# Deliverable 6.4

# CLARIFY Repositories for Open Access

| | |
|---|---|
| **Project** | **CLARIFY – Cloud ARtificial Intelligence For pathologY** |
| **Grant Agreement ID:** | **860627** |
| **Consortium coordinator:** | **UNIVERSITAT POLITECNICA DE VALENCIA** |
| **Start and end date:** | **1 November 2019 - 31 October 2023** |
| **Funded under:** | **H2020-EU.1.3.1.** |
| **Date of issue:** | **7-May-2021** |
| **Due date:** | **30-Apr-2021** |
| **Leader in charge of deliverable:** | **bitYoga AS** |

| **Dissemination level** | |
|---|---|
| **X** | PU = Public |
| | PP = Restricted to other programme participants (including the EC) |
| | RE = Restricted to a group specified by the consortium (including the EC) |
| | CO = Confidential, only for members of the consortium (including the EC) |

European Commission

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska Curie grant agreement No 860627.

# CHANGE REGISTER

| Version | Date | Author | Organisation | Changes |
|---|---|---|---|---|
| A_DRAFT | 06-Apr-2021 | Jiahui Geng | bY | Initialization |
| | 23-Apr-2021 | Russel Wolff | bY | Circulated for comments. |
| | 28-Apr-2021 | Kjersti Engan | UiS | Comments to the previous version |
| | 30-Apr-2021 | Jiahui Geng | bY | Updated version |
| | 5-May-2021 | Sandra Morales | UPV | Comments to the previous version |
| | 6-May-2021 | Jiahui Geng | bY | Updated version |
| | 6-May-2021 | Sandra Morales | UPV | Typos and format review. Minor changes |
| A | 7-May-2021 | Valery Naranjo | UPV | Final version |

## Statement of independence

The work described in this document is genuinely a result of efforts pertaining to the CLARIFY project: any external source is properly referenced.

| | | |
|---|---|---|
| Confirmation by Authors: | Jiahui Geng | bitYoga AS |
| | Russel Wolff | bitYoga AS |
| | Chunming Rong | University of Stavanger |

## Abbreviations

| | |
|---|---|
| API | Application program interface |
| AWS | Amazon Web Services |
| bY | bitYoga AS |
| BSD | Berkeley Software Distribution license |
| BY | Attribution |
| CBIR | Content-Based Image Retrieval |
| CC0 | Creative Commons Public Domain Dedication |
| CDLA | Community Data License Agreement |
| CISH | Chromogenic in situ hybridization |
| CLARIFY | Cloud ARtificial Intelligence For pathology |
| DCC | Digital Curation Centre |
| DID | Decentralized Identity |
| DMP | Data Management Plan |
| DNN | Deep neural network |
| EMC | Erasmus Universitair Medisch Centrum Rotterdam |
| ESFRI | European Strategy Forum on Research Infrastructures |
| ESR | Early Stage Researcher |

| | |
|---|---|
| EU | European Union |
| FAIR | Findable, Accessible, Interoperable and Re-usable |
| GPL | GNU General Public License |
| GDPR | General Data Protection Regulation |
| HR-NMIBC | high-risk non-muscle invasive bladder cancer |
| IDR | Image Data Resource |
| INCLIVA | Fundación Para la Investigación del Hospital Clínico de la Comunitat Valenciana |
| IPR | Intellectual property rights |
| ITR | Image Tool Resource |
| LGPL | GNU Lesser General Public License |
| MDE | Metadata Editor |
| MIT | MIT (Massachusetts Institute of Technology) License |
| NC | Non-commercial |
| ND | Non-derivative works |
| NGS | Next-generation sequencing |
| NMBC | non-muscle invasive bladder cancer |
| OA | Open Access |
| ORE | Open Europe Research |
| ODC | Open Data Commons |
| ODC-ODbL | Open Data Commons Open Database License |
| ODC-BY | Open Data Commons Attribution License |
| ODC-PDDL | Open Data Commons Public Domain Dedication and License |
| PDDL | Public Domain Dedication License |
| OpenAIRE | Open Access Infrastructure for Research in Europe |
| SA | Share-alike |
| SML | Spitzoid melanocytic lesions |
| SUH | Stavanger University Hospital |
| TNBC | Triple negative breast cancers |
| TY | Tyris Software SL |
| UGR | Universidad de Granada |
| UiS | Universitetet i Stavanger |
| UPV | Universitat Politecnica De Valencia |
| UvA | Universiteit van Amsterdam |
| VAE | Virtual Analysis Environment |
| WSI | Whole slide images |

## Table of Contents

# 1 Executive summary

The deliverable D6.4 describes the CLARIFY open access policies and procedures to store and preserve CLARIFY data. Different data sources will be in the frame of this project, like software codes, project documentation, private and sensitive user data, and experimental data, involving relevant scientific research staff from academia, industry, and hospitals. The medical partners are: Helse Stavanger HF (SUH), Erasmus Universitair Medisch Centrum Rotterdam (EMC) and Fundación Para la Investigación del Hospital Clínico de la Comunitat Valenciana (INCLIVA). The industrial partners are: bitYoga AS (bY) and Tyris Software SL (TY). The academic partners are Universiteit van Amsterdam (UvA), Universitat Politècnica de València (UPV), Universitetet i Stavanger (UiS) and Universidad de Granada (UGR). This document includes relevant information about what types of data CLARIFY will generate, how we store and share the data under the guideline of GDPR, and help the efficiency of research and development and disseminate the research results.

Deliverable 6.4 is under the Task T6.3 Data Management & Open Access, within the DoA of the CLARIFY project.

# 2 Introduction

In line with the EU guidelines regarding the Data Management Plan (DMP), the document describes the CLARIFY open access policies and procedures to store and preserve CLARIFY data.

We will explain the basic concepts and procedures of open access in Section 2.1. In Section 2.2, we will explain an important point for open access: license types, which takes into account data ownership and use. Licensing our data and code will protect our copyright and allow others to develop software and algorithms based on our public resources. We will analyze the differences between the various licenses and give a guideline about choosing the proper license. In Section 2.3, we will analyze our data in detail and consider how to keep the data aligned to the FAIR principle.

## 2.1 Open Access

**Open access (OA)** refers to providing online access to scientific information free of charge to the end-user and reusable. Modern research is based on extensive scientific dialogue, and progress is made by improving early work. Barroso[1] emphasizes the central role of knowledge and innovation in promoting growth. Therefore, more comprehensive access to scientific publications and data can help encourage collaboration, avoid duplication of effort, and speed up innovation based on previous research results.

The open access mandate comprises two steps:

- Depositing research data in repositories

  People should deposit research data in a research data repository. These are online research data archives, which can be subject/topic-based, institutional, or centralized. The Open Access Infrastructure for Research in Europe (OpenAIRE) offers information and support to link publications to underlying research data. Some repositories, such as Zenodo (a collaboration between OpenAIRE and CERN), allow researchers to store publications and data simultaneously while providing tools to connect them. Euro-BioImaging is a research infrastructure that provides open access to biological and biomedical imaging technologies, training, and data services. Researchers can freely store, share and access biological images to accelerate scientific discovery.

