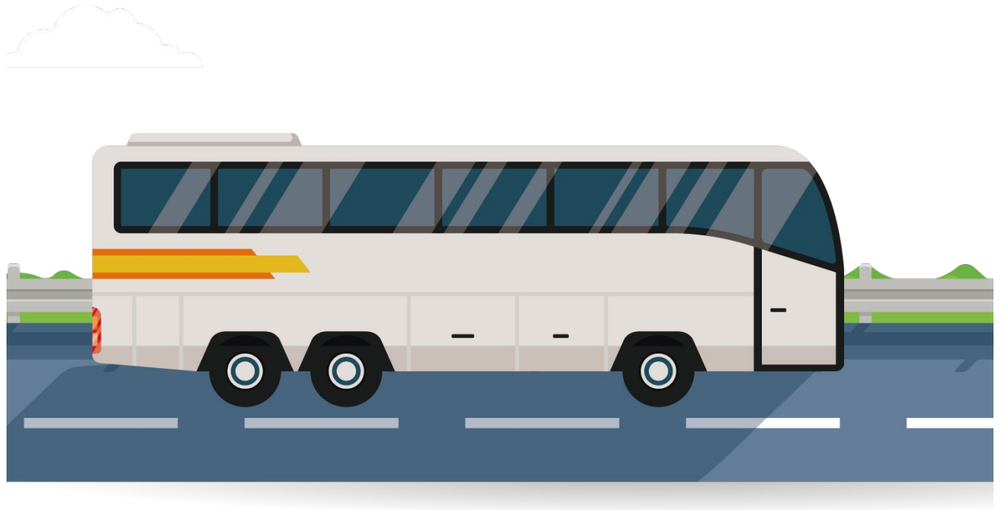
Pablo Panero
pablo.panero@cern.ch

zenodo

Research. Shared.

You are a researcher…



Credit: Dave Hill, licensed under Creative Commons Attribution-NonCommercial-ShareAlike 2.0 Generic.
https://www.flickr.com/photos/dmh650/4031607067/in/gallery-wlef70-72157633022909105/

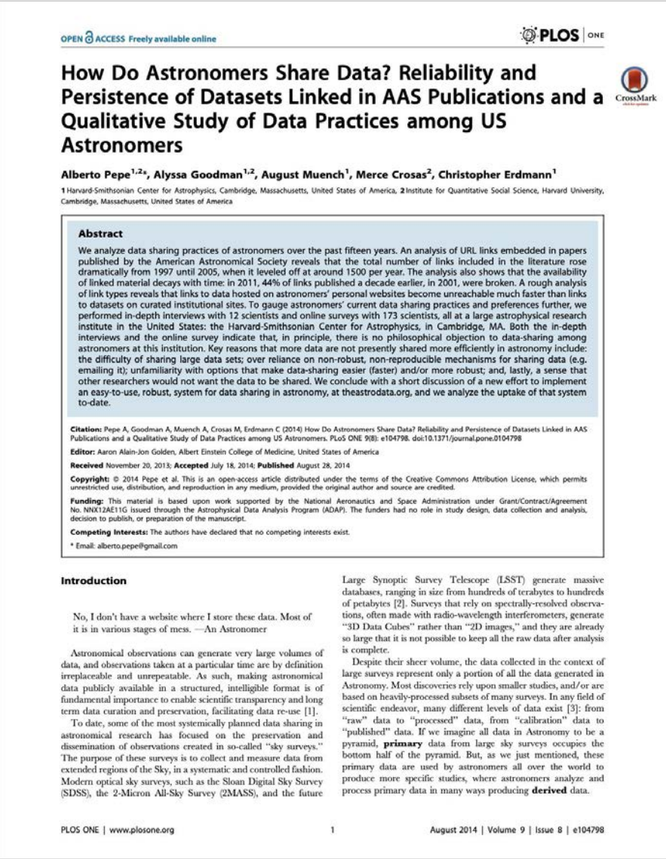
Credit: By Bryan Tong Minh / CC-BY-2.5
(http://commons.wikimedia.org/wiki/File:Brand_bouwkunde_-_TU_Delft_-_13_Mei_2008.jpg)

Credit: By Brian Herzog / licensed under Creative Commons Attribution-NonCommercial-ShareAlike 2.0
— https://www.flickr.com/photos/herzogbr/6756173595/in/gallery-wlef70-72157633022909105/

Source: https://github.com/microsoft/vscode/issues/32405

The result?

50% of the links in papers are **inaccessible** after 10 years



OPEN ACCESS Freely available online

PLOS ONE

How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers

Alberto Pepe[1,2]*, Alyssa Goodman[1,2], August Muench[1], Merce Crosas[2], Christopher Erdmann[1]

1 Harvard-Smithsonian Center for Astrophysics, Cambridge, Massachusetts, United States of America, 2 Institute for Quantitative Social Science, Harvard University, Cambridge, Massachusetts, United States of America

Abstract

We analyze data sharing practices of astronomers over the past fifteen years. An analysis of URL links embedded in papers published by the American Astronomical Society reveals that the total number of links included in the literature rose dramatically from 1997 until 2005, when it leveled off at around 1500 per year. The analysis also shows that the availability of linked material decays with time: in 2011, 44% of links published a decade earlier, in 2001, were broken. A rough analysis of link types reveals that links to data hosted on astronomers' personal websites become unreachable much faster than links to datasets on curated institutional sites. To gauge astronomers' current data sharing practices and preferences further, we performed in-depth interviews with 12 scientists and online surveys with 173 scientists, all at a large astrophysical research institute in the United States: the Harvard-Smithsonian Center for Astrophysics, in Cambridge, MA. Both the in-depth interviews and the online survey indicate that, in principle, there is no philosophical objection to data-sharing among astronomers at this institution. Key reasons that more data are not presently shared more efficiently in astronomy include: the difficulty of sharing large data sets; over reliance on non-robust, non-reproducible mechanisms for sharing data (e.g. emailing it); unfamiliarity with options that make data-sharing easier (faster) and/or more robust; and, lastly, a sense that other researchers would not want the data to be shared. We conclude with a short discussion of a new effort to implement an easy-to-use, robust, system for data sharing in astronomy, at theastrodata.org, and we analyze the uptake of that system to-date.

Citation: Pepe A, Goodman A, Muench A, Crosas M, Erdmann C (2014) How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLoS ONE 9(8): e104798. doi:10.1371/journal.pone.0104798

Editor: Aaron Alain-Jon Golden, Albert Einstein College of Medicine, United States of America

Received November 20, 2013; **Accepted** July 18, 2014; **Published** August 28, 2014

Copyright: © 2014 Pepe et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This material is based upon work supported by the National Aeronautics and Space Administration under Grant/Contract/Agreement No. NNX12AE11G issued through the Astrophysical Data Analysis Program (ADAP). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: alberto.pepe@gmail.com

Introduction

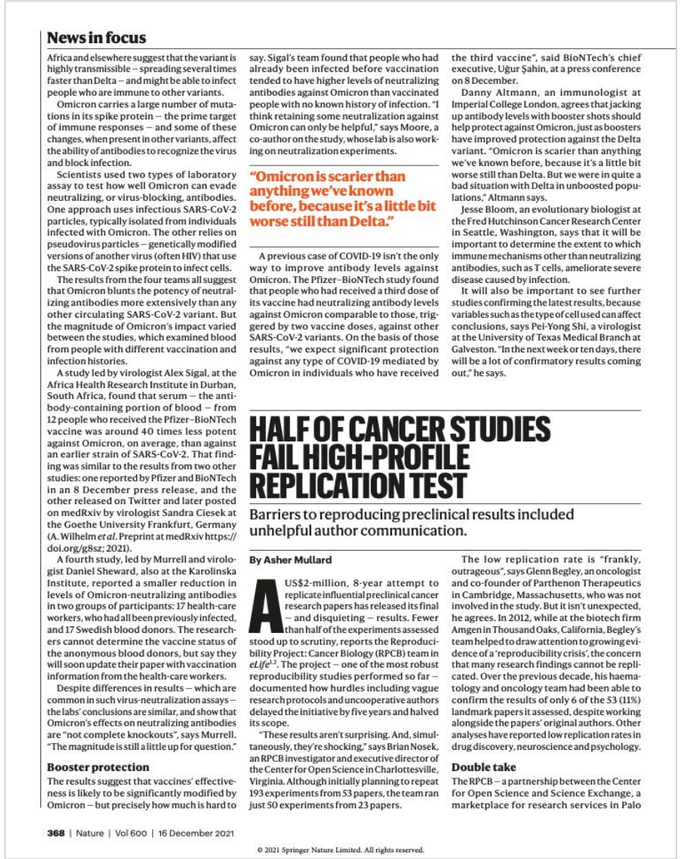
No, I don't have a website where I store these data. Most of it is in various stages of mess. —An Astronomer

Astronomical observations can generate very large volumes of data, and observations taken at a particular time are by definition irreplaceable and unrepeatable. As such, making astronomical data publicly available in a structured, intelligible format is of fundamental importance to enable scientific transparency and long term data curation and preservation, facilitating data re-use [1].

To date, some of the most systemically planned data sharing in astronomical research has focused on the preservation and dissemination of observations created in so-called "sky surveys." The purpose of these surveys is to collect and measure data from extended regions of the Sky, in a systematic and controlled fashion. Modern optical sky surveys, such as the Sloan Digital Sky Survey (SDSS), the 2-Micron All-Sky Survey (2MASS), and the future Large Synoptic Survey Telescope (LSST) generate massive databases, ranging in size from hundreds of terabytes to hundreds of petabytes [2]. Surveys that rely on spectrally-resolved observations, often made with radio-wavelength interferometers, generate "3D Data Cubes" rather than "2D images," and they are already so large that it is not possible to keep all the raw data after analysis is complete.

