

Related to other papers in this special issue	29 (p285); 4 (p40); 9 (p87); 17 (p171); 2 (p10); 16 (p158); 21 (p208); 3 (p30); 7 (p66); 5 (p47)
Addressing FAIR principles	F1, F2, R1, R1.1, R1.2, R1.3

A Generic Workflow for the Data FAIRification Process

Annika Jacobsen^{1†}, Rajaram Kaliyaperumal¹, Luiz Olavo Bonino da Silva Santos²,
Barend Mons^{1,2}, Erik Schultes², Marco Roos¹ & Mark Thompson¹

¹Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

²GO FAIR International Support & Coordination Office (GFISCO), Leiden, The Netherlands

Keywords: FAIR data; FAIRification workflow; FAIR data stewardship; Hands-on FAIRification; FAIR dissemination

Citation: A. Jacobsen, R. Kaliyaperumal, L.O. Bonino da Silva Santos, B. Mons, E. Schultes, M. Roos & M. Thompson. A generic workflow for the data FAIRification process. *Data Intelligence* 2(2020), 56–65. doi: 10.1162/dint_a_00028

ABSTRACT

The FAIR guiding principles aim to enhance the Findability, Accessibility, Interoperability and Reusability of digital resources such as data, for both humans and machines. The process of making data FAIR (“FAIRification”) can be described in multiple steps. In this paper, we describe a generic step-by-step FAIRification workflow to be performed in a multidisciplinary team guided by FAIR data stewards. The FAIRification workflow should be applicable to any type of data and has been developed and used for “Bring Your Own Data” (BYOD) workshops, as well as for the FAIRification of e.g., rare diseases resources. The steps are: 1) identify the FAIRification objective, 2) analyze data, 3) analyze metadata, 4) define semantic model for data (4a) and metadata (4b), 5) make data (5a) and metadata (5b) linkable, 6) host FAIR data, and 7) assess FAIR data. For each step we describe how the data are processed, what expertise is required, which procedures and tools can be used, and which FAIR principles they relate to.

[†] Corresponding author: Annika Jacobsen (E-mail: a.jacobsen@lumc.nl, ORCID: 0000-0003-4818-2360).

1. INTRODUCTION

The FAIR data principles aim to enable efficient and error-free analysis of data from multiple sources by machines and ultimately by humans, through enhancing their Findability, Accessibility, Interoperability and Reusability [1]. Since the initiation of the FAIR principles in 2014, FAIR metrics [2, 3, 4], FAIR infrastructure [5] and FAIR tools [6] have been developed to aid the process of making data FAIR (“FAIRification”). At the same time, a FAIRification workflow emerged, which was first developed for and subsequently matured through numerous “Bring Your Own Data” (BYOD) workshops [7], and FAIRification projects of rare disease patient registries. The aim was a domain-independent workflow that may be used in a wide range of FAIRification efforts. Indeed, the workflow has been applied in BYODs on e.g., tomato, yeast, cancer and several in the rare disease domain. The rare disease domain provided an excellent use case, considering the evident need to efficiently analyze sparse, heterogeneous and privacy-sensitive data from multiple sources across institutes and countries.

Early drafts of the FAIRification workflow originate from the first BYODs before the inception of FAIR. These BYODs had a focus on interoperability following earlier work that applied semantic Web technology as a framework for interoperability and machine-readability [8, 9]. With the advent of FAIR, the workflow was adapted to also cover the other three facets of FAIR: findability, accessibility and reusability, which heavily relies on metadata (i.e., data descriptors). This evolution has resulted in the workflow presented in this paper, which we intend as a template and a guide: in practice we expect that instantiations of the workflow may vary depending on the data source, use case, domain or FAIRification objective.