[1] "Europe 2020 - A European strategy for smart, sustainable and inclusive growth". 2010. https://ec.europa.eu/eu2020/pdf/COMPLET%20EN%20BARROSO%20%20%20007%20-%20Europe%202020%20-%20EN%20version.pdf

- Enabling open access

  The project must then take steps, to the extent possible, to enable third parties to access, mine, exploit, reproduce and disseminate this research data. The most direct and effective way to do this is to attach the appropriate license.

  **Gold open access**

  In general, the critical advantage of gold open access is that publications are freely available from the first publication, which means they can be used immediately. Besides, the open-content license related to gold open access grants a wide range of exploitation rights, coupled with immediate availability, and also achieves a certain degree of visibility. This has a positive impact on the scope of the publication and the frequency of citations.

  **Green open access**

  Green open access does not offer the same legal framework for content licensing. As a result, scientific exploitation is only permitted within the confines of the legal restrictions of copyright law. That means that the author's contract has to be carefully reviewed to enable an article to be reused in a way that fulfills all the legal stipulations. There is no uniform rule governing the open accessibility of publications because different publishing houses impose different embargo periods before making the articles freely available.

  Figure 1 how enabling open access to scientific publication and research data.

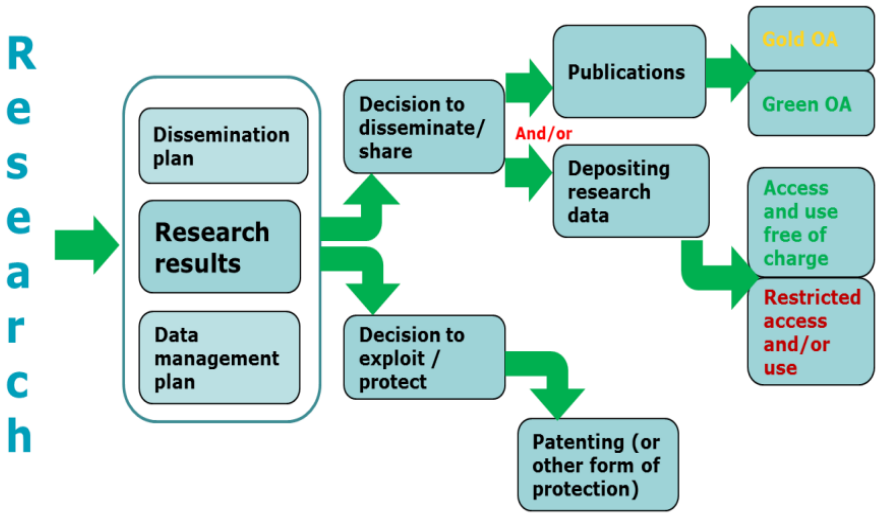


*Figure 1: Open access to scientific publication and research data in the wider context of dissemination and exploitation.*[2]

[2] "H2020 Programme. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020". 2017. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

## 2.2 Licenses

A license is a legal instrument that specifies a standard set of terms and conditions regarding sharing and reusing research data and software. If researchers want to publish their research data in a data repository, they must choose a license to accompany their data. Each data repository has its licensing options. Some repositories require a specific license if people want to deposit their data with them.

### 2.2.1 Software License

An open-source license is a legal license. Through it, the copyright owner explicitly allows users to use, modify, and distribute copyright-protected software for free. Copyright law prohibits sharing by default; in other words, software without a license is equivalent to retaining the copyright. Although it is an open-source code, users can only view the source code and not use it. Once used, the copyright will be infringed. Therefore, if the software is open source, the user must be explicitly granted an open-source license. Free software licenses could be divided into two categories: **permissive licenses**, which allow the software to be reused in any project, even closed source projects; and **copyleft licenses**, which require software derivatives to be licensed under the same terms.

**Permissive licenses** allow unrestricted reuse of code, including the possibility of building commercial software, and the new code will not be disclosed. The user can modify the code to make it a closed source software. It has three essential characteristics：

(1) No restrictions on use. Users can use the code to do whatever they want.

(2) No guarantee. The code quality is not guaranteed, and the user is at his own risk.

(3) Notice requirement. The user must disclose the original author.

**Copyleft license** stipulates that when a software based on previous copyleft works is released to the public, all source code must be distributed. Copyleft ensures that it is not feasible for any organization to attempt to use open-source code in commercial activities. Although the Copyleft license allows the sale of modifications to the open-source code, it removes any restrictions on how the recipient can use it and forces the developer to share all the source code with the recipient. Copyleft grants subsequent users the same rights, making it illegal to keep the source code library secret under the license if the team releases software.

As an antonym of copyright, copyleft means that users can copy at will without permission. However, it comes with preconditions and is more restrictive than a loose license.

- If people distribute the binary format, they must provide the source code.
- The modified source code must be consistent with the license before the modification.
- No other restrictions beyond the original license

The core of the above three conditions is: the modified copyleft code must not be closed source.

## License Selection

There are already more than 100 open source licenses globally (continuously increasing), of which six are the most popular: GPL (General Public License), BSD (Berkeley Software Distribution), MIT (Massachusetts Institute of Technology), Mozilla, Apache, and LGPL (Lesser General Public License). Figure 2 below shows a simple strategy for selecting software open-source licenses.