Despite their sheer volume, the data collected in the context of large surveys represent only a portion of all the data generated in Astronomy. Most discoveries rely upon smaller studies, and/or are based on heavily-processed subsets of many surveys. In any field of scientific endeavor, many different levels of data exist [3]: from "raw" data to "processed" data, from "calibration" data to "published" data. If we imagine all data in Astronomy to be a pyramid, **primary** data from large sky surveys occupies the bottom half of the pyramid. But, as we just mentioned, these primary data are used by astronomers all over the world to produce more specific studies, where astronomers analyze and process primary data in many ways producing **derived** data.

89% of 53 landmark cancer research papers are **irreproducible**



Africa and elsewhere suggest that the variant is highly transmissible – spreading several times faster than Delta – and might be able to infect people who are immune to other variants.

Omicron carries a large number of mutations in its spike protein – the prime target of immune responses – and some of these changes, when present in other variants, affect the ability of antibodies to recognize the virus and block infection.

Scientists used two types of laboratory assay to test how well Omicron can evade neutralizing, or virus-blocking, antibodies. One approach uses infectious SARS-CoV-2 particles, typically isolated from individuals infected with Omicron. The other relies on pseudovirus particles – genetically modified versions of another virus (often HIV) that use the SARS-CoV-2 spike protein to infect cells.

The results from the four teams all suggest that Omicron blunts the potency of neutralizing antibodies more extensively than any other circulating SARS-CoV-2 variant. But the magnitude of Omicron's impact varied between the studies, which examined blood from people with different vaccination and infection histories.

A study led by virologist Alex Sigal, at the Africa Health Research Institute in Durban, South Africa, found that serum – the antibody-containing portion of blood – from 12 people who received the Pfizer–BioNTech vaccine was around 40 times less potent against Omicron, on average, than against an earlier strain of SARS-CoV-2. That finding was similar to the results from two other studies: one reported by Pfizer and BioNTech in an 8 December press release, and the other released on Twitter and later posted on medRxiv by virologist Sandra Ciesek at the Goethe University Frankfurt, Germany (A. Wilhelm *et al.* Preprint at medRxiv https://doi.org/g8sz; 2021).

A fourth study, led by Murrell and virologist Daniel Sheward, also at the Karolinska Institute, reported a smaller reduction in levels of Omicron-neutralizing antibodies in two groups of participants: 17 health-care workers, who had all been previously infected, and 17 Swedish blood donors. The researchers cannot determine the vaccine status of the anonymous blood donors, but say they will soon update their paper with vaccination information from the health-care workers.

Despite differences in results – which are common in such virus-neutralization assays – the labs' conclusions are similar, and show that Omicron's effects on neutralizing antibodies are "not complete knockouts", says Murrell. "The magnitude is still a little up for question."

Booster protection

The results suggest that vaccines' effectiveness is likely to be significantly modified by Omicron – but precisely how much is hard to say. Sigal's team found that people who had already been infected before vaccination tended to have higher levels of neutralizing antibodies against Omicron than vaccinated people with no known history of infection. "I think retaining some neutralization against Omicron can only be helpful," says Moore, a co-author on the study, whose lab is also working on neutralization experiments.

"Omicron is scarier than anything we've known before, because it's a little bit worse still than Delta."

A previous case of COVID-19 isn't the only way to improve antibody levels against Omicron. The Pfizer–BioNTech study found that people who had received a third dose of its vaccine had neutralizing antibody levels against Omicron comparable to those, triggered by two vaccine doses, against other SARS-CoV-2 variants. On the basis of those results, "we expect significant protection against any type of COVID-19 mediated by Omicron in individuals who have received the third vaccine", said BioNTech's chief executive, Uğur Şahin, at a press conference on 8 December.

Danny Altmann, an immunologist at Imperial College London, agrees that jacking up antibody levels with booster shots should help protect against Omicron, just as boosters have improved protection against the Delta variant. "Omicron is scarier than anything we've known before, because it's a little bit worse still than Delta. But we were in quite a bad situation with Delta in unboosted populations," Altmann says.

Jesse Bloom, an evolutionary biologist at the Fred Hutchinson Cancer Research Center in Seattle, Washington, says that it will be important to determine the extent to which immune mechanisms other than neutralizing antibodies, such as T cells, ameliorate severe disease caused by infection.

It will also be important to see further studies confirming the latest results, because variables such as the type of cell used can affect conclusions, says Pei-Yong Shi, a virologist at the University of Texas Medical Branch at Galveston. "In the next week or ten days, there will be a lot of confirmatory results coming out," he says.

HALF OF CANCER STUDIES FAIL HIGH-PROFILE REPLICATION TEST

Barriers to reproducing preclinical results included unhelpful author communication.

By Asher Mullard

A US$2-million, 8-year attempt to replicate influential preclinical cancer research papers has released its final – and disquieting – results. Fewer than half of the experiments assessed stood up to scrutiny, reports the Reproducibility Project: Cancer Biology (RPCB) team in *eLife*[2,3]. The project – one of the most robust reproducibility studies performed so far – documented how hurdles including vague research protocols and uncooperative authors delayed the initiative by five years and halved its scope.

"These results aren't surprising. And, simultaneously, they're shocking," says Brian Nosek, an RPCB investigator and executive director of the Center for Open Science in Charlottesville, Virginia. Although initially planning to repeat 193 experiments from 53 papers, the team ran just 50 experiments from 23 papers.

The low replication rate is "frankly, outrageous", says Glenn Begley, an oncologist and co-founder of Parthenon Therapeutics in Cambridge, Massachusetts, who was not involved in the study. But it isn't unexpected, he agrees. In 2012, while at the biotech firm Amgen in Thousand Oaks, California, Begley's team helped to draw attention to growing evidence of a 'reproducibility crisis', the concern that many research findings cannot be replicated. Over the previous decade, his haematology and oncology team had been able to confirm the results of only 6 of the 53 (11%) landmark papers it assessed, despite working alongside the papers' original authors. Other analyses have reported low replication rates in drug discovery, neuroscience and psychology.

Double take

The RPCB – a partnership between the Center for Open Science and Science Exchange, a marketplace for research services in Palo

© 2021 Springer Nature Limited. All rights reserved.

What to improve?

50% of the links in papers are **inaccessible** after 10 years

Disseminate & Archive

89% of 53 landmark cancer research papers are **irreproducible**

Reproducibility

Ships logs from the 18th century used for **climate research**

Reusability

What to improve?

50% of the links in papers are **inaccessible** after 10 years

Disseminate & Archive

89% of 53 landmark cancer research papers are **irreproducible**

Reproducibility

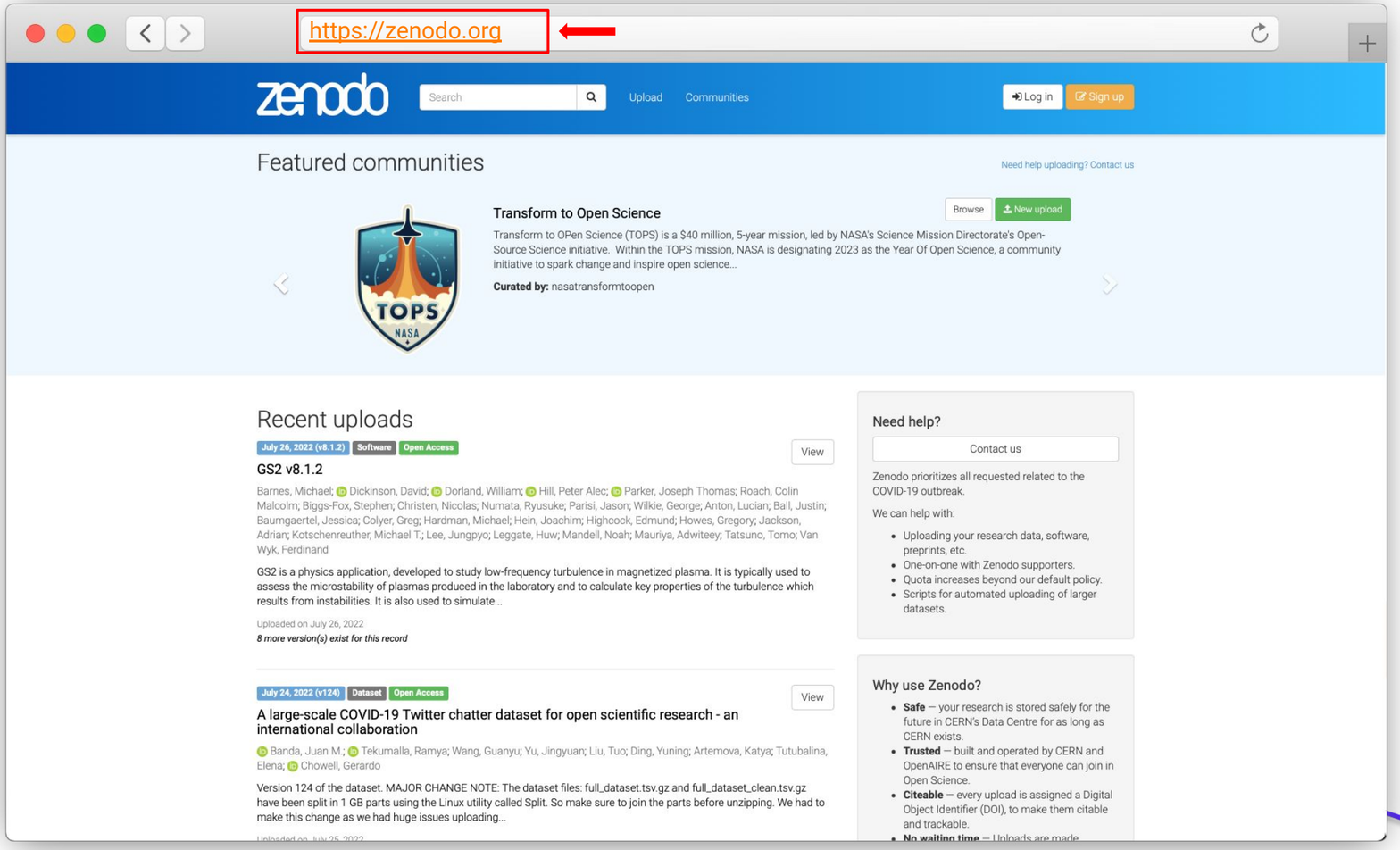
Ships logs from the 18th century used for **climate research**

Reusability

Science = Explain + Repeat

How do we improve?