2. A GENERIC FAIRIFICATION WORKFLOW

The details of the generic step-by-step FAIRification workflow can be seen in Figure 1. The workflow is divided into three phases: pre-FAIRification, FAIRification and post-FAIRification, which are further divided into seven steps. The steps include: 1) identify FAIRification objective, 2) analyze data, 3) analyze metadata, 4) define semantic model for data (4a) and metadata (4b), 5) make data (5a) and metadata (5b) linkable, 6) host FAIR data, and 7) assess FAIR data. The steps do not always need to be followed in a strict sequential order and may be iterated. Please note that with each step there are multiple smaller “steps” that themselves also may be iterated. Data sets differ and practical constraints or new insights may lead to a different order of execution and some steps are often visited multiple times. Each step attempts to enable the implementation of the FAIR principles and aims to enhance the FAIR status (i.e., “FAIRness”) of the data set.

Data FAIRification requires different types of expertise and should therefore be carried out in a multidisciplinary team guided by FAIR data steward(s). The different sets of expertise are on i) the data to be FAIRified and how they are managed, ii) the domain and the aims of the data resource within it, iii) architectural features of the software that is (or will be) used for managing the data, iv) access policies applicable to the resource, v) the FAIRification process (guiding and monitoring it), vi) FAIR software services and their deployment, vii) data modelling, viii) global standards applicable to the data resource, and ix) global standards for data access. A good working approach is to organize a team that contains or

has access to the required expertise. The core of such a team may be formed by data stewards, with at least expertise of the local environment and of the FAIRification process in general.