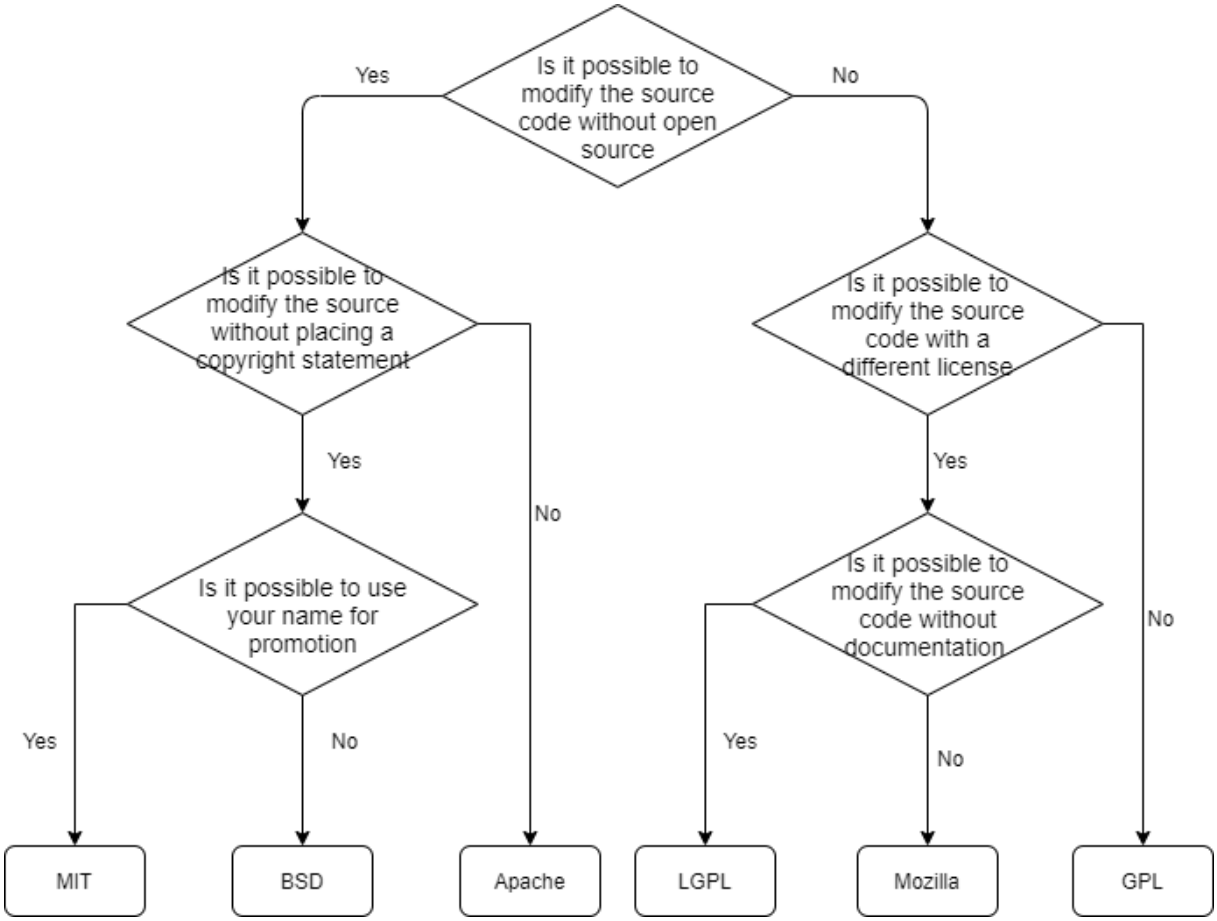


*Figure 2. How to choose a proper license*

### 2.2.2 Data License

There are three primary data licenses framework as follows:

- **Open Data Commons** (ODC).[3]

  Open Data Commons provides a set of legal tools and licenses to help publish, provide and use open data. It contains Open Data Commons Open Database License (ODC-ODbL), Open Data Commons Attribution License (ODC-BY), and Open Data Commons Public Domain Dedication and License (ODC-PDDL).

- **Community Data License Agreement** (CDLA).[4]

  There are two initial CDLA licenses. **CDLA-Sharing** is designed to reflect the principle of reproduction rights in the licensing of data. Generally, if someone shares their data, CDLA-Sharing sets forth terms to ensure that downstream researchers can use and modify that data and are also required to share their changes. **CDLA-Permissive** is similar to a permissive open source license. The data publisher allows anyone to use, modify, and do what they want with no obligation to share any changes or modifications.

- **Creative Commons** (CC).[5]

  The table below describes the difference among CC licenses, including CC Zero or CC0 Public Domain and the combination of the four basic rights[6]: BY (Attribution), SA (ShareAlike), ND (No Derivative Works), NC (Non-Commercial).

| **Licence** | Can I copy & redistribute the work? | Is it required to attribute the author? | Can I use the work commercially? | Am I allowed to adapt the work? | Can I change the licence when redistributing? |
|---|---|---|---|---|---|
| CC0 | Y | N | Y | Y | Y |
| CC BY | Y | Y | Y | Y | Y |
| CC BY-SA | Y | Y | Y | Y | N |
| CC BY-ND | Y | Y | Y | N | Y |

[3] "Open Data Commons: legal tools for open data". https://opendatacommons.org/
[4] "Community Data License Agreement (CDLA)". https://cdla.dev/
[5] "Creative Commons". https://creativecommons.org/
[6] "Creative Commons license – Wikipedia". https://en.wikipedia.org/wiki/Creative_Commons_license

| | | | | | |
|---|---|---|---|---|---|
| CC BY-NC | Y | Y | N | Y | Y |
| CC BY-NC-SA | Y | Y | N | Y | N |
| CC BY-NC-ND | Y | Y | N | N | Y |

*Table I. Creative Commons Licences*[7]

## 2.3 FAIR

In 2016, "FAIR Guiding Principles for Scientific Data Management and Stewardship" was published in Scientific Data. The authors intend to provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital assets. These principles emphasize machine operability (i.e., the ability of computational systems to find, access, interoperate, and reuse data without or with minimal human intervention), as humans increasingly rely on computational support to process data as the volume, complexity, and speed of creation of data increases.

Before we put forward the practice of FAIR, we need to analyze and summarize our data.

### 2.3.1 CLARIFY Data

#### 2.3.1.1 Data Types

There are four primary sources of data in the frame of the project:

- Software codes organized according to the Intellectual Property Rights (IPR) codes in repositories (e.g., Bitbucket, Github) using an Open source software, such as Apache license when possible;
- Project documentation, including deliverables, working documents, scientific papers, posters, videos, photos, and meeting notes, linked to the website and the open science repository according to their access rules;
- Private and sensitive user data, pseudonymized and under user control
- Experimental data, retrieved during the evaluation and validation

**Medical partners**: Whole slide images (WSIs) and connected metadata will be available for the academic partners through a Cloud database.

**Academic partners** will collect the pseudonymized data from the blockchain for clustering and classification. Alternatively, until the blockchain framework is deployed, each partner is

[7] "Licensing your data - CESSDA TRAINING." https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Licensing-your-data

responsible for initializing the pseudonymized data collection into private databases hosted in their ecosystem. A migration process will integrate existing anonymous data repositories into the blockchain once it is deployed. The processed information will serve other applications by providing intelligence and reasoning about the network and decisions therein. During the infrastructure deployment and operation, they will collect system logs.

**Industrial partners**

**BY** will maintain the software code developed for secure sharing of data using the open-source software repository. Additionally, BY also provides RESTFUL APIs to the blockchain for enabling integration with other modules, including frontend applications. BY maintains extensive documentation of both the developed code and APIs for easy knowledge transfer, usage, and integration. A Cloud database will securely store encrypted user-sensitive data and offer separate channels with authorized parties to share relevant keys required for decryption.

**TY** will lead the work on developing Content-Based Image Retrieval (CBIR) methods that will take preprocessed image features as input data and a Computer-aided diagnosis tool. Thus the two primary sources of information will be all the software code developed and the groundtruth data (processed features) used to train the algorithms. Tyris intends to use Bitbucket, a solid commercial web-based repository that allows git-based management for the proprietary code. Open-source options will be considered regarding the code to be accessed by the consortium, such as Github or Gogs an open-source self-hosted repository. The groundtruth data used will be anonymized. In case of sensitive data, it will be stored in secure cloud storage, such as Oracle Cloud, which uses zero-knowledge encryption.

