This project is co-financed by the European Union

Grant Agreement No.: 824603
Call: H2020-SwafS-2018-1
Type of action: RIA
Starting date: 1/02/2019

ACTION

OPEN DATA PORTAL

Coordinator: Esteban González
Quality reviewer: Gloria Re Calegari (CEFRIEL)

Deliverable nature	DEMONSTRATOR
Dissemination level	PUBLIC
Work package and Task	WP4
Contractual delivery date	30/04/2020
Actual delivery date	

Authors

Author name	**Organization**	**E-Mail**
Esteban González Guardia	**UPM**	**egonzalez@fi.upm.es**
Óscar Corcho	**UPM**	**ocorcho@fi.upm.es**

Abstract	This deliverable describes the first version of the Data Portal of our project. Our Data Portal not only will be a data repository, it will be a place where we publish all the outputs of our project, including the outputs of our pilots. The solution adopted is based on Zenodo and a website developed over Zenodo to access our resources.
Keywords	Data management, data portal, citizen science

Disclaimer

The information, documentation and figures available in this deliverable, is written by the ACTION project consortium under EC grant agreement 824603 and does not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

How to quote this document

Gonzalez, E. and Corcho, O. (2020), ACTION Data Portal

TABLE OF CONTENTS

EXECUTIVE SUMMARY

One of the objectives of ACTION is: "*Create a digital infrastructure to help citizen scientists easily set up and manage projects in all their online and offline manifestations, manage and share their data openly, and comply withRRI(Responsible Research and Innovation) practices*". In this context, ACTION will provide to the projects (pilots) a repository to publish their data.

The objective of this deliverable/document is to describe and to implement a technological solution to deploy a data portal. FAIR principles (Findable, Accessible, Interoperable and Re-usable) have inspired us to design and to develop this solution.

Our data portal is not only a repository of data, where all users upload datasets, it is a repository for all project´s outputs that can be used for researching. This is aligned with future deliverables such as the research object catalogue. A research object is a collection of resources that contains information about your research. In a research object, you can include papers, slides, data and software used in your research.

To develop our data portal we have followed a novel approach. We have used the well-known platform Zenodo as a repository. The projects and their pilots will be able to upload and to publish their digital resources in the communities created for this purpose. Pilots will be the owners of these communities and they can manage them. A guideline (in Annex I) has been created and a webinar was given to train them. This is a way to form newcomers in the use of open science tools.

At the same time we have deployed a simple website to display the resources of our project (the community of ACTION and their pilots). This website is a tool to visualize the results published by the communities of ACTION. This offers a unified vision of ACTION and its pilots, even when the relation between them has officially finished.

The solution adopted is simple and scalable when the number of pilots and resources deposited will grow as time passes. Using Zenodo as a repository, it helps us to ensure the long-term sustainability of the projects.

The data portal is available here: https://data.actionproject.eu/

1 INTRODUCTION

The objective of this deliverable is to design and to deploy an open data portal that allows citizens to access the outcomes generated in ACTION and all of their pilots.

According to the European Commission[1] , an Open Data Portal (ODP) is a *web-based interface designed to make it easier to find re-usable information*. To make this information re-usable, it must be published with proper licenses and with a set of appropriate metadata.

FAIR principles[2] have guided us when designing the portal. FAIR principles are:

- **Findable**: Metadata and data must be easy to use both for humans and computers. Metadata must be described properly and it must be indexed in a searchable infrastructure.
- **Accessible**: Metadata and data must be accessible with a persistent identifier using any open communication protocol.
- **Interoperable**: Metadata and data must be integrated with other data. For that, vocabularies, ontologies and references are recommended mechanisms to describe the data.
- **Reusable**: In order to be replicated, it must be registered with a clear data usage license, as well as the origin of the data. Also, it is highly recommended to follow the domain standards.

In section 2, the solution adopted to implement and to deploy an ODPI is described.

In section 3, the structure of the ODP website is analyzed.

In Annex I, we have included a guideline for uploading resources to our ODP.

[1] https://ec.europa.eu/digital-single-market/en/open-data-portals
[2] https://www.go-fair.org/fair-principles/

2. Data Portals

A data portal is not only a data repository but also allow users to discover, analyse and visualize them using a set of tools.

Some of its most relevant features are:

- Addition of metadata.
- Integration with other data repositories.
- Notification of the inclusion of new datasets.
- Integration with tools for visualizing and manipulating data.

There are some solutions to implement open data portals like CKAN, Socrata and OpenData Soft.

CKAN[3] is an open source solution for managing data portals. With CKAN, you can share and publish your data, making it findable and accessible. Also, CKAN has a set of plugins to extend its functionality, including the generations of maps and charts. Plus, you can modify the user interfaces to adapt it to your project´s branding.

According to wikipedia, *Zenodo[4] is a general-purpose open-access repository developed under the European OpenAIRE program and operated by CERN*. Metadata follows the FAIR principles and a Digital Object Identifier (DOI) is generated for each resource (make it findable). Zenodo is based on the framework Invenio[5] and can be deployed in your own servers, as it can be done with CKAN. In the case of Zenodo, they provide storage capacity for communities in their servers, hosted in CERN facilities.