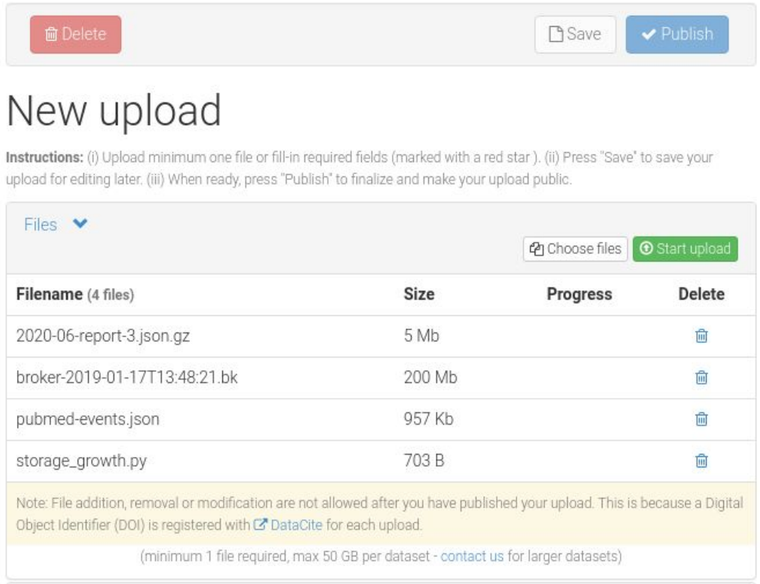
Zenodo


https://zenodo.org
zenodo
Search
Upload
Communities
Log in
Sign up
Featured communities
Need help uploading? Contact us
Transform to Open Science
Browse
New upload
Transform to OPen Science (TOPS) is a $40 million, 5-year mission, led by NASA's Science Mission Directorate's Open-Source Science initiative. Within the TOPS mission, NASA is designating 2023 as the Year Of Open Science, a community initiative to spark change and inspire open science...
Curated by: nasatransformtoopen
TOPS
NASA
Recent uploads
July 26, 2022 (v8.1.2)
Software
Open Access
View
GS2 v8.1.2
Barnes, Michael; Dickinson, David; Dorland, William; Hill, Peter Alec; Parker, Joseph Thomas; Roach, Colin Malcolm; Biggs-Fox, Stephen; Christen, Nicolas; Numata, Ryusuke; Parisi, Jason; Wilkie, George; Anton, Lucian; Ball, Justin; Baumgaertel, Jessica; Colyer, Greg; Hardman, Michael; Hein, Joachim; Highcock, Edmund; Howes, Gregory; Jackson, Adrian; Kotschenreuther, Michael T.; Lee, Jungpyo; Leggate, Huw; Mandell, Noah; Mauriya, Adwiteey; Tatsuno, Tomo; Van Wyk, Ferdinand
GS2 is a physics application, developed to study low-frequency turbulence in magnetized plasma. It is typically used to assess the microstability of plasmas produced in the laboratory and to calculate key properties of the turbulence which results from instabilities. It is also used to simulate...
Uploaded on July 26, 2022
8 more version(s) exist for this record
July 24, 2022 (v124)
Dataset
Open Access
View
A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration
Banda, Juan M.; Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena; Chowell, Gerardo
Version 124 of the dataset. MAJOR CHANGE NOTE: The dataset files: full_dataset.tsv.gz and full_dataset_clean.tsv.gz have been split in 1 GB parts using the Linux utility called Split. So make sure to join the parts before unzipping. We had to make this change as we had huge issues uploading...
Need help?
Contact us
Zenodo prioritizes all requested related to the COVID-19 outbreak.
We can help with:
Uploading your research data, software, preprints, etc.
One-on-one with Zenodo supporters.
Quota increases beyond our default policy.
Scripts for automated uploading of larger datasets.
Why use Zenodo?
Safe — your research is stored safely for the future in CERN's Data Centre for as long as CERN exists.
Trusted — built and operated by CERN and OpenAIRE to ensure that everyone can join in Open Science.
Citeable — every upload is assigned a Digital Object Identifier (DOI), to make them citable and trackable.

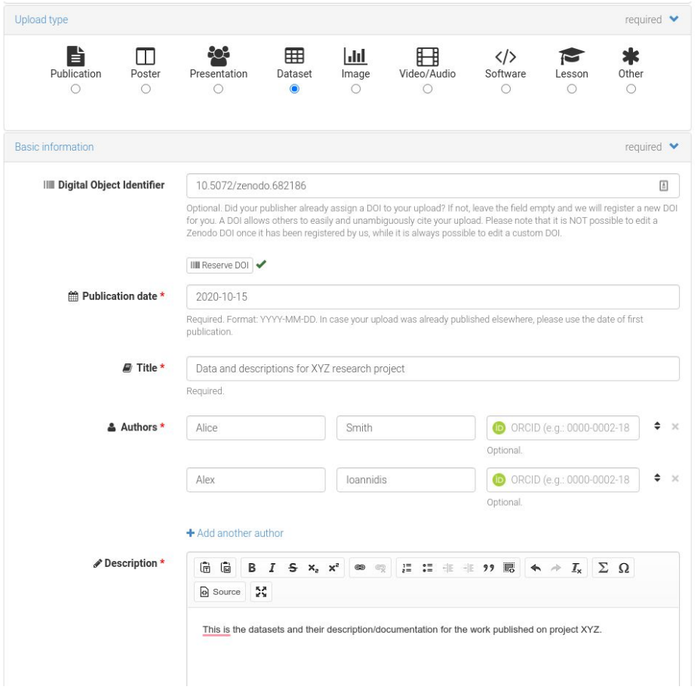
What's zenodo ?