### 2.3.1.2 Data Usage

This section relates to how the data is generated, used or shared in relation to the projects' core objectives.

**Medical partners**

The **SUH** and **INCLIVA** pathologists will provide pseudonymized whole slide images (WSI) and pseudonymized clinical data from patients to other CLARIFY partners. They will annotate the diagnostically and prognostically relevant images within each WSI through an appropriate annotation tool that will be supplied by academic partners and tested and validated. SUH will also further analyze Triple-negative breast cancers (TNBCs) by molecular biological techniques like NGS, methylation, and CISH for microRNAs). According to Dutch GCP guidelines, whole slide images of HR-NMIBC patients stored at **EMC** will be scanned to perform image analysis with CLARIFY developed annotation tools. Histopathological patterns will be correlated to available molecular and clinicopathological data by using CLARIFY developed deep learning algorithms. Molecular and clinicopathological data have already been stored in a pseudo-anonymized database.

### Academic partners

**UvA**: During the project, UvA will produce software code, WebAPI, and system specifications for the distributed data fabric. More specifically, the following assets will be considered in the context of the data management plan:

- Software code will follow Apache license and is available via Github;
- Web API of the services will be documented and accessible via Swagger
- Schema and ontology for semantic alignment will be online accessible with persistent identifier
- Publications, deliverables and internal reports will follow general guidelines of the project

**UPV** will need to access and store data provided by CLARIFY's medical partners (WSI) for artificial-intelligence algorithm development and to accomplish the objectives of WP3. In particular, UPV's main goal is to develop different feature extraction methods from WSI to identify significant patterns and use them for both automatic diagnosis and image retrieval with the focus on Triple negative breast cancer (TNBC) and Spitzoid melanocytic lesions (SML).

**UiS** will need to access and store data provided by CLARIFY's medical partners (WSI) for artificial intelligence algorithm development and accomplish the objectives of WP3. The goal of UiS is to develop methods, algorithms and learn Deep Neural Network (DNN) prediction models for automatic WSI interpretation. UiS will propose algorithms and models for preprocessing, segmentation, and anonymization of WSI and algorithms and models for extracting diagnostic and prognostic information from histological images of non-muscle invasive bladder cancer.

**UGR**'s goal is to research and develop methods for automatic WSI interpretation through advanced image processing techniques and artificial intelligence. UGR will develop algorithms and software code for preprocessing and standardization of WSIs through advanced image processing, automatic significant feature extraction from WSI, and crowdsourcing models for automatic diagnosis and assistance tools for digital pathology applications. UGR will also generate documentation related to the project development.

### Industrial partners

**BY**'s goal is to facilitate and provide a novel cloud-oriented data infrastructure. BY will develop software code for easy deployment of a blockchain and create a secure consortium that can deliver the objectives of this project. To facilitate this, we will propose different algorithms to store, retrieve and share sensitive data securely.

**TY** will work on the development of different strategies for content-based image retrieval. Tyris will make use of the features extracted in Task 3.2 of WP3 to identify the characteristic patterns and search for matches in a features database. Note that Tyris will not necessarily need to access raw data from CLARIFY's medical partners (WSI) since data will be provided as preprocessed image features. Also, Tyris's code will be the base of a diagnosis tool (Task 4.4 and D4.3 of WP4).

### 2.3.2 FAIR Practice

GO FAIR[8] offers some recommendations to guide the fairness of data management, and we will analyze how to implement these recommendations in the context of our data.

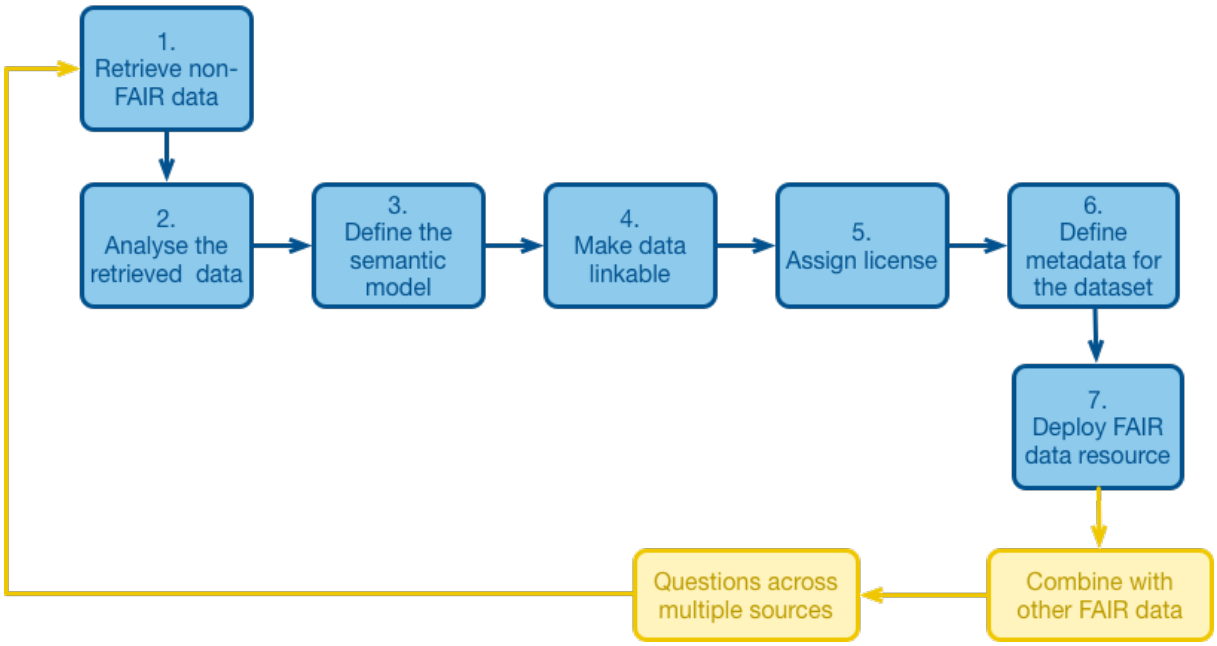


*Figure 3. FAIRification process adopted by GO FAIR[8]*

#### 2.3.2.1 Findability

The first step in reusing data is to find them. Data and metadata should be easy to find for both humans and computers. Machine-readable metadata is essential for the automatic discovery of data sets and services, so this is an important part of the FAIRization process.

**Data and metadata are assigned a globally unique and persistent identifier**

According to FAIR principles, the CLARIFY project will specify and integrate data identification mechanisms during the project architecture redesign. Medical and Academic partners will use pseudonymized data with system-generated serialized Process Identifiers (PID) that cannot hamper privacy by design features. Industrial partners' data does not involve identification issues.