In the following table, a comparative between the two systems can be found:

Feature	CKAN	Zenodo
Storage capacity	ACTION Server	Unlimited
High Availability	No (1)	Yes
Funds	ACTION	Unlimited (2)
Usability to upload datasets	Easy	Easy
Internal structure	Compact	Scattered (3)
Plugins	Yes	No
Search tools	Yes	Yes

[3] https://ckan.org/
[4] https://zenodo.org/
[5] https://inveniosoftware.org/

API endpoint	Yes	Yes (4)
Private datasets	Yes	Restricted access option

Table 1: Comparative between CKAN and Zenodo

(1) Depending on your infrastructure
(2) Zenodo has funding for 20 years
(3) All resources are listed in the same page without not internal structure. CKAN is based in organization, and you can create sub-organizations
(4) The API only has methods to upload resources. Nevertheless, resources are exposed through the protocol OAI-PMH

Zenodo, at this moment, can provide us with unlimited storage capacity in their servers. In this case, it is not necessary to deploy the system in our servers, avoiding the limitation of capacity, and the maintenance of the platform (and its scalability who relies in external systems)

In the case of the sustainability of the project, has funding for the next years provided by the OpenAire project, which offers an excellent solution for persisting the data in the future. Plus, a Digital Object Identifier (DOI) can be generated, which offers a global and persistent identifier. This identifier can be used to cite our data in scientific papers or other digital objects.

New functionalities can be adopted in CKAN through the use of plugins. There are software extensions that can be added to the core of CKAN. For example, you can adapt the templates of your metadata to your necessities or store the data in databases.

Regarding the access to the platform, CKAN implements an API to create, modify and delete datasets, as well as resources. If you have the extension of the database deployed (datastore plugin), you can access the data (not only to the metadata) stored in the database. In the case of Zenodo, the API[6] only can be used to deposit files, and to create/update/delete the metadata through the POST/PUT methods. Nevertheless, Zenodo implements an OAI-PMH endpoint, so you can consult the metadata of each dataset present in the system.

Both systems have the possibility to deposit private datasets. In the case of Zenodo, an embargo period can be established to make published resources after a date.

Regarding the internal organization, CKAN provides more functionalities such as the creation of hierarchical organizations and groups of datasets in the portal. In Zenodo, only can be created communities which are not related to each other.

[6] https://developers.zenodo.org/

3. Architecture of the system

In Table 1, we have analyzed the functionalities of two platforms to build/deploy data repositories/portals, CKAN and Zenodo.

Although CKAN is a very popular solution for data portals, the infrastructure depends on the resources of your project (present and future), which has a big impact on the sustainability of the solution (storage capacity, hosting infrastructure, updates).

As a H2020 project, the indexing of our outputs in OpenAire is an important task to carry out. This makes our outputs more visible to the community, as well as they will be visible in the participant portal of the EU Commission for reporting projects. Zenodo can provide us with this functionality.

A requirement of our system is to make our outputs or resources of our project citables. Zenodo can generate a DOI for each resource uploaded which is an important feature to take into account.

One of the missions of ACTION is to be an accelerator of citizen science projects (pilots). In the case of the data management, we want to teach them how to manage their data and publish it following the GO FAIR principles. Zenodo is perfectly aligned with this initiative which can be an excellent choice to deploy our data portal.

For these reasons, we have chosen Zenodo as the solution adopted to build our data portal. But to build our solution, we have to solve some problems first.

Zenodo has some limitations (API requests, graphic interface, internal structure) that have to be improved in our solution. Thus, we have developed and integrated three modules:

- A website based on HTML + Javascript to interact with the information deposit on Zenodo. It is like a frontend of Zenodo. It allows us to represent and to organize the resources based on the structure of our project (pilots). Also, some statistics can be integrated in the platform.
- An Elastic Search engine to index all the information related to our project present on Zenodo. Also, we have added some statistics such as number of visualizations, downloads, etc … This will be extended in future releases with new metrics.
- A harvester (ODP_ZenodoHarvester) to get the metadata from Zenodo and to index them in the ElasticSearch engine. This information is updated every hour, in order to avoid API limitations in Zenodo. We use the OAI-PMH protocol from Zenodo to do it.

In the following diagram, an overview of the architecture can be found.