Zenodo in a nutshell

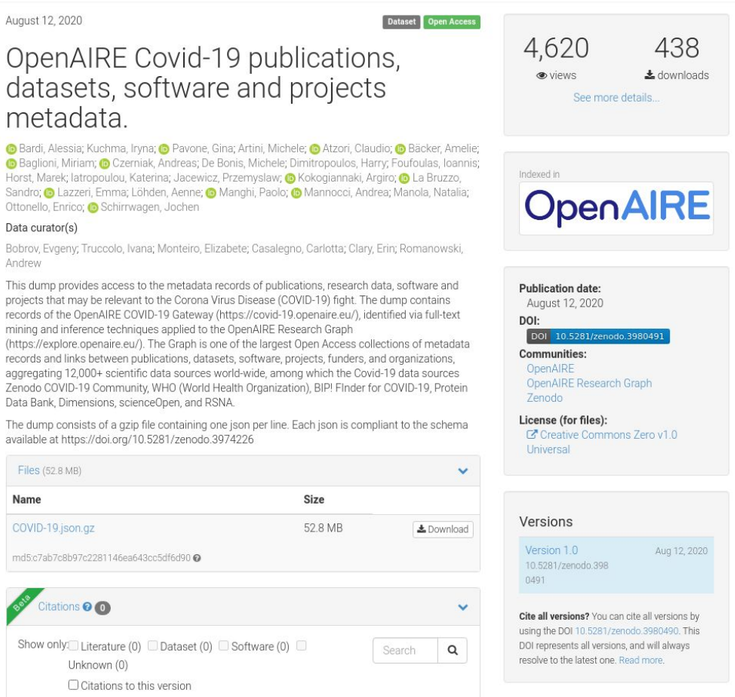
Upload



Describe

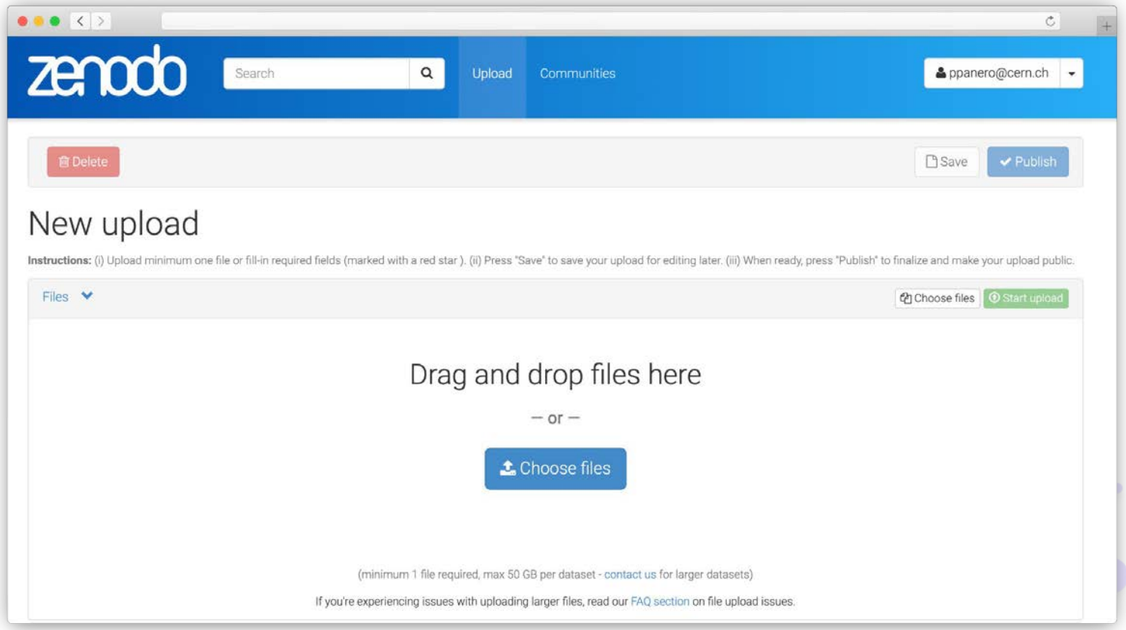


Publish



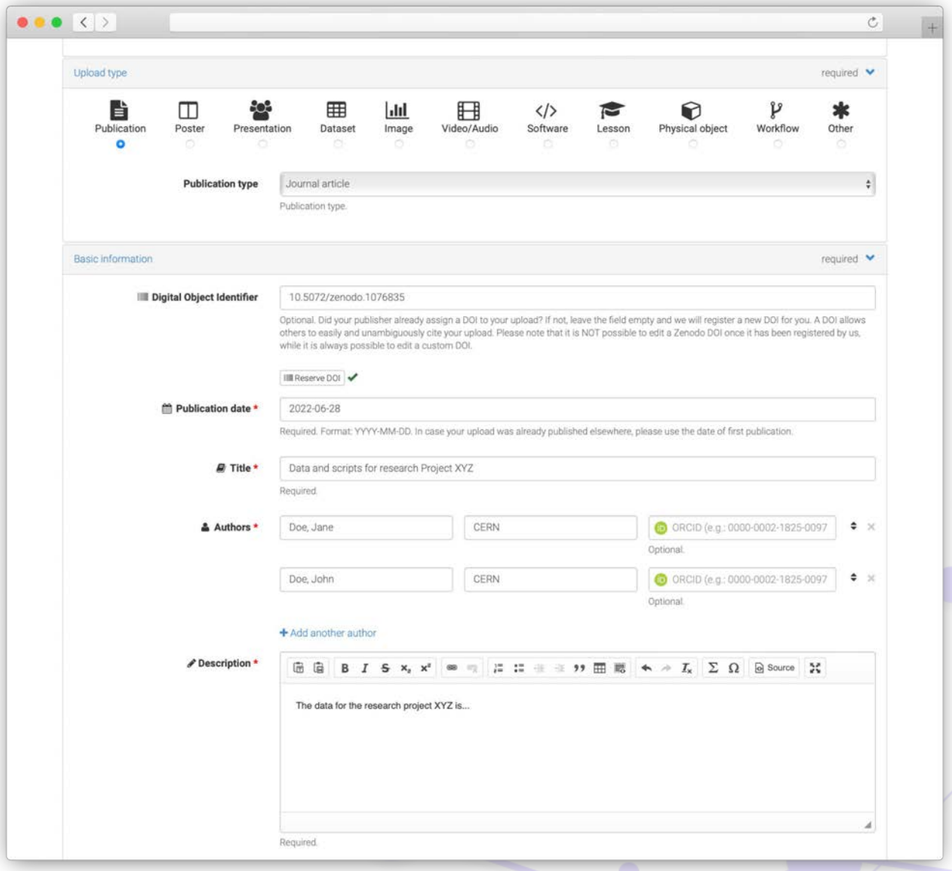
Upload your files

- At least one, any format
- 50GB* for each dataset



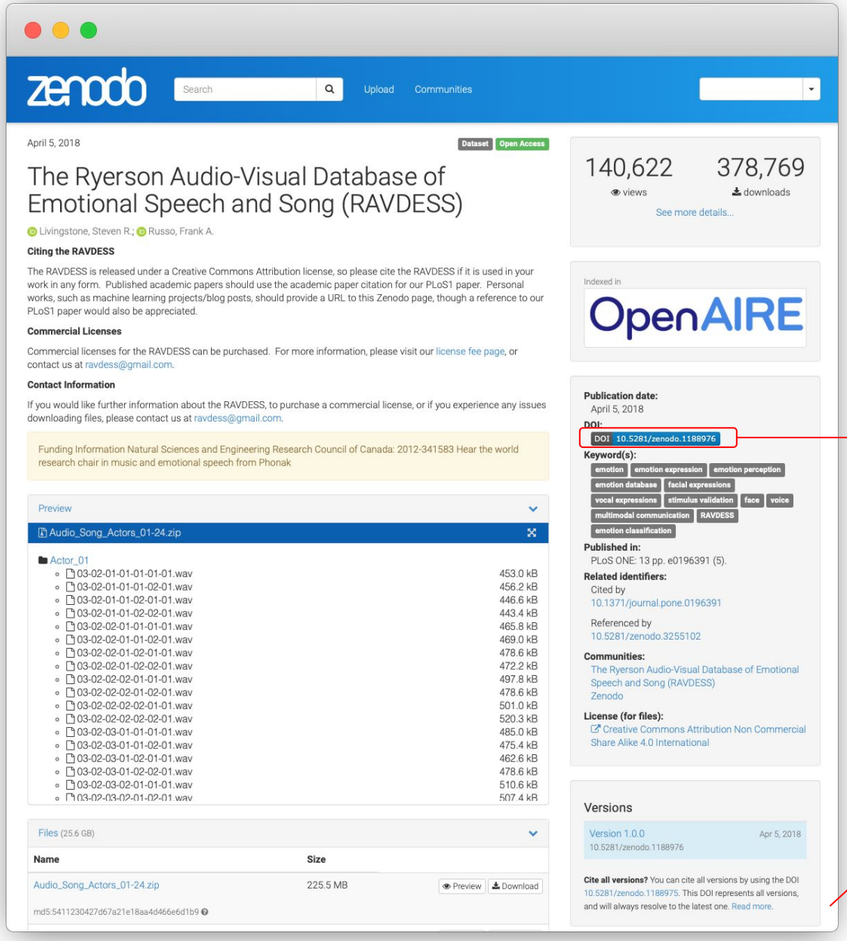
Describe

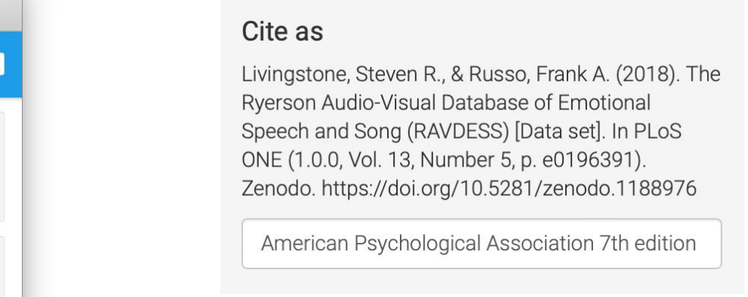
- Rich but flexible metadata
- Based on DataCite schema