[8] "FAIRification Process - GO FAIR." https://www.go-fair.org/fair-principles/fairification-process/

**Data are described with rich metadata**

When creating FAIR digital resources, metadata can be generous and extensive, including descriptive information about context, quality and condition, or characteristics of the data. Abundant metadata allows computers to automate routine and tedious classification and prioritization tasks that currently require a great deal of researcher attention. The reason behind this principle is that someone should be able to find data based on the information provided by its metadata, even if there is no identifier for the data.

We can refer to the existing open-source tools to make our metadata, manage our data, and make it fit the requirements of the FAIR principle. Tools[9] are developed to support the FAIR process. The Metadata Editor (MDE) is a software tool that allows non-technical users to easily define and publish the required metadata; FAIRifier is a complex application that allows users to merge data and metadata, data licenses data models, and selected ontology and identifiers.

**Data and metadata are registered or indexed in a searchable resource**

Medical and Academic partners have no applicable search keyword requirements. Industrial partners will follow Google search quality guidelines to generate search keywords. We can also achieve this goal with the help of some existing open-source tools.

### 2.3.2.2 Accessibility

Partners allow access to evolving data over the API as soon as the interface is complete and while the services are running. The repository will be available for a specified amount of time. Access to scientific or other relevant documents due to publication would be available after publication for reuse and derivative works.

### 2.3.2.3 Interoperability

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

Humans should be able to exchange and interpret each other's data. This also applies to computers, that is, data that should be read by machines without the need for specialized or special algorithms, translations, or mappings. Interoperability usually means that each computer system has at least an understanding of the data exchange format of other systems. To do this, and to ensure the automatic searchability and interoperability of the data set, the key is to use (1) commonly used controlled vocabulary, ontology, glossary, and (2) a good data model (3) a clear framework to describe and structure data and metadata.

**Data model** accurately and unambiguously describes the meaning of entities and relationships in a dataset in a computer-operable manner. Depending on the dataset, defining an appropriate semantic model can take much effort, even for experienced data modelers. An

[9] "Find FAIR Data tools - Dutch Techcentre for Life Sciences." https://www.dtls.nl/fair-data/find-fair-data-tools/

excellent semantic model should represent the consensus view for a specific purpose in a given domain. Therefore, it is a good practice to search for existing models. Data models usually contain multiple terms from existing ontologies and vocabularies. A glossary is a computer-readable document that captures terms, their URIs, and descriptions. Ontology can be simply described as a vocabulary with a hierarchical structure, meaningful relationships between concepts, and constraints. These conceptual models allow us to classify data models and data items using the provided terminology, concepts, and conceptual structure.

#### 2.3.2.4 Reusability

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

We have explained different licenses for data and software in Section 2.2. Partners can choose suitable licenses for their open access work.

As a recommendation, partners will consider the Creative Commons Attribution license for all publicly available documents. Licensees regulate the copy, distribution, display and performance of the work and allow derivative works only upon crediting the author or licenser. Proprietary nonsensitive data will be licensed under Apache version 2 license.

**Data and metadata are associated with detailed provenance**

In order to facilitate the reuse of our data by other researchers, we need to indicate the source of the data, the history of the data, and how it has been processed, including a description of the workflow that led to the data. We also need to explain whether our data contains the other party's data, how we deal with them, and finally, we have to tell others how to cite our data.

**Data and metadata meet domain-relevant community standards**

If the datasets are similar, it is easier to reuse them: the same type of data, data organized in a standardized way, complete and sustainable file formats, documents (metadata) that follow a standard template, and use a common vocabulary. If there are community standards or best practices for data archiving and sharing, these standards and practices should be followed. For example, many communities have minimum information standards (such as MIAME, MIAPE). FAIR data should at least meet these standards.

# 3 Open Access Tools and Repositories

In this section, we will introduce ORCID and several platforms and repositories for open access.

## 3.1 ORCID

**What is an ORCID?**

ORCID[10] (Open Researcher and Contributor ID) provides a digital identifier that individuals own and control and that distinguishes them from other researchers. They can link their ID to their professional information - relationships, grants, publications, peer reviews, and more. Researchers can share their knowledge with other systems with ORCID, ensuring that all contributions are recognized. At the same time, we will save you time and hassle and reducing the risk of errors.

**Benefits of ORCID**

ORCID allows identification beyond names. Names can be prevalent globally, they can be changed, they can be phonetically translated into other letters, so reliably associating researchers with their research and organizations can be difficult - the unique ORCID iD addresses this.

An ORCID also allows maintaining a digital CV that is constantly updated. A researcher can decide to register which research activities to link to their ID, which organizations to access, what information to make public, what information to share with trusted parties, and what privacy to retain. Individuals control their profiles and can change these settings and permissions at any time.

## 3.2 Open Europe Research

To help improves the quality, efficiency, and responsiveness of research, the European Commission has launched a new Open Access publishing platform Open Research Europe[11] (ORE) which is dedicated to providing all Horizon 2020 and Horizon Europe beneficiaries and their collaborators with an accessible, high-quality venue to publish their research at no cost to themselves.

Once accepted, articles are published promptly after passing a series of pre-publication checks to assess originality, readability, authorship, and compliance with European Open Research policies and ethical guidelines. Peer review, by invited experts recommended by the authors, takes place publicly after publication. Articles will continue to be published regardless of the reviewers' reports.

Authors are encouraged to respond publicly to the peer review reports published with the article and publish a revised version of their article if they wish. Researchers can track their papers,

[10] "ORCID.org." https://orcid.org/
[11] "Open Europe Research ." https://open-research-europe.ec.europa.eu/

drafts and version their articles within the ORE platform[12]. ORE also provide guidelines for data under the FAIR principle, researchers can consult FAIRSharing.org[13] for details of data standards specific to the research topic. Depending on the field of study, there may already be standards in place that will help guide how the data should be structured, formatted, and annotated. OER also provides a list of open repositories as a reference. See Figure 4.

**Health data (restricted access to protect anonymity of participants possible)**

| DATA TYPE | WHERE TO SUBMIT | WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE |
|---|---|---|
| Addiction and HIV data | National Addiction & HIV Data Archive Program | Title, DOI, Route of access |
| Cancer imaging | Cancer Imaging Archive | Title, DOI, Route of access |
| Cancer-related clinical trial data | Project Datasphere | Title, DOI, Route of access |
| Clinical trial data | Vivli | Title, DOI, Route of access |

**Neuroimaging data**

| DATA TYPE | WHERE TO SUBMIT | WHAT TO INCLUDE IN THE DATA AVAILABILITY SECTION OF YOUR ARTICLE |
|---|---|---|
| | | |
| Raw fMRI datasets | OpenfMRI | Title and accession number(s) |
| MRI and PET unthresholded statistical maps | NeuroVault* | Title and URL (which includes a unique data ID) |

*Figure 4: Part of open repositories for different domains.*[14]

However, according to the privacy, sensitivity, and purpose of the data, we will place the data and documents into public repositories or internal repositories.