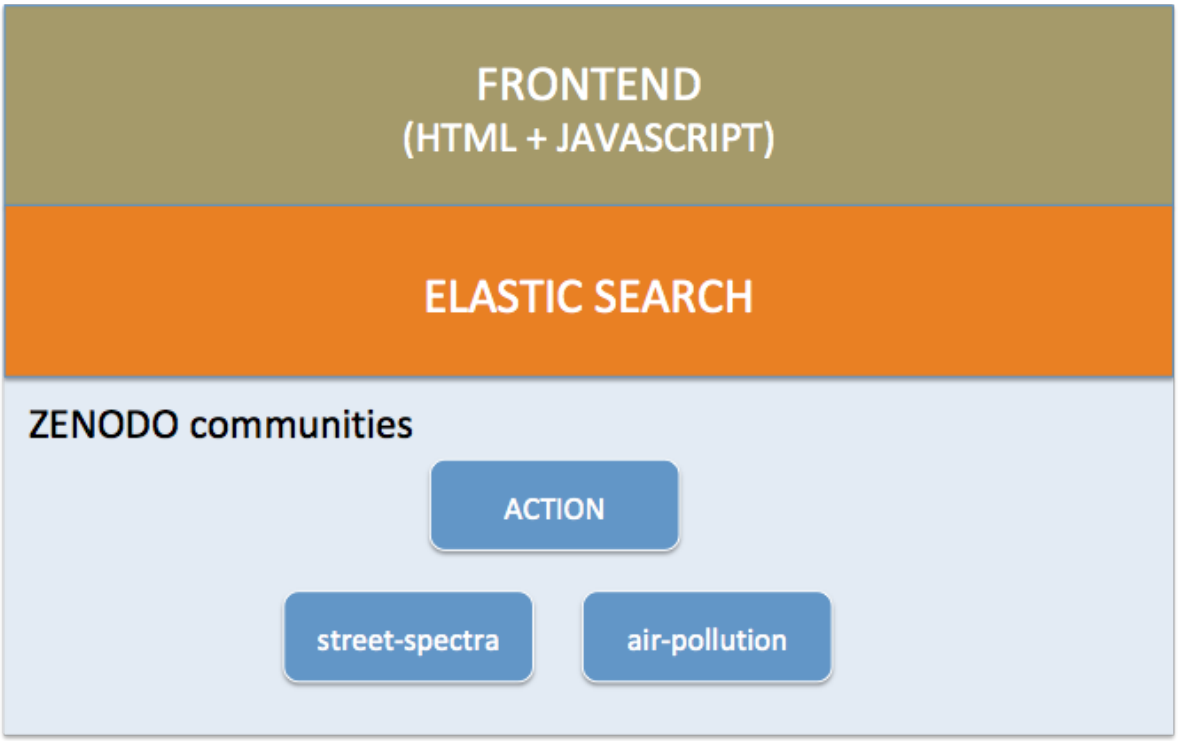

Fig 1. Open Data Portal Architecture

As was mentioned previously, Zenodo is used as a resources repository. Zenodo is organized in communities. In our case, we have created one community for each ACTION pilot, plus the community of the project ACTION.

Each pilot will upload their resources in their own community (see Annex I), linking them with the community of ACTION and OpenAire. In this way, resources can be organized under the umbrella of ACTION. At the same, the future independence of the pilots will be preserved without losing the community formed in ACTION. This will be an important factor for the sustainability of the projects.

The code developed for the project can be found here:

- Frontend (website) -> https://github.com/actionprojecteu/ODP_frontend
- OpenDataHarvester[7] -> https://github.com/actionprojecteu/ODP_ZenodoHarvester

[7] This module is responsible for loading the data of the Zenodo communities in the Elastic Search.

4. INTERFACE DESIGN

As was mentioned in the previous section, the interface of our data portal is built in HTML and Javascript. The objective of this webpage is to offer a common access interface of the different ACTION communities built in Zenodo. Also, dashboards and statistics results may be added in the future.

As a requirement, this interface should be responsive, adapting their content and design to the device used to visualize it.

The website is based on a free template named JobFinder. This template has a creative commons license with attribution.

The website (https://data.actionproject.eu/) can be divided in three sections:

- Searching tool.
- Projects.
- Results.

3.1 Searching tool

In the first section (see Fig 2), users will be able to search a resource based on a pattern over the metadata used in Zenodo. The search is executed on all the communities related with ACTION. Also, you can make searches based on the category of the resources (software, datasets, documents, etc ….). Results will be displayed on results section.