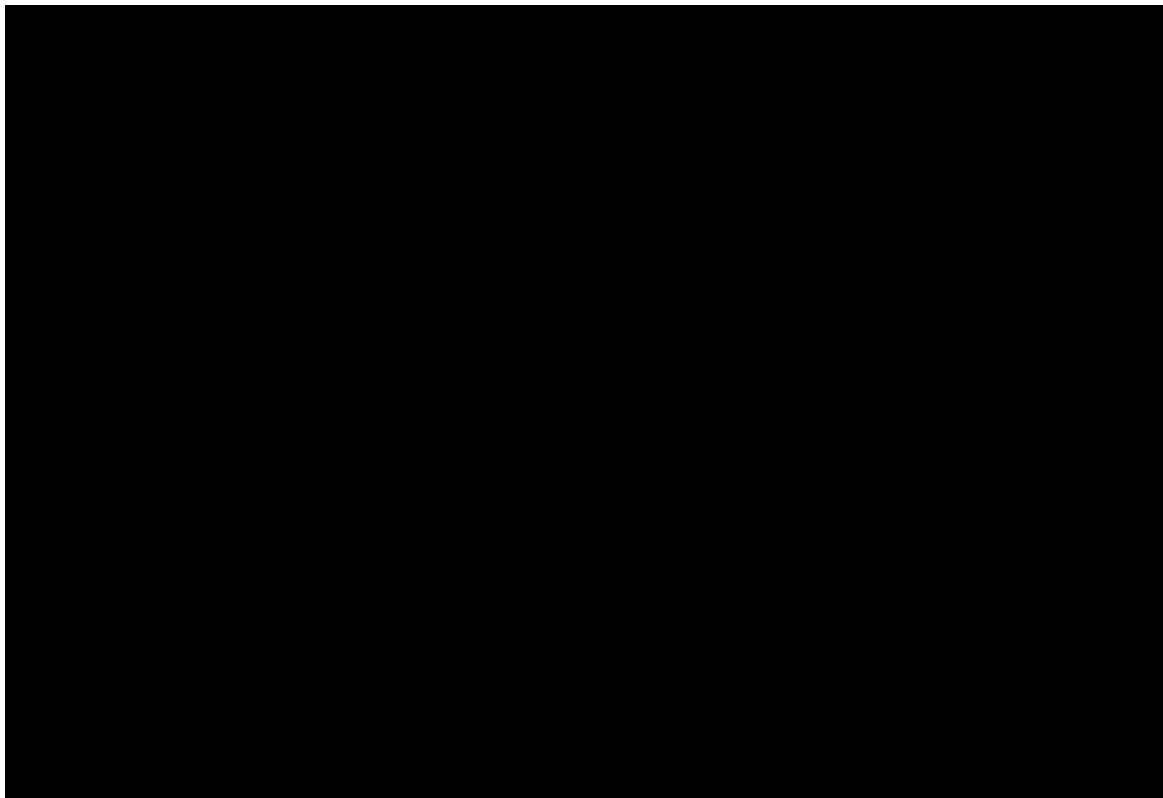


Figure 1. A generic workflow for the Data FAIRification process. The order is not strict and can be iterative. The workflow consists of the following steps: 1) identify the FAIRification objective, 2) assess the current state of the data, 3) develop a FAIRification plan, 4) implement the FAIRification plan, 5a) make data linkable, 5b) make metadata linkable, 6) host FAIR data, and 7) assess FAIR data.

2.1 Identify FAIRification Objective (Step 1)

The first step is to identify the FAIRification objective and is within the pre-FAIRification phase of the workflow. This step requires having access to the data, or in the case of privacy-sensitive data (when even the data steward should not get access to the actual information), a sample of anonymized or mocked data may be used. This step also requires having a general knowledge and understanding of the data set, as well as being familiar with the FAIR principles in general. Objectives for FAIRification could be specific requirements of publishers, funders [10] or stakeholder communities [11, 12], or to increase the efficiency of using data from multiple sources. We recommend to first focus the FAIRification on a subset of the data

elements in line with available resources for FAIRification (e.g., time). The workflow can be iterated so more data elements may be included later. A good way to select the subset is by defining domain relevant “driving user question(s)” that require at least two data resources. This should be done in a team with both domain and data modelling expertise. Other good drivers for implementing FAIR principles are to enhance the findability, accessibility or reusability of the data, e.g., by improving the metadata. Eventually, the FAIRification objective depends on the availability of 1) expertise, 2) FAIR solutions that may be reused [11, 12], and 3) data management tools with FAIRification features that are appropriate to the data set [13].

2.2 Analyze Data (Step 2)

The second step is to analyze the data to prepare for subsequent FAIRification (e.g., improving interoperability) and is within the pre-FAIRification phase of the workflow. This process may include: 1) investigating the data in whatever form(s) it is available (specified in Step 1) and checking whether both the data representation (format) and the meaning of the data elements (the data semantics) are clear and unambiguous, and 2) checking whether the data already contain FAIR features, such as persistent unique identifiers for data elements [14] (FAIR principle F1 [1]) by e.g., using FAIRness assessment tooling [2, 3, 4]. It is evident that this step is tightly connected with Step 1 since e.g., selecting a relevant subset of the data and defining driving user questions(s) are highly relying on being familiar with the data.

2.3 Analyze Metadata (Step 3)

The third step is to analyze the availability of metadata regarding findability, accessibility, and reusability, and is within the pre-FAIRification phase of the workflow (note, metadata is made interoperable in Steps 4b and 5b). This process may include: 1) investigating the metadata describing the data, or if no metadata exists, identifying what metadata should be gathered (which may be very unique for each stakeholder community), and 2) checking whether the metadata already contains FAIR features, such as rich metadata and provenance descriptions (FAIR principles F2 and R1, R1.1, R1.2, R1.3 [1]) by e.g., using FAIRness assessment tooling [2, 3, 4]. Improving metadata regarding findability, accessibility, and reusability requires including details such as: license, copyright, contributions statements (e.g., funders, data set creator, publisher), and description of use conditions and access of data.

2.4 Define Semantic Data and Metadata Model (Step 4a and 4b)

The fourth step is to define a semantic model of the data (4a) and the metadata (4b) and is within the FAIRification phase of the workflow. The semantic models are templates for the next step transforming the data (5a) and metadata (5b) into a machine-readable format. Generating a semantic model is often the most time-consuming step of data FAIRification. However, we expect the modelling effort to diminish as more and more models are made available for reuse over time, especially if such models are treated as FAIR digital objects themselves. Thus, it is important to first check whether a semantic model already exists for the data and the metadata that may be reused. For cases where no semantic model is available a new one needs to be generated. We briefly describe this process below.