[12] "Article Guidelines" https://open-research-europe.ec.europa.eu/for-authors/article-guidelines-new-versions/

[13] FAIRsharing" https://fairsharing.org/

[14] "Article Guidelines" https://open-research-europe.ec.europa.eu/for-authors/article-guidelines-new-versions/

## 3.3 Euro-BioImaging

Euro-BioImaging [15] is recognized by the European Strategic Forum for Research Infrastructures (ESFRI) as a landmark European research infrastructure for biological and biomedical imaging technologies. Through Euro-BioImaging, life scientists have access to cutting-edge imaging instruments, expertise, training opportunities and image data services that they may not find at their own institutions or partners. All scientists, regardless of affiliation, area of expertise or field of activity, can benefit from these open services, which are provided to high quality standards by leading imaging facilities across Europe. Euro-BioImaging is a research infrastructure that provides open access to imaging technology, training and data services in biological and biomedical imaging. Euro-BioImaging consists of imaging facilities called nodes, which open their doors to all life science researchers.

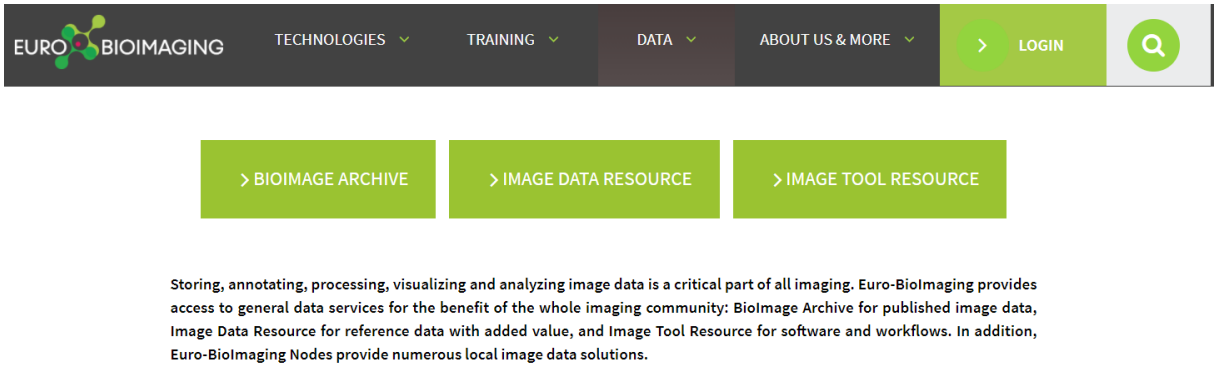


*Figure 5: Interface of Euro-Bioimaging*[16]

When you visit the website of Euro-Bioimaging, you will notice the three primary services it provides, namely technologies, training, and data. You can apply for the technologies by submitting your research proposal and you can take abundant online courses from the first two columns separately. Here we focus on the data column.

There are three options under the data column, the first two Bioimage Archive and Image Data Resource (IDR) are data resource platforms, and the last one is a software resource platform Image Tool Resource (ITR).

**ITR** is an index of open-source bioimage informatics software. With ITR, Euro-BioImaging aims to publish tools that have links to accessible datasets, benchmarking resources, and public workflows to aid in training and reproducibility for image processing and analysis. We can

[15] "Euro Bioimaging." https://www.eurobioimaging.eu/
[16] "Euro Bioimaging Data" https://www.eurobioimaging.eu/data

contact Euro-Bioimaging[17] to include our software or publish software code by submitting a tool to be included in ITR.

When our software is accepted by ITR, other interested researchers have the opportunity to use our software, modify our code according to the license, or cite our work in their papers.

**Tags Key :** F Feature data R Regions of interest (ROI) T Track data N Notebooks O Other data

| Name | Domains | Tags | Data and Results |
|---|---|---|---|
| Icy | bio-imaging<br>medical-imaging | stand-alone<br>manual | |
| Osirix | medical-imaging | volume rendering<br>stand-alone | |
| wnd-charm | bio-imaging<br>medical-imaging | feature extraction<br>image classification<br>automated | |
| XNAT | medical-imaging | stand-alone | |
| CellCognition | screening<br>bio-imaging | | T |
| CellProfiler | screening<br>bio-imaging<br>medical-imaging | feature extraction<br>image classification<br>automated<br>manual | F<br>R<br>N |
| EBImage | screening | | F<br>R<br>N |

*Figure 6. Examples of software already included in ITR.*[18]

**IDR**[19] is a biological image data integration and publishing platform. Users can upload and download image resources as directed. The integration of experimental, imaging, and analytical metadata also provides an opportunity to include new capabilities for data visualization and analysis and add value to original studies and datasets. As with most modern online resources, IDR makes data available through a web user interface and a web-based JSON API. This encourages third parties to use IDR on their websites. To further expand the possibilities of IDR data reuse, we use the open-source tool WND-CHARM to compute comprehensive feature vector sets of IDR image data.

[17] "Euro Bioimaging Contact address." info@eurobioimaging.eu
[18] " Image Tool Resource (ITR) - Euro Bioimaging." https://www.eurobioimaging.eu/data/itr
[19] "IDR: Image Data Resource " , https://idr.openmicroscopy.org/about/

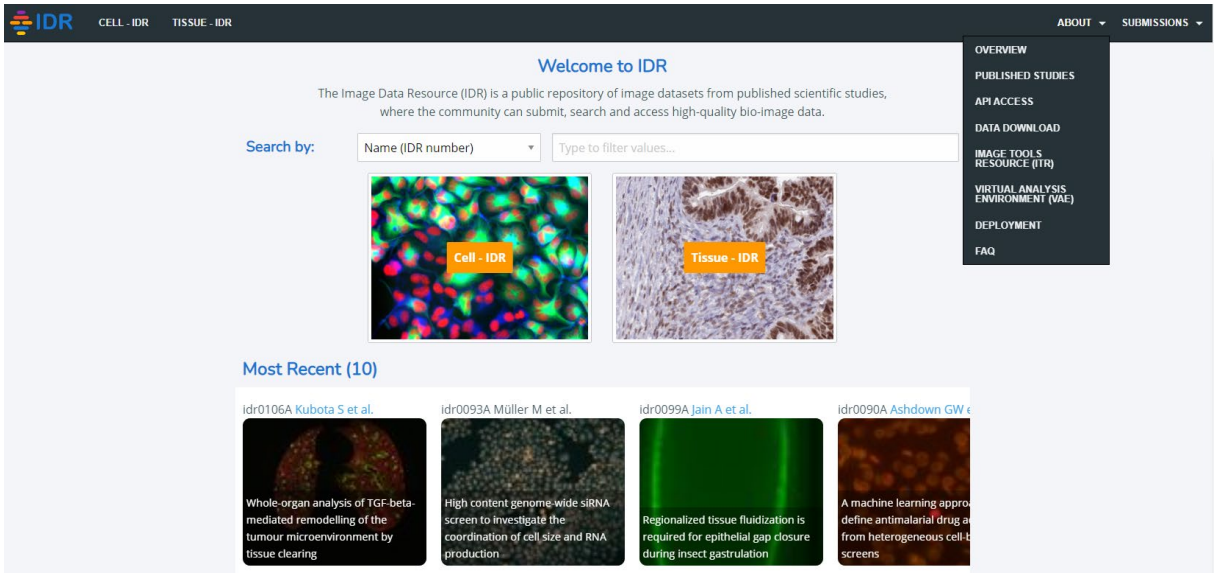


*Figure 7. Overview of webpage of Image Data Resource .*[20]

On the IDR home page, users can browse two types of image resources: cells and tissues. Additional options are hidden behind the "About" option.