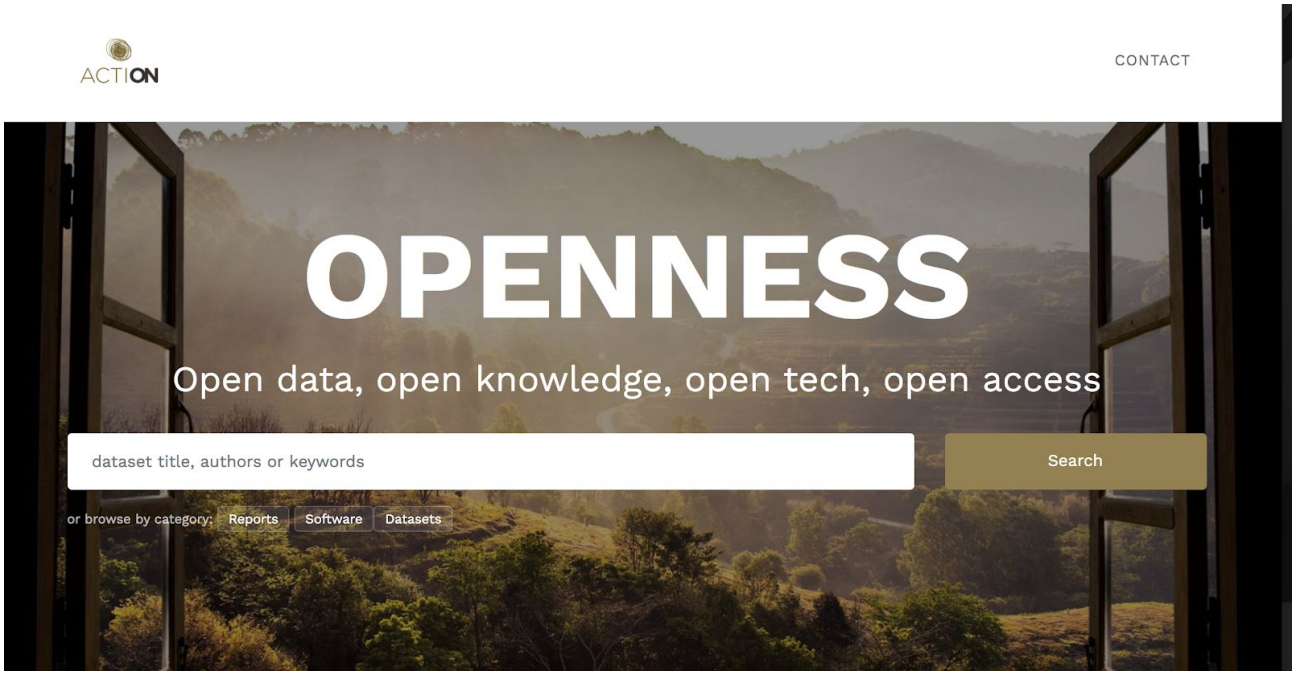

Fig 2. Search section

3.2 Project section

In the next section, users can see cards with the different projects/pilots of ACTION. If a user clicks on one of them, will see the resources of this pilot (resources of its community in Zenodo). Results will show up in the results section.

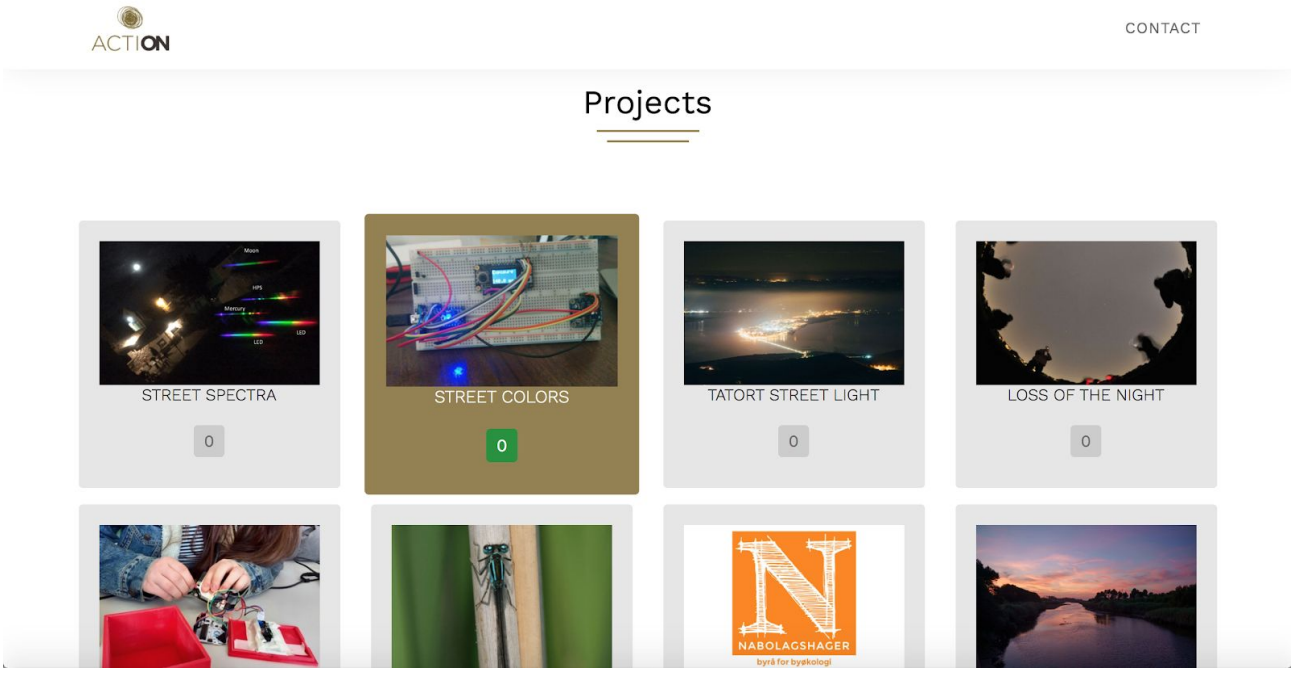

Fig 3. Projects section

3.3 Results section

The last section (see Fig. 4) is where the results are displayed. Users can see the following information related with the resource:

- Title
- Authors
- Project
- License used
- Number of visualizations in Zenodo
- Number of downloads in Zenodo