Building a semantic data model (4a) can be defined in three steps: 1) creating a conceptual model, 2) searching for ontology terms, and 3) creating a semantic data model from Steps 1) and 2). This requires domain expertise on the data set and expertise in semantic data modelling. The domain expert(s) make sure that the exact meaning of the data is understood by the modeler and the modeler ensures that the semantic representation correctly represents the domain knowledge. It is important that both the data representation (format) and the meaning of the data elements (the data semantics) are clear and unambiguous (as mentioned in Step 2). A good vehicle for this discussion is to first create an abstract “conceptual model”, which lists the main concepts and relationships between data elements to be FAIRified, e.g., related to the subset specified in the driving user question(s).

Next, the concepts and relations between the data elements in the data set are substituted with the machine-readable classes and properties from ontologies, vocabularies and thesauri. While acknowledging the differences between the latter types of resources, we will subsequently use the ontologies as proper ontologies generally best serve the FAIRification process. Ontologies, and the concepts and properties that they describe, can be found using search engines, such as the Ontology Lookup Service (OLS)[†], BioPortal[‡] and BARTOC[¶]. We have found that making optimal choices, demands good searching skills and experience. For instance, it is generally insufficient to just choose the first ontology in the list provided by ontology search tools by definition. Instead one should also check the usability license, usage statistics, update activity, whether the ontology contains a good class and property structure (which generally facilitates data integration), and whether a general ontological framework is used (such as OBO Foundry [15]). Nevertheless, it may be very difficult to decide which term from which ontology should be used, i.e., to match the detail in domain specific ontologies with the detail that is needed to describe data elements correctly. Terms used in human narrative do not always match directly with the ontological representation of the term. If the search is unsuccessful, new ontology terms could be defined and added to existing ontologies or new ontologies could be developed. This is however a time-consuming process that should be undertaken with a team of experts from both the domain of the study as well as in consultation with ontology experts.

Finally, the conceptual model and the ontology terms are used to create a detailed semantic data model that in contrast to the conceptual model, distinguishes between the data items (instances and their values) and their types (classes). This model is an exact representation of the data and exposes the meaning of the data in machine-readable terms (ideally in the most universal form possible). This enables the transformed FAIR data set to be efficiently incorporated in other systems, analysis workflows, and unforeseen future applications.

Finally, the conceptual model and the ontology terms are used to create a detailed semantic data model that in contrast to the conceptual model, distinguishes between the data items (instances and their values) and their types (classes). This model is an exact representation of the data and exposes the meaning of the data in machine-readable terms (ideally in the most universal form possible). This enables the transformed

[†] <https://www.ebi.ac.uk/ols/index>.

[‡] <https://bioportal.bioontology.org/>.

[¶] <https://bartoc.org>.

FAIR data set to be efficiently incorporated in other systems, analysis workflows, and unforeseen future applications.

For metadata (4b), semantic models describing generic items are available to be reused, e.g., DCAT to describe data set description (see 5b describing tools with reusable semantic models). Domain-specific items (e.g., described by principles F2 and R1, R1.1, R1.2, R1.3 [1]) should be decided by each individual self-identified domain [11,12], and need thereafter to be described in a semantic metadata model.

2.5 Make Data and Metadata Linkable (Step 5a and 5b)

The fifth step is to make the data (5a) and metadata (5b) linkable i.e., transformed to a FAIR representation and is within the FAIRification phase of the workflow. The method for making data and metadata linkable is highly application and use case dependent. However, it is crucial that a description of the data and metadata is available in a representation framework that is globally understood by machines. Further, the semantic model should be associated with the data and metadata so that it is available for unforeseen future applications and scalable interoperability across all types of data. An example of a linkable machine-readable global framework is the Resource Description Framework (RDF). It provides a common and straightforward underlying model and creates a powerful global virtual knowledge graph.

In order to transform the data into a machine-readable form (Step 5a) the semantic data model defined (or chosen) in Step 4a is required. Specialized tools are available for this process such as the FAIRifier, which provides insight into the transformation process and makes the process reproducible by tracking intermediate steps [6]. Other similar tools are Karma [16], Rightfield [17], and OntoMaton [18].