The integration of image-based phenotypic and computational features makes IDR an attractive candidate for computational reanalysis. To facilitate access to IDR's terabyte-scale datasets, IDR is connected to a computational resource based on Jupyter Notebook.

The IDR Virtual Analysis Environment (VAE) supports open and reproducible analysis of IDR's data. It is built on JupyterHub and can be used by anyone interested in exploring and mining the diverse and extensive image data and metadata in IDR.

There exist several notebooks in the VAE that can be run as a tutorial.

[20] "IDR: Image Data Resource", https://idr.openmicroscopy.org/about/

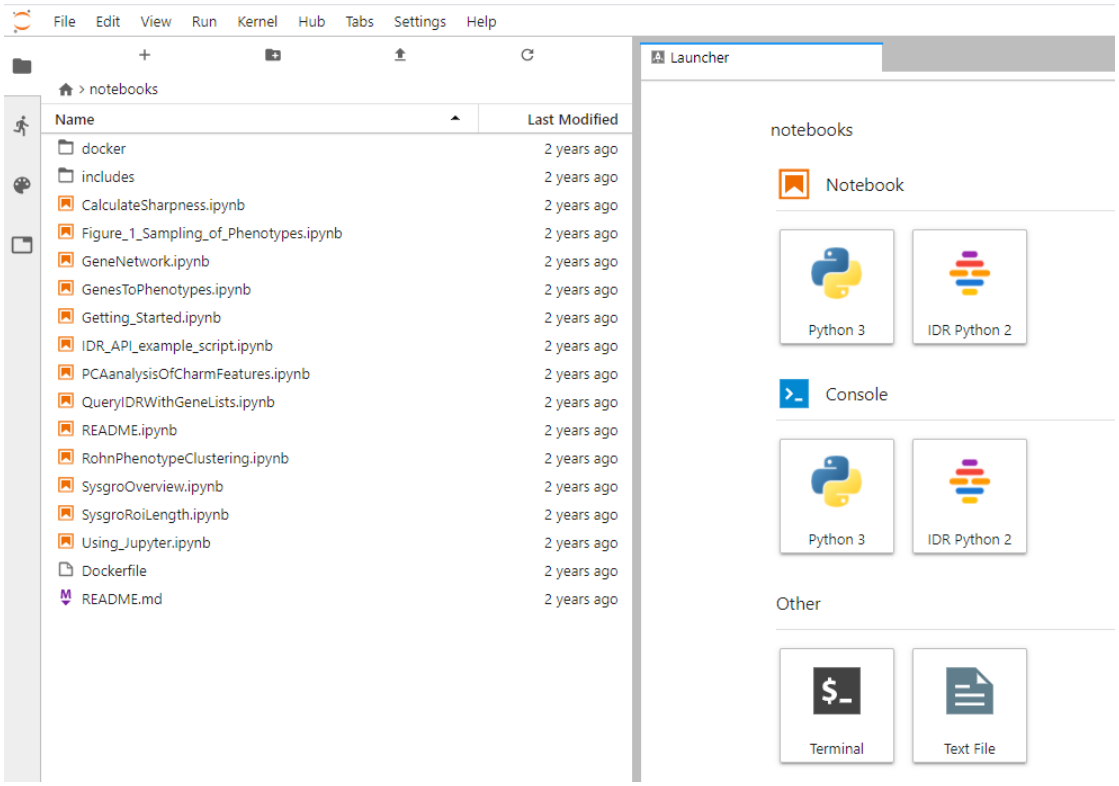


*Figure 8. IDR Virtual Analysis Environment.*[21]

## 3.4 OpenAIRE

OpenAIRE.[22] (Open Access Infrastructure for Research in Europe) stands for a pan-European research information system for the presentation and linking of research results, which aggregates metadata from repositories, archives, scientific journals and other infrastructures.

The Zenodo.[23] research data repository is a product of OpenAIRE. There are now totally 10 different types of resources that can be uploaded. See Figure 9.

[21] "IDR: Image Data Resource Virtual Analysis Environment." https://idr-analysis.openmicroscopy.org/
[22] "OpenAIRE." https://www.openaire.eu/.
[23] "Zenodo." https://zenodo.org/.