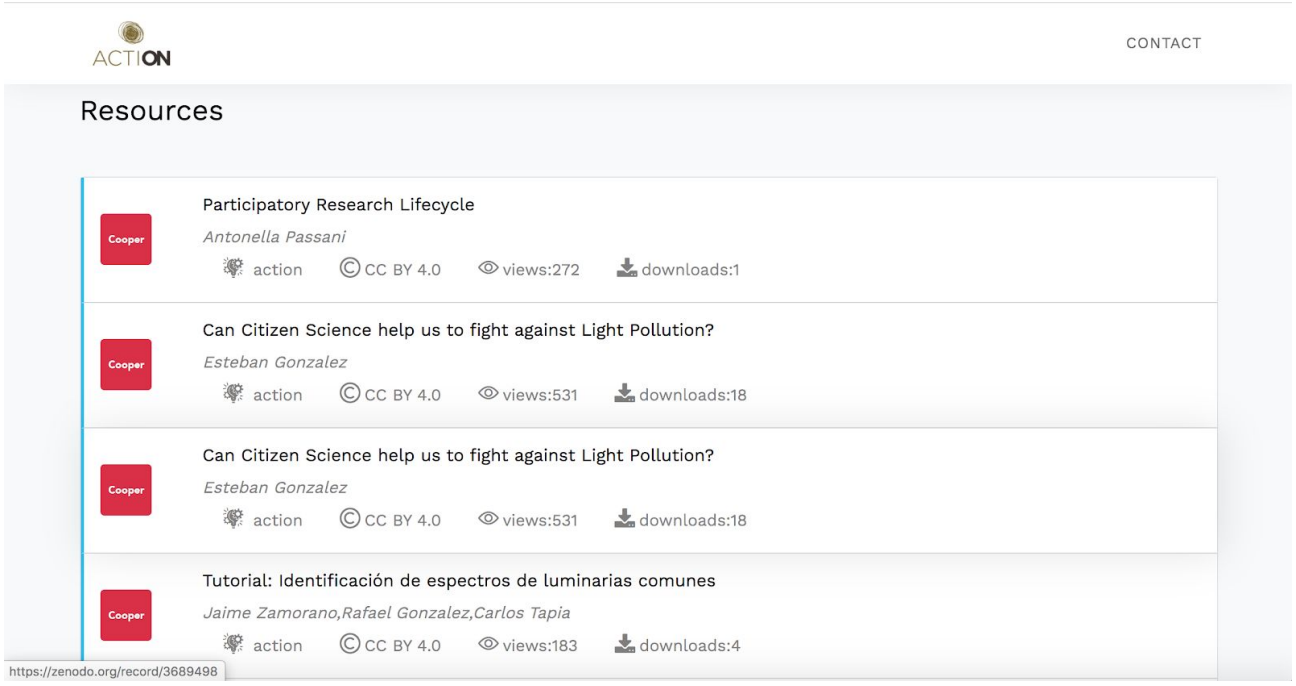

Fig 4. Results section

4 CONCLUSIONS

The objective of this deliverable is to deploy a platform where the data of ACTION can be deposited and published. After the study of some solutions present in the market, we have chosen a solution based on Zenodo.

Four aspects have been considered to make the decision: i) alignment with OpenAire project, ii) the sustainability of the project, iii) to give pilots a *science facet* using a researchers' repository, iv) each resource will have a Digital Object Identifier and v) integration with Github (to deposit software automatically).

Pilots of ACTION will have their own community in Zenodo where they can upload their outcomes. Plus, the community will belong to the pilot, being maintained in the future beyond the project. The Open Data Portal offers a unified vision of ACTION and its pilots, even when the relation between them has officially finished.

The solution adopted is easy and scalable when the number of pilots and resources deposited will grow as time passes. Plus, this solution allows us not only to publish data, but also other resources such as presentations, software, posters, deliverables, etc ...

Looking ahead, the use of Zenodo will help us in the development of the future Research Objects catalogue deliverable.

ANNEX I

Guideline for publishing resources in Zenodo

Zenodo is a platform hosted by CERN oriented to deposit, archive and publish scientific material with open licenses, it means, open to the public. Furthermore, a DOI (Digital Object Identifier) can be generated for each resource, which allows it to be easily cited and shared.

First, you have to think about the resources that you want to publish on this platform. It includes:

- Any relevant document (papers, deliverables, guidelines, etc …).
- Presentations.
- Images or Videos.
- Dissemination material (digital)
- Datasets

Once you have identified the material you want to store, you can start to use the platform. For that, follow the next steps:

1. You have to create an account on the platform. Goto the Zenodo website (https://zenodo.org/) and click on the Sign Up button. You will have to provide an email address, a username and a password or using another alternative authentication method.
2. Once you are logged into the platform, you will have to create a community for your project. Click on the Community option in the menu and on the New button in the Community section.