For the transformation of the metadata into a machine-readable form (Step 5b) the semantic metadata model defined (or chosen) in Step 4b is required. For some generic metadata items there are several tools available that support this transformation process such as the FAIR Metadata Editor [6], CEDAR [19], and BioschemasGenerator. The FAIR Metadata Editor is a free online tool that demonstrates the concept of structuring metadata in a FAIR-supporting way. Good metadata increases the potential to make a resource more findable. We mention two additional mechanisms to increase the findability of a resource. First, we recommend registering a resource in a domain-relevant registry or index, preferably one that strives for FAIR-compliance. Second, to enable indexing of the data set by general purpose Web search engines such as Google, we recommend including Schema.org markup (or a domain specific variant like Bioschemas) for example using the DataCatalog and Dataset profiles.

2.6 Host FAIR Data (Step 6)

The sixth step is to host the FAIR data i.e., make it available for consumption and is within the FAIRification phase of the workflow. This enables human and machine use through different interfaces, such as an Application Programming Interface (API), RDF triple store, or Web application. Please note that “FAIR does not mean open” [20] and that access restrictions may be applied at any level of (meta)data on each of the

interfaces. There are many different ways to deploy a FAIR resource online and to provide (and manage) access [21, 22]. One of these is the general-purpose FAIR data accessor as provided by the FAIR Data Point (FDP) software component [6]. It is developed as an exemplar tool to demonstrate the critical step of using global standards to provide access to structured metadata, and to demonstrate compliance with the FAIR guiding principles [1]. An FDP facilitates transparent, controlled access in a stepwise manner to increasingly detailed information about the data set and eventually the data records to both humans and machines. The human interface consists of a simple Web page providing links to the relevant layers of metadata provided by the FDP. The FDP machine interface will return a machine-readable RDF document.

2.7 Assess FAIR Data (Step 7)

The seventh step is to assess the FAIR data and is within the post-FAIRification phase of the workflow. This process may include: 1) an evaluation to check whether the original objectives as defined in Step 1 have been achieved (if not, some of the steps in the workflow may need to be revisited), and 2) checking the FAIR status of the data and metadata by e.g., using FAIRness assessment tooling [2, 3, 4], and compare it with the FAIR status assessed in Steps 2 and 3.

If driving user question(s) were defined in Step 1 it should be “answered” in this step. The results of these question(s) are gathered by processing the FAIR machine-readable data. If RDF is the machine-readable format used, then RDF data stores (triple stores) are used to store the machine-readable data, and SPARQL queries are used to retrieve the data required to answer the driving user question(s).

3. DISCUSSION AND CONCLUSIONS

In this paper, we have described a generic workflow for the data FAIRification process. It mainly describes the technical hands-on part, but can also be used for other purposes, such as planning, training and dissemination. The purpose of this workflow is to make FAIRification easier. However there are specific decisions beyond this workflow that need to be made by stakeholders who are part of an organization, institution, consortium, or other relevant collective supporting the FAIRification. This for instance pertains to decisions regarding: 1) a common standard for FAIR (meta)data collection and storage within a given community, 2) the FAIRification plan prior to data capture, and 3) sharing semantic data models within and across communities and domains.

Stakeholders also need to consider managerial aspects of FAIRification, i.e., the required expertise, and building and maintaining capacity on FAIR data stewardship for the longer term. The way a FAIR project is approached depends on the available budget, and on the type and size of an organization. Capacity could be established by organizing a dedicated team of specialists within a larger organization, or by organizing collaboration with experts who are willing to contribute expertise and part-time consultancy to guide the process. In either case, we recommend that a small team of experts comprised of one or more trained FAIR data stewards is formed to maintain a full overview of the FAIRification process as it is implemented. The FAIRification workflow, including its required budget, should be explicitly incorporated in Data Management

Plans (DMPs) [13]. Because of the interdisciplinary nature of FAIRification and the early stage of development of support for FAIR data stewards, a DMP typically involves interdisciplinary and cross-organizational collaboration.

