processing, will be linked with the translational research platforms cBioPortal and tranSMART established at the university hospital. This will ensure a FAIR and translational data management.

### Conflict of Interest

None declared.

### Acknowledgments

This research is supported by MIRACUM that is funded by the German Federal Ministry of Education and Research (BMBF) within the "Medical Informatics Funding Scheme" (FKZ 01ZZ1801A). The authors thank the contributors of the Base, Ensembl, Galaxy, GenePattern, ISA-Tools, LabKey Server, openBIS, SABIO-RK and SEEK platforms who answered our request for information. The present work was performed in fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." from the Friedrich-Alexander-Universität Erlangen-Nürnberg (PP).

## References

1 Grütz R, Mathieu N, Löhnhardt B, Weil P, Krawczak M. Archivierung von Genomdaten. Medgen 2013;25(03):388–394
2 Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6(269):pl1
3 Scheufele E, Aronzon D, Coopersmith R, et al. tranSMART: an open source knowledge management and high content data analytics platform. AMIA Jt Summits Transl Sci Proc 2014;2014:96–101
4 Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data 2016;3:160018
5 GO FAIR International Support and Coordination Office. FAIR principles. Available at: https://www.go-fair.org/fair-principles/. Accessed June 13, 2019
6 Holub P, Kohlmayer F, Prasser F, et al. Enhancing reuse of data and biological material in medical research: from FAIR to FAIR-health. Biopreserv Biobank 2018;16(02):97–105
7 Sahoo SS, Valdez J, Kim M, Rueschman M, Redline S. ProvCaRe: characterizing scientific reproducibility of biomedical research studies using semantic provenance metadata. Int J Med Inform 2019;121:10–18
8 National Center for Biotechnology Information. PubMed. Available at: https://www.ncbi.nlm.nih.gov/pubmed/. Accessed June 15, 2019
9 Google LLC. Google Scholar. Available at: https://scholar.google.de. Accessed June 15, 2019
10 Wruck W, Peuker M, Regenbrecht CRA. Data management strategies for multinational large-scale systems biology projects. Brief Bioinform 2014;15(01):65–78
11 Canuel V, Rance B, Avillach P, Degoulet P, Burgun A. Translational research platforms integrating clinical and omics data: a review of publicly available solutions. Brief Bioinform 2015;16(02):280–290
12 FAIRDOM Association e. V. FAIRDOM. Available at: https://fairdom.org. Accessed June 13, 2019
13 University of Oxford. FAIRsharing. Available at: https://fairsharing.org. Accessed June 13, 2019
14 Tan W-C. Provenance in databases: past, current, and future. IEEE Data Eng Bull 2007;30(04):3–12
15 ELIXIR. Galaxy Community: To promote FAIR principles in Galaxy. Available at: https://elixir-europe.org/communities/galaxy. Accessed June 13, 2019
16 Pennsylvania State University & Emory University. Galaxy Project. Available at: https://galaxyproject.org/. Accessed June 13, 2019
17 Giardine B, Riemer C, Hardison RC, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 2005;15(10):1451–1455
18 Afgan E, Baker D, Batut B, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res 2018;46(W1):W537–W544
19 Broad Institute. GenePattern. Available at: http://software.broadinstitute.org/cancer/software/genepattern/. Accessed June 13, 2019
20 Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet 2006;38(05):500–501
21 Reich M, Tabor T, Liefeld T, et al. The GenePattern notebook environment. Cell Syst 2017;5(02):149–151.e1
22 Xu S, Rogers T, Fairweather E, Glenn A, Curran J, Curcin V. Application of data provenance in healthcare analytics software: information visualisation of user activities. AMIA Jt Summits Transl Sci Proc 2018;2017:263–272
23 Lund University. BASE. Available at: http://base.thep.lu.se. Accessed June 13, 2019
24 Vallon-Christersson J, Nordborg N, Svensson M, Häkkinen J. BASE—2nd generation software for microarray data management and analysis. BMC Bioinformatics 2009;10:330
25 Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biol 2002;3(08):E0003
26 Häkkinen J, Nordborg N, Månsson O, Vallon-Christersson J. Implementation of an open source software solution for laboratory information management and automated RNAseq data analysis in a large-scale Cancer Genomics initiative using BASE with extension package Reggie. bioRxiv. 2016:38976. Doi:10.1101/038976
27 European Bioinformatics Institute and Sanger Centre. e!Ensembl. Available at: http://www.ensembl.org/index.html. Accessed June 13, 2019
28 Hubbard T, Barker D, Birney E, et al. The Ensembl genome database project. Nucleic Acids Res 2002;30(01):38–41
29 Cunningham F, Achuthan P, Akanni W, et al. Ensembl 2019. Nucleic Acids Res 2019;47(D1):D745–D751
30 ISA-Tools. ISA. Available at: https://isa-tools.org. Accessed June 13, 2019
31 Rocca-Serra P, Brandizi M, Maguire E, et al. ISA software suite: supporting standards-compliant experimental annotation and enabling curation at the community level. Bioinformatics 2010;26(18):2354–2356