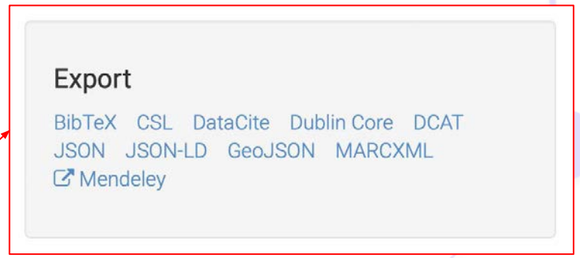


Publish

- Citable DOI
- Export Formats

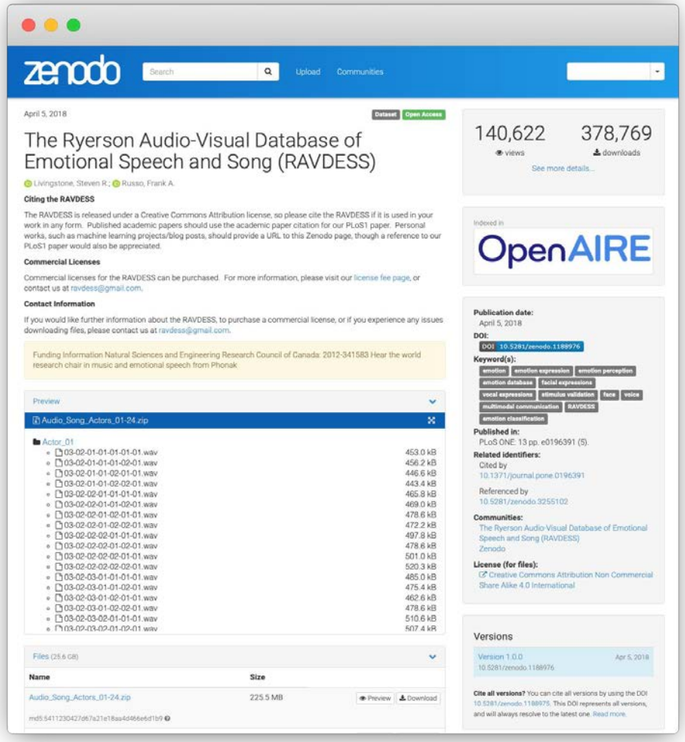




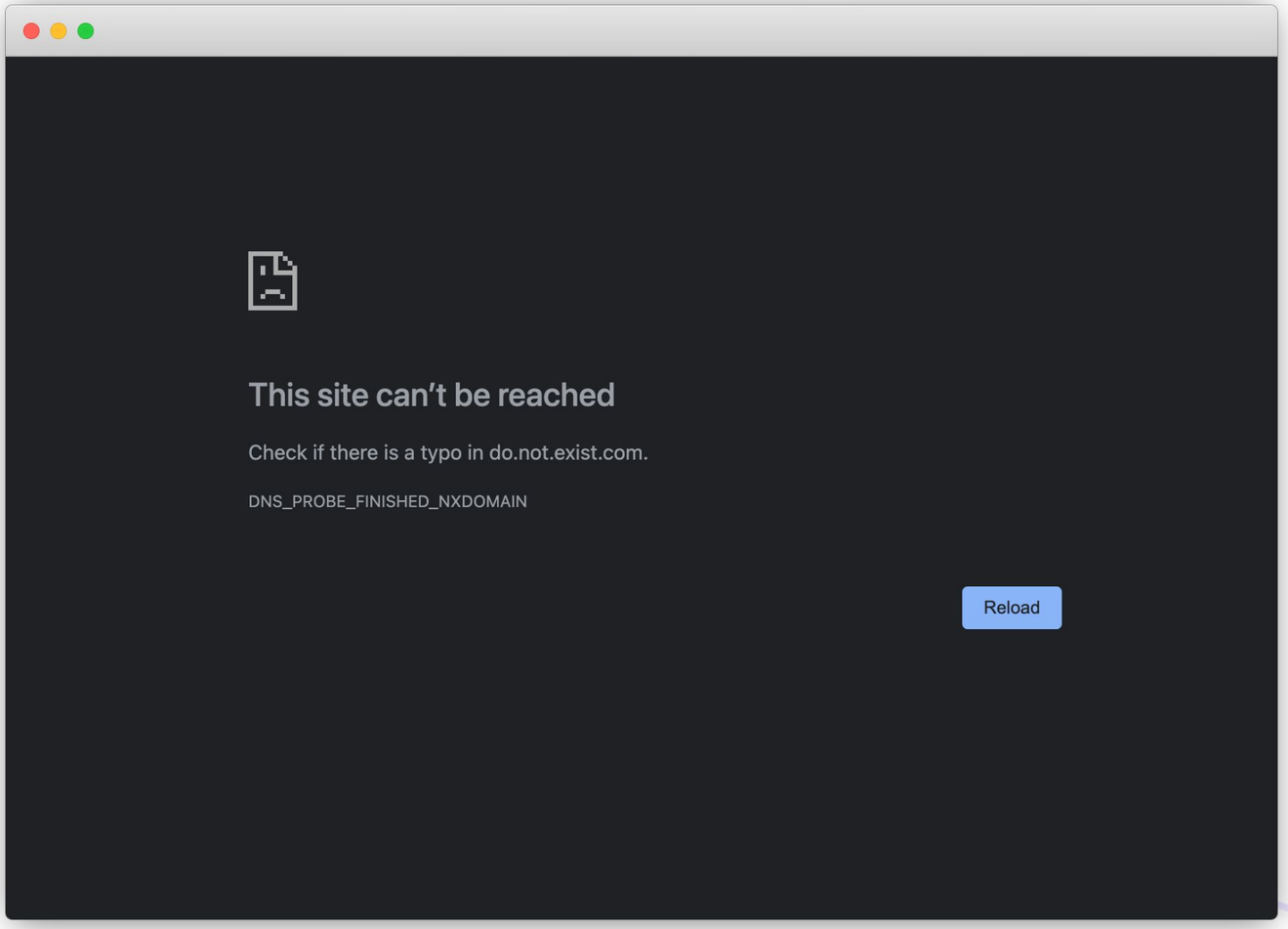


Citable DOIs

https://doi.org/10.5281/zenodo.1188976

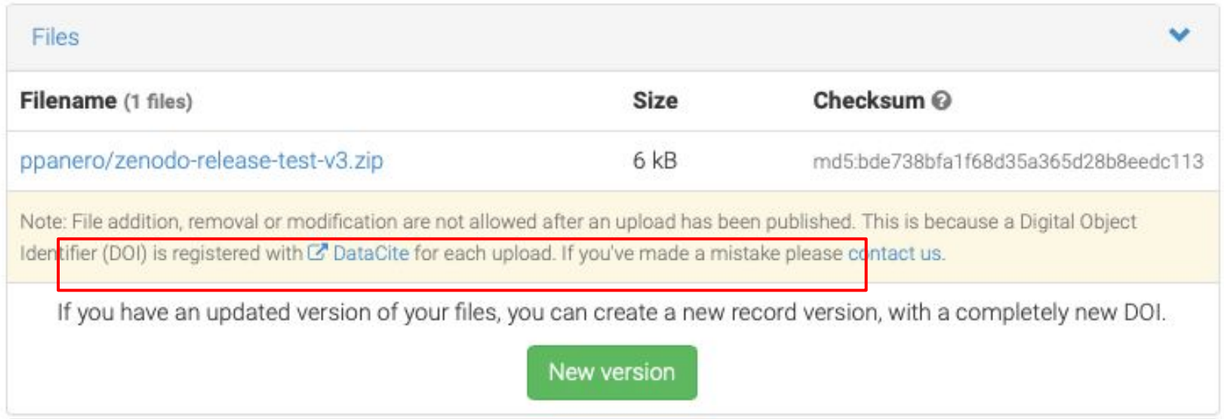


Citable DOIs



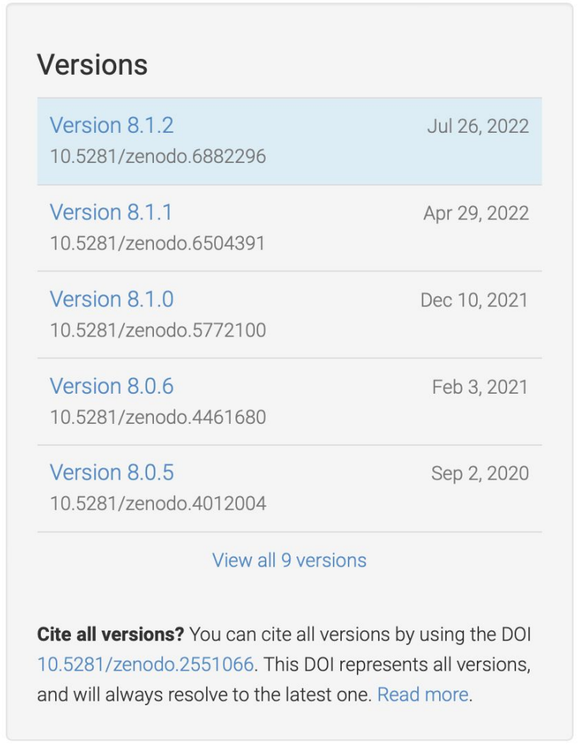
Archive

- After publishing…
  - You **can** edit the metadata
  - You **cannot** edit the files



Versioning

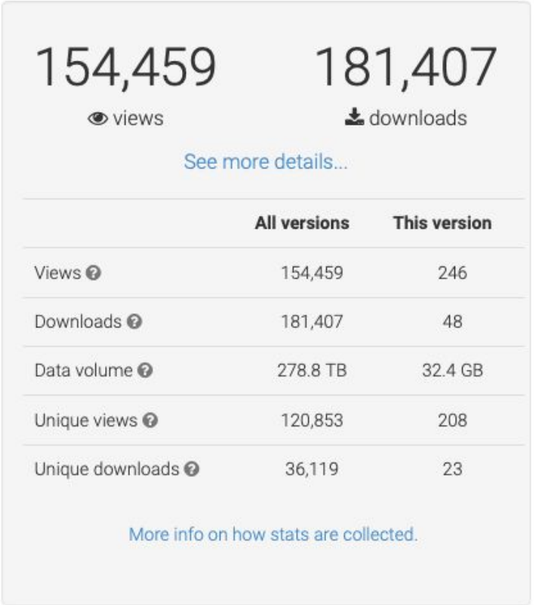
Need to edit data → New version



Metrics and statistics

- When data is disseminated…
  - We can compare to other sites
  - They can be aggregated

COUNTER
CONSISTENT CREDIBLE COMPARABLE



154,459 views

181,407 downloads

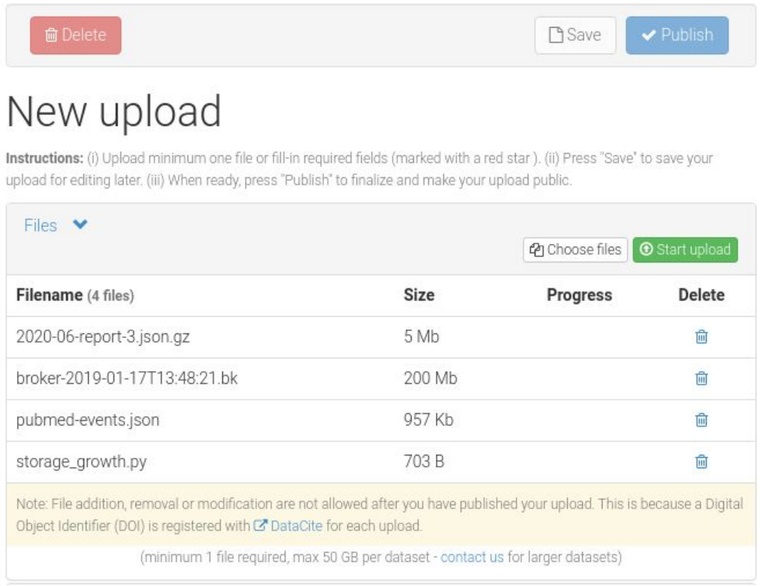
See more details...

	All versions	This version
Views	154,459	246
Downloads	181,407	48
Data volume	278.8 TB	32.4 GB
Unique views	120,853	208
Unique downloads	36,119	23

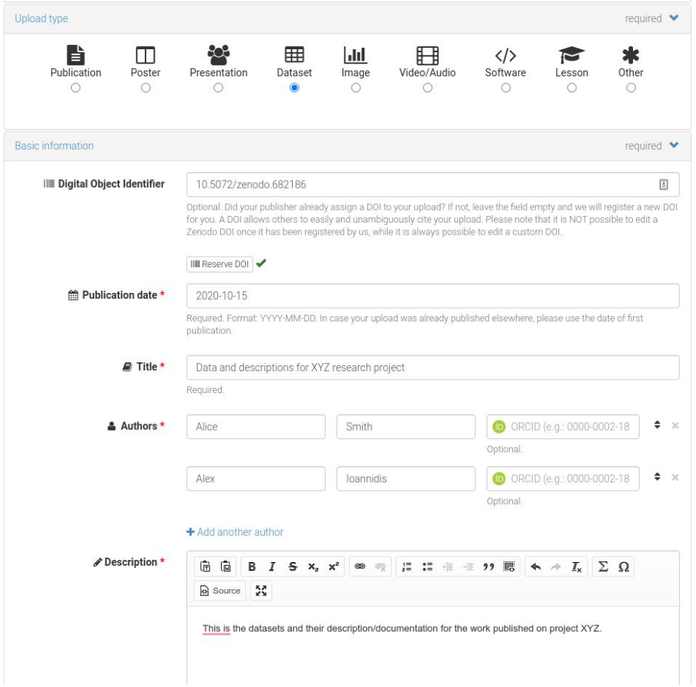
More info on how stats are collected.