The FAIRification workflow presented in this paper is generic, thus intended to be used in any domain. Also, we note that our current workflow representation is not intended as a normative or final workflow for the FAIRification process. It should be used as a template, and we expect continuous evolution of the workflow as the awareness and understanding of specific data stewardship issues increases in application communities: for example, we are currently considering whether a planning phase and inventory phase should be added as explicit separate steps, or whether the workflow should change to accommodate FAIRification-by-design (e.g., before starting data capture) as opposed to FAIRification of an existing data set (post-hoc-FAIRification) [23]. The future composition of the workflow highly depends on which implementation decisions are made [11], and domains start to reuse solutions from other domains, we will see a creolization in the workflow [12].

AUTHOR CONTRIBUTIONS

The workflow presented in the manuscript is a result of many years of experience by all authors, A. Jacobsen (a.jacobsen@lumc.nl), R. Kaliyaperumal (R.Kaliyaperumal@lumc.nl), L.O. Bonino da Silva Santos (luiz.bonino@go-fair.org), B. Mons (barend.mons@go-fair.org), E. Schultes (erik.schultes@go-fair.org), M. Roos (m.roos@lumc.nl) and M. Thompson (m.thompson@lumc.nl). A. Jacobsen and M. Thompson are the lead in writing the manuscript. All authors contributed to the writing and provided critical feedback to help shape the manuscript.

ACKNOWLEDGEMENTS

The work of A. Jacobsen, R. Kaliyaperumal, M. Roos and M. Thompson is supported by funding from the European Union's Horizon 2020 research and innovation program under the EJP RD COFUND-EJP N° 825575. The work of A. Jacobsen, R. Kaliyaperumal, M. Roos and M. Thompson is supported by funding from ELIXIR EXCELERATE, H2020 grant agreement number 676559. M. Roos and M. Thompson received funding from NWO (VWData 400.17.605) and H2020-EU 824087. The work of B. Mons and L.O. Bonino da Silva Santos is funded by the H2020-EU 824068 and the GO FAIR ISCO grant of the Dutch Ministry of Science and Culture.

REFERENCES

- [1] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, ... & B. Mons. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.