32 LabKey-Server. LabKey support. Available at: https://labkey.org. Accessed June 13, 2019
33 Nelson EK, Piehler B, Eckels J, et al. LabKey Server: an open source platform for scientific data integration, analysis and collaboration. BMC Bioinformatics 2011;12:71
34 Zurich ETH. openBIS. Available at: https://labnotebook.ch. Accessed June 13, 2019
35 Bauch A, Adamczyk I, Buczek P, et al. openBIS: a flexible framework for managing and analyzing complex data in biology research. BMC Bioinformatics 2011;12:468
36 KTF. BMBF, DFG LIS. SABIO-RK. Available at: http://sabio.h-its.org/. Accessed June 13, 2019
37 Wittig U, Kania R, Golebiewski M, et al. SABIO-RK—database for biochemical reaction kinetics. Nucl Acids Res 2012;40(Database issue):D790–D796
38 Wolstencroft K, Owen S, Krebs O, et al. SEEK: a systems biology data and model management platform. BMC Syst Biol 2015;9:33
39 FAIRDOM Association e. V. seek4science. Available at: https://seek4science.org. Accessed June 13, 2019

**Appendix 1** Detailed queries used for the interrogation of PubMed database (queries last run on July 28, 2019)

| Query number | Query | Items found |
|---|---|---|
| 1 | (((data management[Title/Abstract]) OR (database[Title/Abstract] OR database[Title/Abstract])) OR (data repository[Title/Abstract] OR data repositories[Title/Abstract])) OR (translational platform [Title/Abstract] OR translational platforms[Title/Abstract] OR translational research platform[-Title/Abstract] OR translational research platforms[Title/Abstract]) | 283 033 |
| 2 | ((system biology[Title/Abstract] OR systems biology[Title/Abstract])) OR (bioinformatic[-Title/Abstract] OR bioinformatics[Title/Abstract]) | 66 933 |
| 3 | ((((((((((data sharing[Title/Abstract]) OR data analysis[Title/Abstract]) OR FAIR[Title/Abstract]) OR (Findable[Title/Abstract] AND Accessible[Title/Abstract] AND Interoperable[Title/Abstract] AND Reusable[Title/Abstract])) OR (metadata standard[Title/Abstract] OR metadata standards[-Title/Abstract])) OR (metadata[Title/Abstract] OR meta data[Title/Abstract])) OR open source [Title/Abstract]) OR software[Title/Abstract]) OR (solution[Title/Abstract] OR solutions[-Title/Abstract])) OR (standard[Title/Abstract] OR standards[Title/Abstract])) OR (strategy[-Title/Abstract] OR strategies[Title/Abstract]) | 2 647 727 |
| 4 | (#1 AND #2) | 6 830 |
| 5 | (#3 AND #4) | 2 263 |
| 6 | *Filters: Best match AND published in the past 5 years AND English [Language]* | **1 111** |