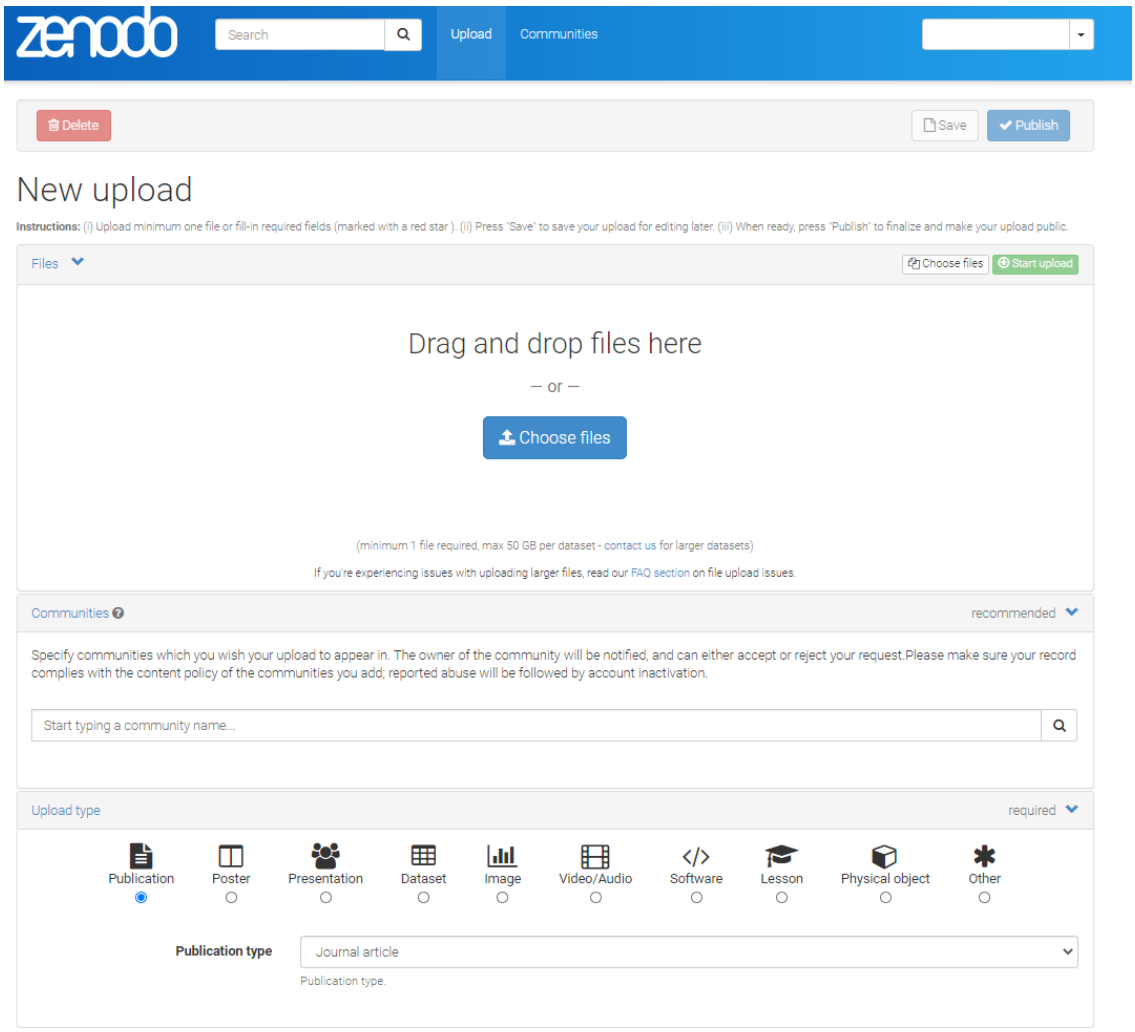


*Figure 9. Uploading page of Zenodo*

Users have a lot of freedom to upload materials. These materials can be publications, videos or pictures, etc. It is worth noting that users can determine their own access rights and licenses for uploading materials.

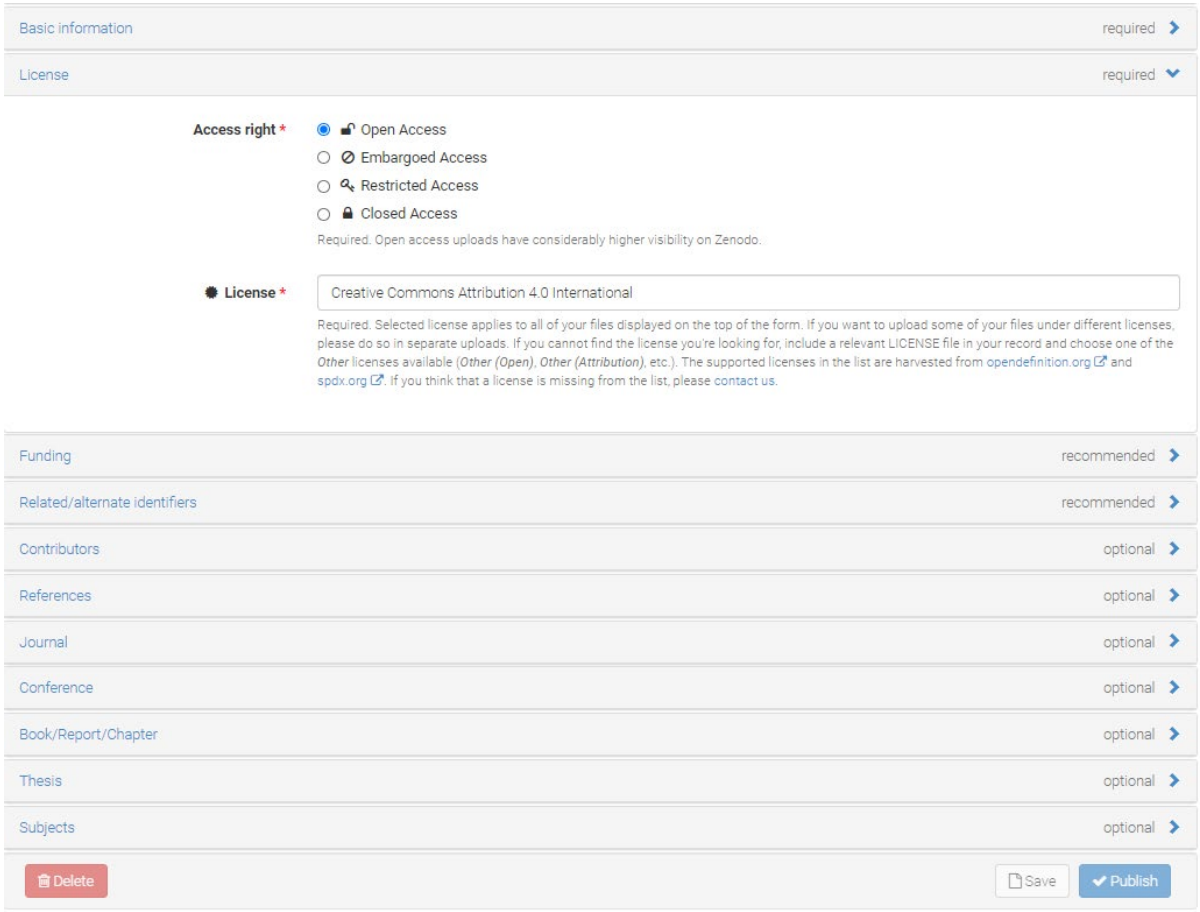


*Figure 10. Metadata about the uploaded resource*

## 3.5 Other Repositories

Sensitive data will only be available to the project researchers using common repositories, internal communication tools and the secure blockchain consortium.

Medical and Academic partners will store the datasets on researchers' **data server,** protected in-line with other confidential data following GDPR, given the low-sensitivity in terms of the Data Protection Act of the anonymized and pseudo-anonymized data. The software repositories and the CLARIFY website will store the other documents. Industrial partners store the source code and the data on a reliable **service provider**, such as Amazon Web Services (AWS). The code will be available through **Github**, using a repository.

# 4 Conclusion

In this document, we describe some essential concepts related to open access, like licenses for software and documents. We will use a combination of public and internal storage to support our research and help us to improve the efficiency of our collaboration and help disseminate the research results. The most relevant CLARIFY outputs will support Gold Open Access to assure maximum impact of the project's results and the rest of the research publications Green Open Access via OpenAire and partner-owned repositories. Before publishing any result, we will analyse what is the most convenient open access practice from those described in this deliverable subject to ethical restrictions that may exist since CLARIFY research involves the use of human subject data. We will organize the CLARIFY's outputs to be findable, accessible, interoperable and reusable (FAIR principle).