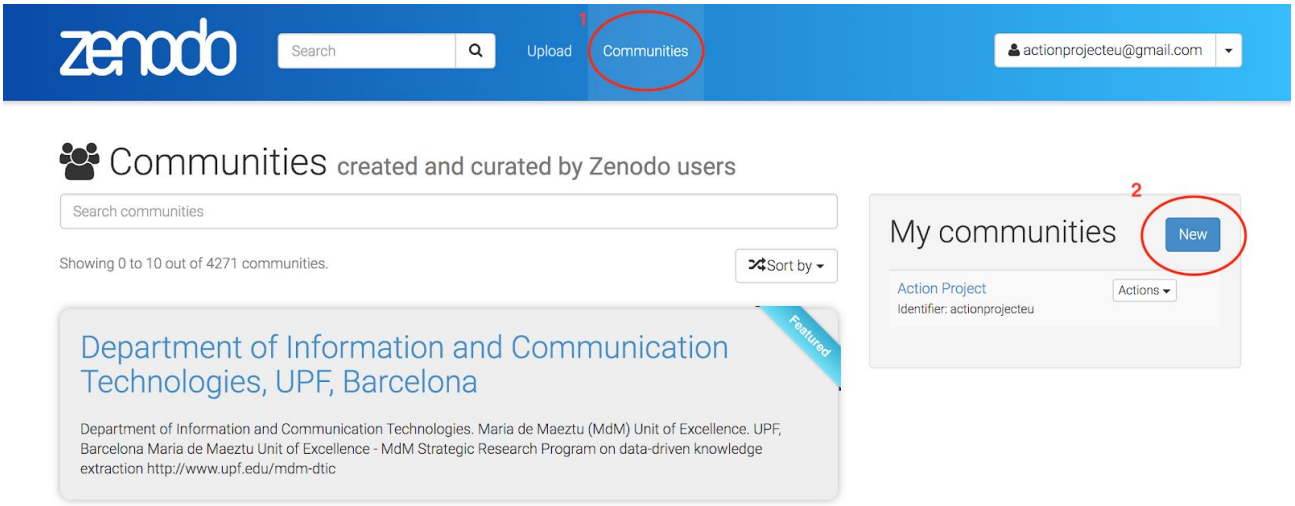

Fig 5. Community screen

3. We should add some extra information for configuring the community. It includes an identifier (it must be unique and it will be part of the url), a title, description and a logo of the community.
4. Once the community is created, we can start to upload our resources. For it, you have to click in the **Upload** option in the menu and later in the **New Upload** button. x

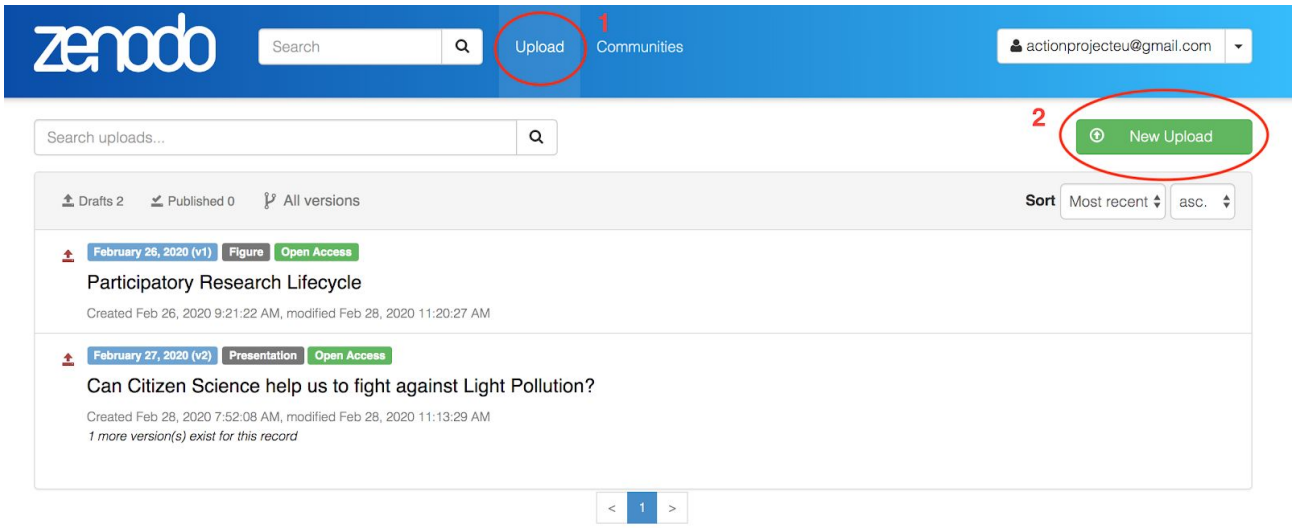

Fig 6. Upload screen

5. A list of fields to add the metadata will be displayed.
6. The first information you will have to complete is the file that you want to deposit. You can drag & drop from the files explorer or select one from your computer. Once the file has been selected, remember to upload it clicking in the Start upload button (red oval). The size limit for the file is 50GB