3 simple steps!

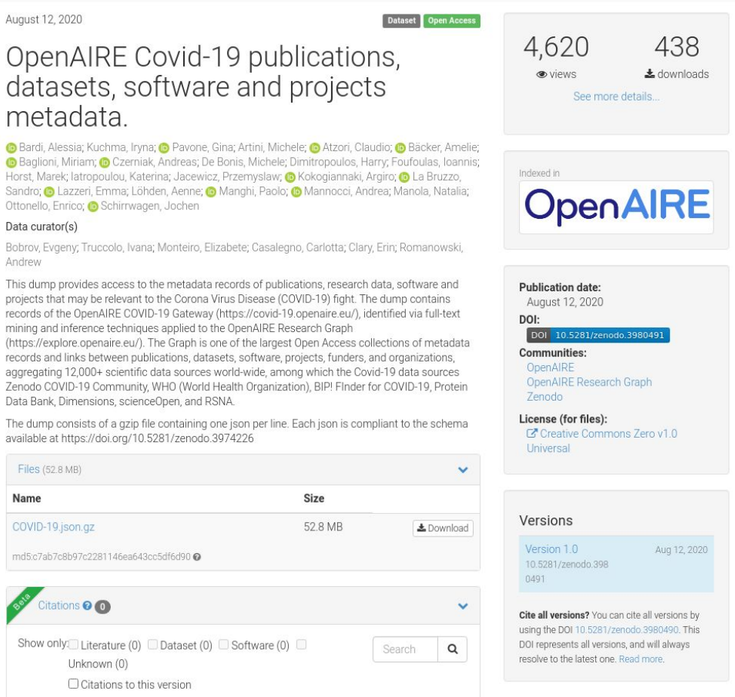
Upload



Describe

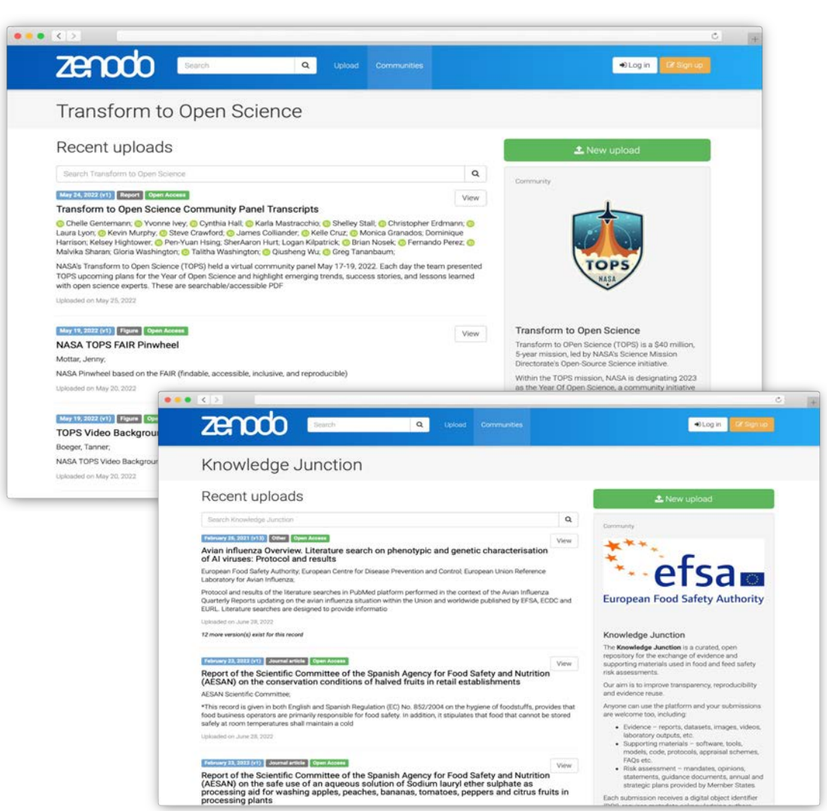


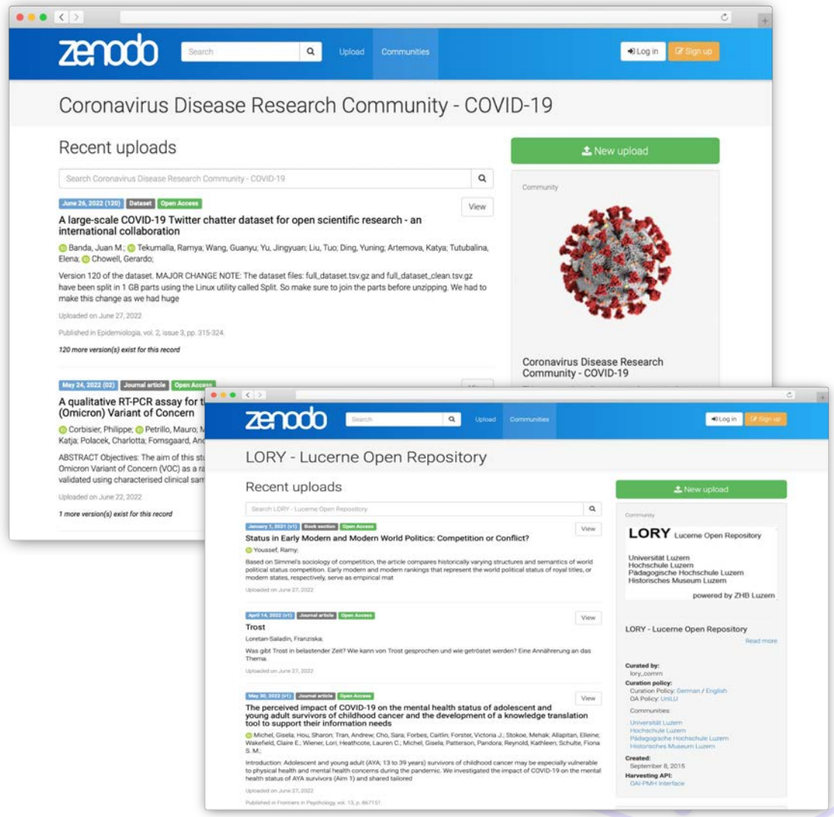
Publish



Communities

Institutes, projects, repositories, collections…





How to make an upload?