- [2] M.D. Wilkinson, S.A. Sansone, E. Schultes, P. Doorn, L.O. Bonino da Silva Santos & M. Dumontier. Comment: A design framework and exemplar metrics for FAIRness. *Scientific Data* 5(2018), 1–4. doi: 10.1038/sdata.2018.118.
- [3] M.D. Wilkinson, M. Dumontier, S.A. Sansone, L.O. Bonino da Silva Santos, M. Prieto, D. Batista, ... & E. Schultes. Evaluating FAIR maturity through a scalable, automated, community-governed framework. *bioRxiv preprint*, 2019. doi: 10.1101/649202.
- [4] R. de Miranda Azevedo & M. Dumontier. Considerations for the conduction and interpretation of FAIRness evaluations. *Data Intelligence* 2(2020), 285–292. doi: 10.1162/dint_a_00051.
- [5] T. Weigel, U. Schwardmann, J. Klump, S. Bendoukha & R. Quick. Making data and workflows findable for machines. *Data Intelligence* 2(2020), 40–46. doi: 10.1162/dint_a_00026.
- [6] M. Thompson, K. Burger, R. Kaliyaperumal, M. Roos & L.O. Bonino da Silva Santos. Making FAIR easy with FAIR tools: From creolization to convergence. *Data Intelligence* 2(2020), 87–95. doi: 10.1162/dint_a_00031.
- [7] B. Hooft, C. Goble, C. Evelo, M. Roos, S. Sansone, F. Ehrhart, ... & B. Mons. ELIXIR-EXCELERATE D5.3: Bring Your Own Data (BYOD). doi: 10.5281/zenodo.3207809.
- [8] C. Haupt, A. Waagmeester, M. Zimmermann & E. Willighagen. Guidelines for exposing data as RDF in Open PHACTS. Available at: <http://www.openphacts.org/specs/2013/WD-rdfguide-20131007/>.
- [9] S.A. Sansone, P. Rocca-Serra, D. Field & E. Maguire. Toward interoperable bioscience data. *Nature Genetics* 44(2)(2012), 121–126. doi: 10.1038/ng.1054.
- [10] M. Bloemers & A. Montesanti. The FAIR funding model: Providing a framework for research funders to drive the transition toward FAIR data management and stewardship practices. *Data Intelligence* 2(2020), 171–180. doi: 10.1162/dint_a_00039.
- [11] A. Jacobsen, R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, ... & E. Schultes. FAIR principles: Interpretations and implementation considerations. *Data Intelligence* 2(2020), 10–29. doi: 10.1162/dint_r_00024.
- [12] L. Lannom, D. Koureas & A.R. Hardisty. FAIR data and services in biodiversity science and geoscience. *Data Intelligence* 2(2020), 122–130. doi: 10.1162/dint_a_00034.
- [13] S. Jones, R. Pergl, R. Hooft, T. Miksa, R. Samors, J. Ungvari, R.I. Davis & T. Lee. Data management planning: How requirements and solutions are beginning to converge. *Data Intelligence* 2(2020), 208–219. doi: 10.1162/dint_a_00043.
- [14] N. Juty, S.M. Wimalaratne, S. Soiland-Reyes, J. Kunze, C.A. Goble & T. Clark. Unique, persistent, resolvable: Identifiers as the foundation of FAIR. *Data Intelligence* 2(2020), 30–39. doi: 10.1162/dint_a_00025.
- [15] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, ... & S. Lewis. The OBO Foundry: Coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology* 25(2007), 1251–1255. doi: 10.1038/nbt1346.
- [16] Karma: A data integration tool. Available at: <https://usc-isi-i2.github.io/karma/>.
- [17] K. Wolstencroft, S. Owen, M. Horridge, O. Krebs, W. Mueller, J.L. Snoep, F. du Preez & C. Goble. RightField: Embedding ontology annotation in spreadsheets. *Bioinformatics* 27(14)(2011), 2021–2022. doi: 10.1093/bioinformatics/btr312.
- [18] E. Maguire, A. González-Beltrán, P.L. Whetzel, S.A. Sansone & P. Rocca-Serra. OntoMaton: A Bioportal powered ontology widget for Google Spreadsheets. *Bioinformatics* 29(4)(2013), 525–527. doi: 10.1093/bioinformatics/bts718.
- [19] M.A. Musen, C.A. Bean, K.-H. Cheung, M. Dumontier, K.A. Durante, O. Gevaert, ... & the CEDAR team. The center for expanded data annotation and retrieval. *Journal of the American Medical Informatics Association* 22(6)(2015), 1148–1152. doi: 10.1093/jamia/ocv048.

- [20] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L.O. Bonino da Silva Santos & M. D. Wilkinson. Cloudy, increasingly FAIR; Revisiting the FAIR data guiding principles for the European Open Science Cloud. *Information Services & Use* 37(1)(2017), 49–56. doi: 10.3233/ISU-170824.
- [21] C. Brewster, B. Nouwt, S. Raaijmakers & J. Verhoosel. Ontology-based access control for FAIR data. *Data Intelligence* 2(2020), 66–77. doi: 10.1162/dint_a_00029.
- [22] A. Landi, M. Thompson, V. Giannuzzi, F. Bonifazi, I. Labastida, L.O. Bonino da Silva Santos & M. Roos. The “A” of FAIR – as open as possible, as closed as necessary. *Data Intelligence* 2(2020), 47–55. doi: 10.1162/dint_a_00027.
- [23] E.A. Schultes, A. Jacobsen, K. Hettne, M. Thompson, M. Kuzak, R. Hooft, ... & C. Evelo. Essential steps of the FAIRification Process. OSF, 2019. osf.io/avrys.