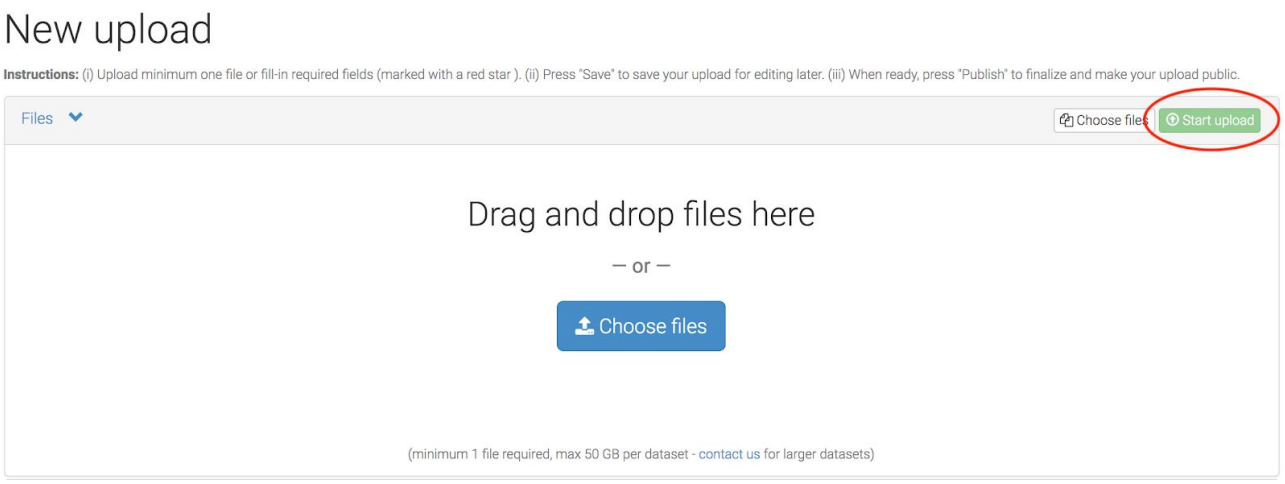

Fig 7. Upload process

7. The second field is the communities where the resource will be indexed. The field works like an autocomplete field, where you can search the different communities inside of Zenodo. In this option, you should select the following communities.
    a. Your community
    b. OpenAire: It is an European project for supporting Open Science
    c. Actionprojecteu: It is our project. Resources will be indexed in our Open Data Portal.

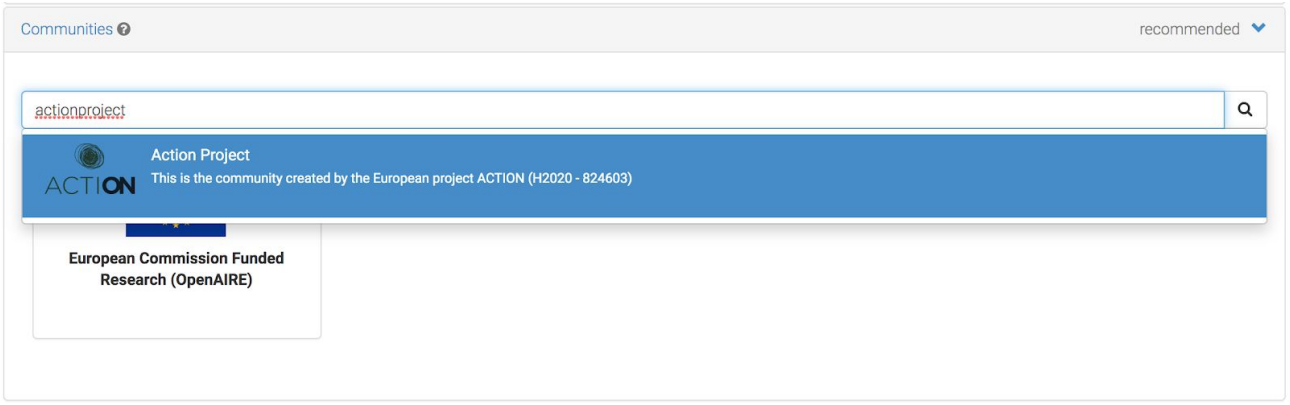

Fig 8. Selection of your communities

8. In the next metadata field, you will have to define the type of resource you are going to upload. The options are: Publication, Poster, Presentation, Dataset, Image, Video, Software, Lesson and Others. Some recommendations:
    a. If you upload a presentation in PDF format, you can visualize it in Zenodo, otherwise, you will have to download it first to view it.
    b. Datasets are better to upload them in JSON or CSV format to be viewed from the Zenodo.
    c. If you have a collection of images, it is better to generate a ZIP file and upload it. In this case, you will have only one DOI for all images.
    d. In the case of software, we strongly recommend generating a ZIP file with the whole project. Zenodo can be linked with your Github account to publish in Zenodo automatically your Github releases.