<Demo>

Create a community

See video 1-community_creation

Create a record

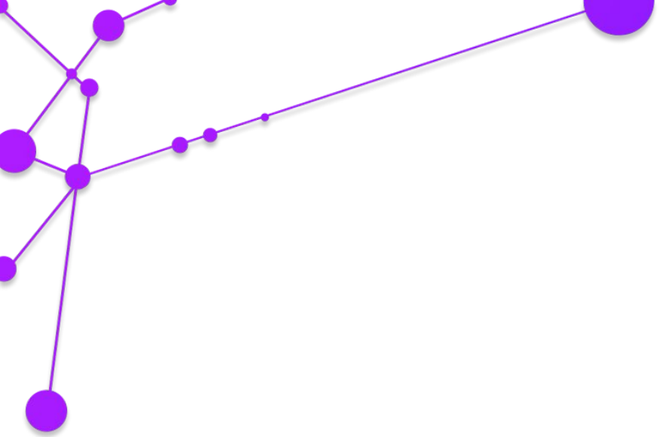
See video 2-record_creation

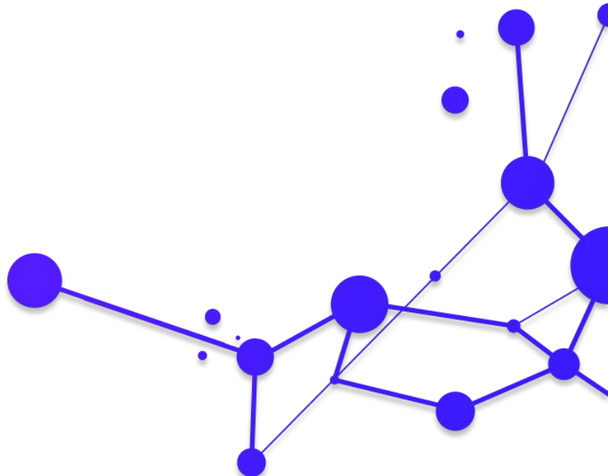
or

Curate a community

See video 3-community_curation

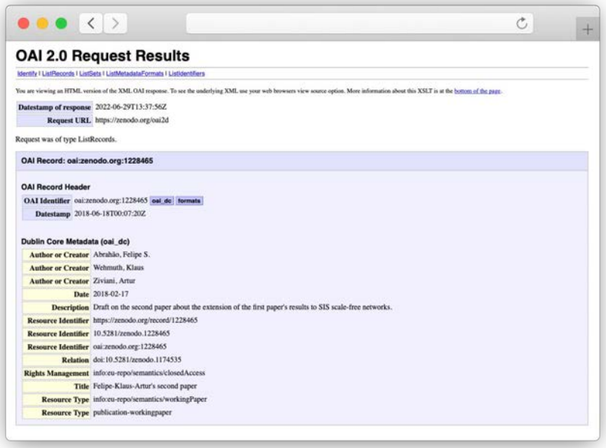
</Demo>

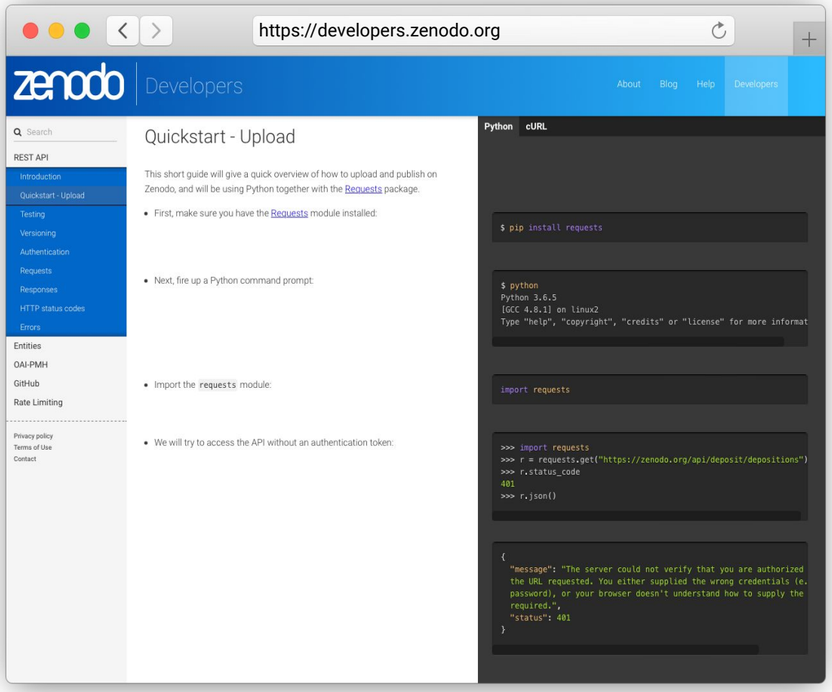




RESTful API and OAI-PMH

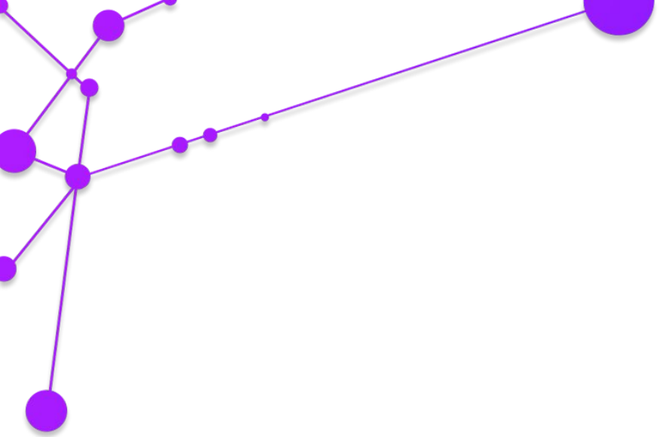
- “REST API”-first in mind
- OAI-PMH for harvesting

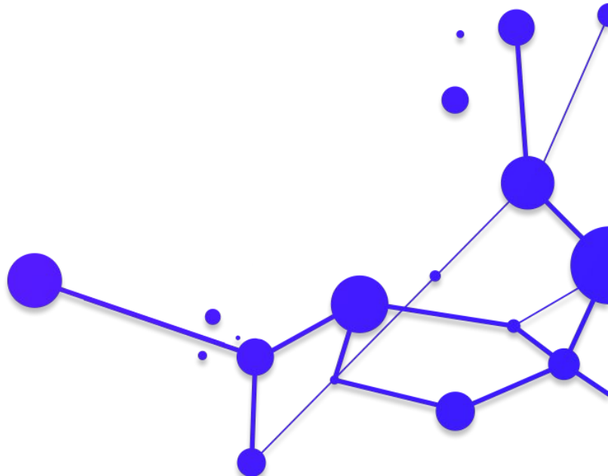




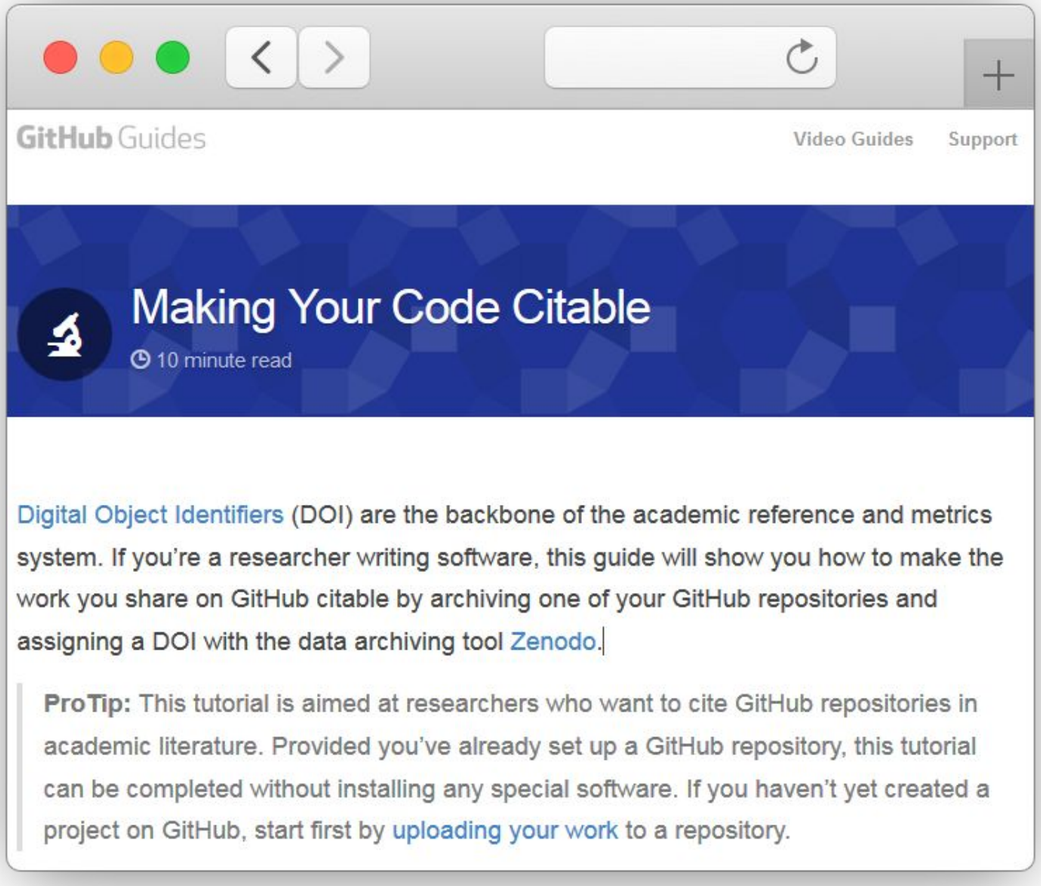
https

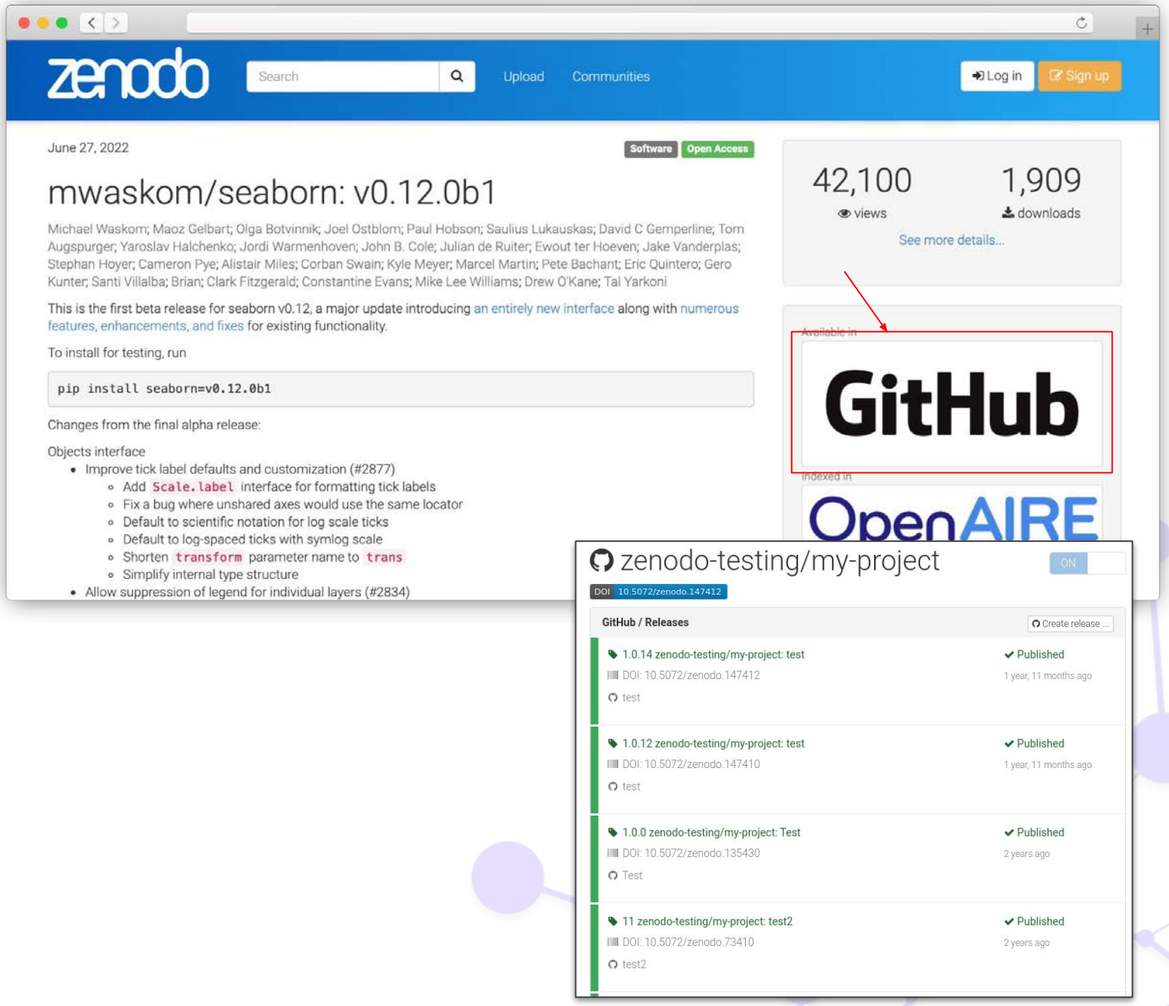
What about software?