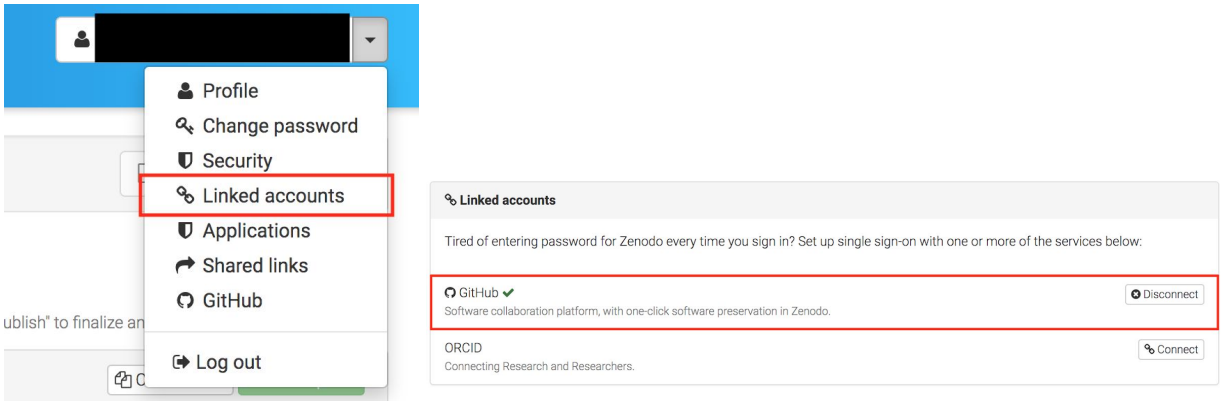

Fig 9. Github integration

9. Next metadata fields are grouped in the tab Basic Information. The first metadata is the Digital Object Identifier or DOI. This is one of the main features that Zenodo provides us. It is a persistent identifier of your resource that can be used to be cited in papers or any other publications. This DOI will be assigned automatically when you publish the resource. Remember that once assigned, it won't be changed. If you already have a DOI, you can explicitly indicate it in this field.
10. The next four fields add basic information to the dataset (publication date, title, authors and description). In the case of the title, try to be as descriptive as possible. Try not to use

excessive long titles. In the case of authors you will have to add the main authors of the resource. Remember to include the ORCID of the author (if exists). If you have many authors, think if some of them are collaborators (there is an special metadata field for this purpose)

11. The next field is the version of your resource. In the software domain, we usually use the SemVer notation with three digits separated by dots. In the rest of resources, a simple number will be more than enough. Our recommendation is to add the version number to the name of the file, and use the mechanism of Zenodo for incorporating new versions. In the upload section, you can upload a new version of the resource. Zenodo will incorporate this new version to the initial DOI, and it will generate a new DOI for this version.
12. Select the language of the resource. Remember to use the ISO 639.2 codes (https://www.loc.gov/standards/iso639-2/php/code_list.php)
13. At this moment, you will have to define the keywords that describe your dataset. Remember to use some generic words (domain related) to get more publicity to your dataset and other more specific (to describe it). These keywords will be used to search your resources. Note that these tags don't belong to a controlled vocabulary, they are free text.
14. In additional notes you can add additional information such as how you obtain this data, a process to assure the quality of them, etc …
15. Next fields determine the license of your resource. You have the following options:
    a. **Open Access**: Your resource will be public in the repository. We strongly recommend the use of the **Creative Commons Attribution 4.0 International** license. With this license, everybody will be able to use your resource, mix it, and make commercial use of it, always indicating the origin. Plus, any copy or derived product will have the same license that the original.
    b. **Embargoed Access**: In this case, the resource will not be available until a specific date, that you will have to define in another field. After this date, the resource will be released with the license selected.
    c. **Restricted Access**: In this case, users can find the resource but they will have to request for access under some conditions (that you will have to define in another field).
    d. **Closed Access**: The resource won't be visible.
16. In the next field, you will have to add your funding sources. In our case, remember to add ACTION (number 824603). You can add more funding sources if it is necessary.

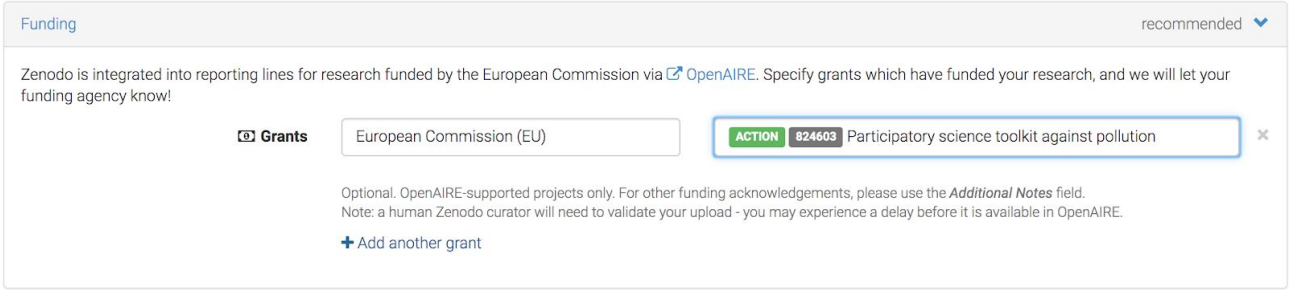

Fig 10. Funding field

17. Next fields (related identifiers) allows us to connect our resources with other resources. For example, if our dataset has been generated or processed with a software, you can indicate

it here. Or if a paper is using an specific dataset, you can indicate here too. It will help you in the future to discover the relation among all your research products. Remember to select the type of the relation (cites, is cited by, etc ..) and the type of the related resource (Image, Book, Dataset, etc ...)

18. Next field is contributors. In this part, you can add more people who help you to create the resource. In the case of CS projects, this is the place where we can add the volunteers and recognize their work.
19. The rest of the fields are related to the type of publication where you are going to use (journal, book, conference, etc …).
20. Finally, **subjects** field is used in case you want to use a vocabulary or a taxonomy to describe the resource (mainly datasets).This is a good practice to share your data.