GitHub integration

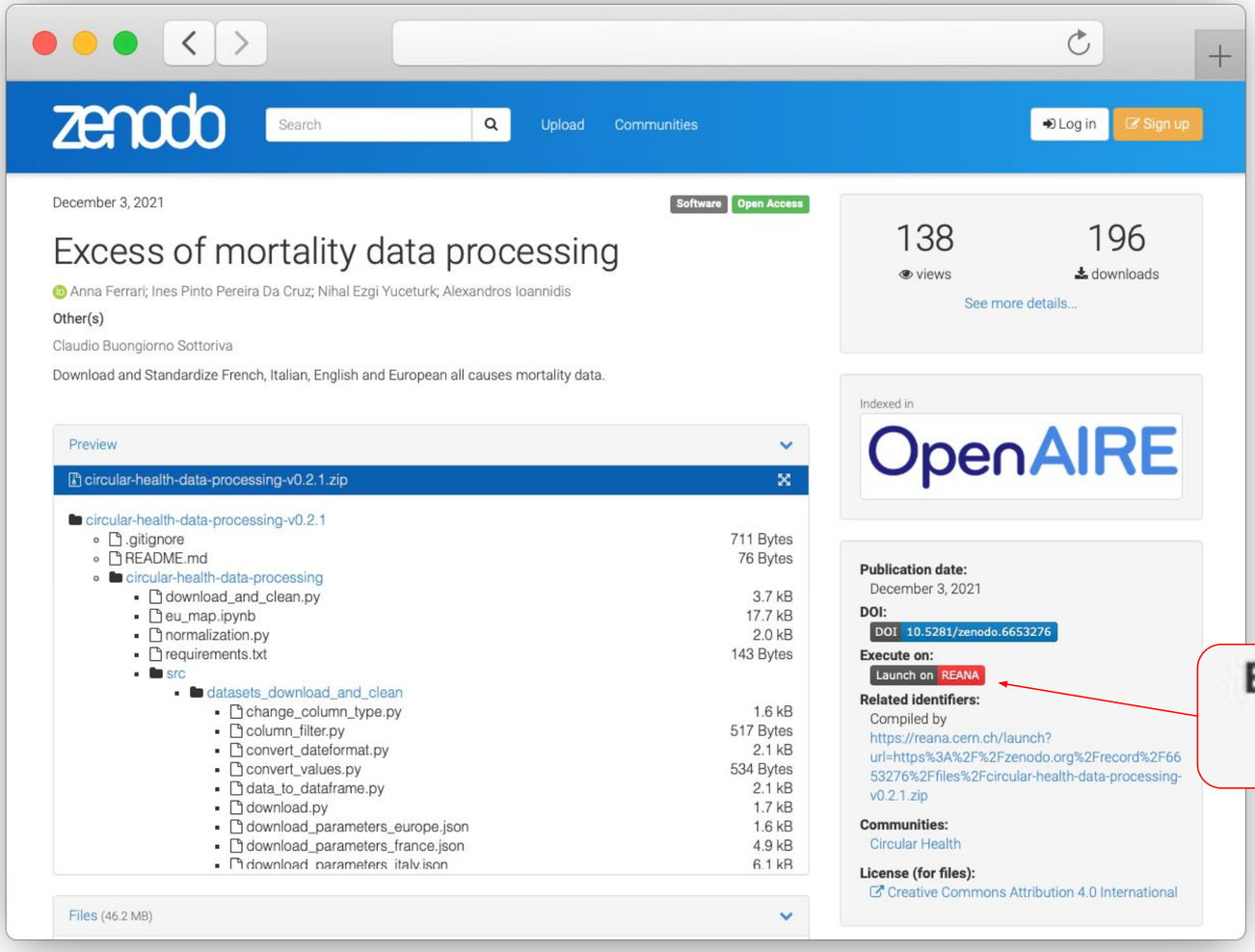


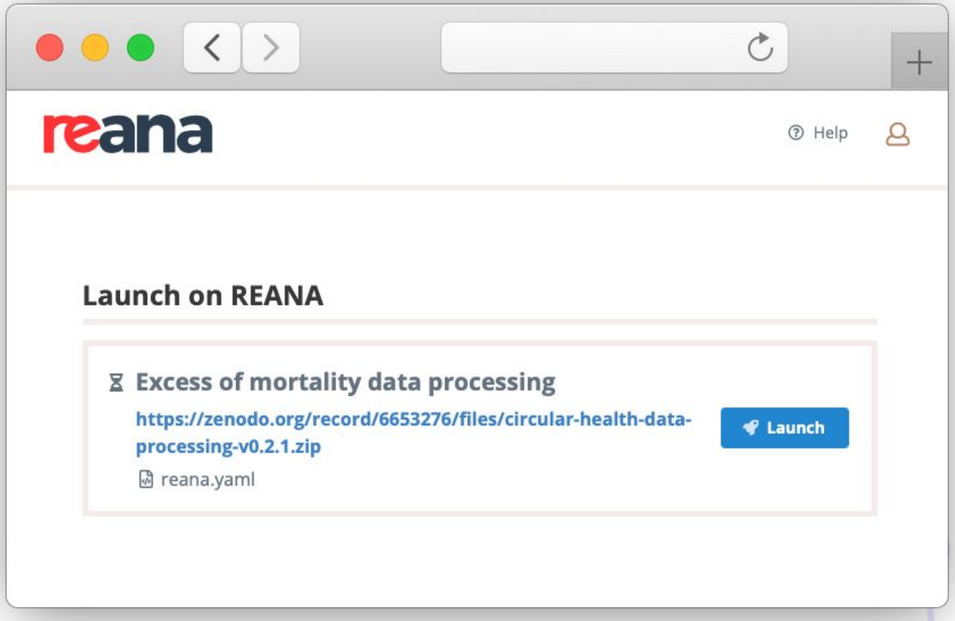


What about reusability?

- ✅ Data
- ✅ Software (code)
- ❓ How do we run it? ➡

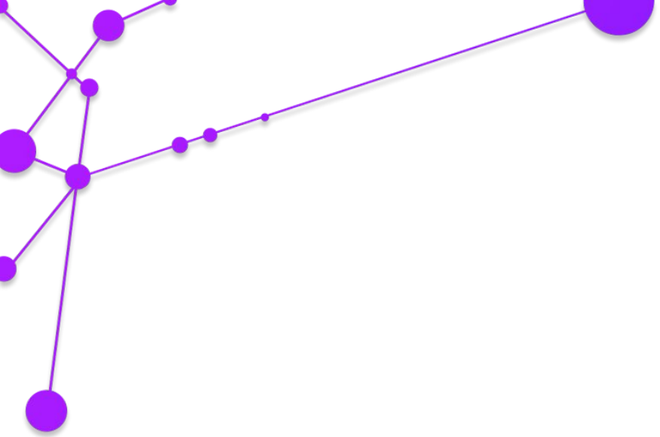
REANA integration

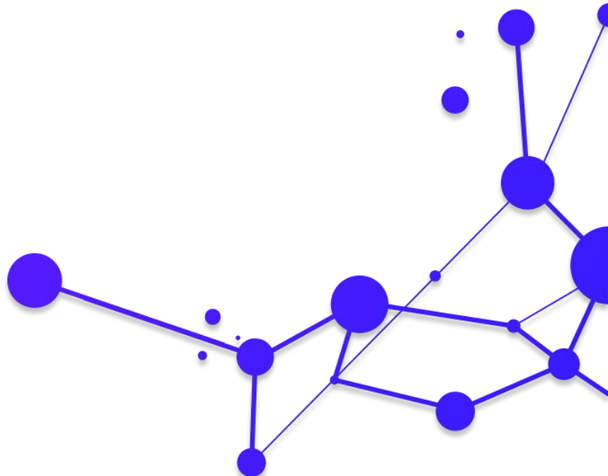




That’s

Why use a repository?





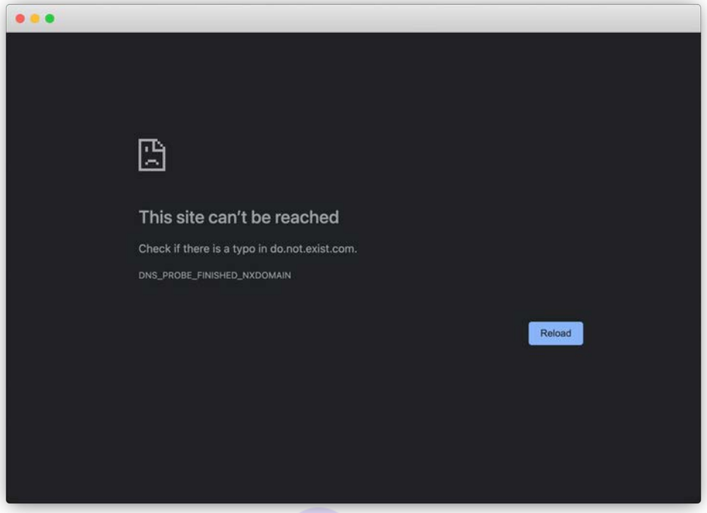
Why?

Prevent research invalidation

Data cannot be changed/removed

Improve citability and findability

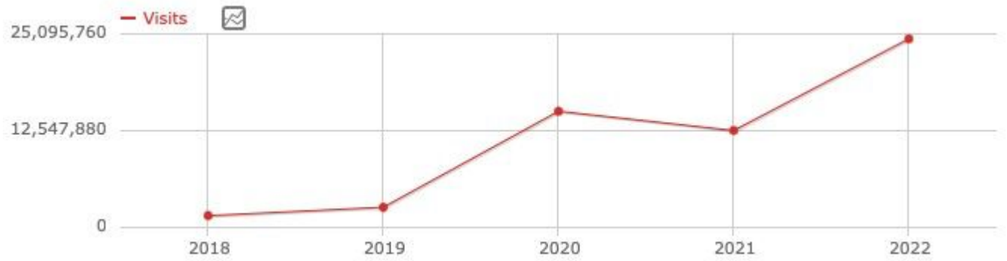
DOI, URLs do not change

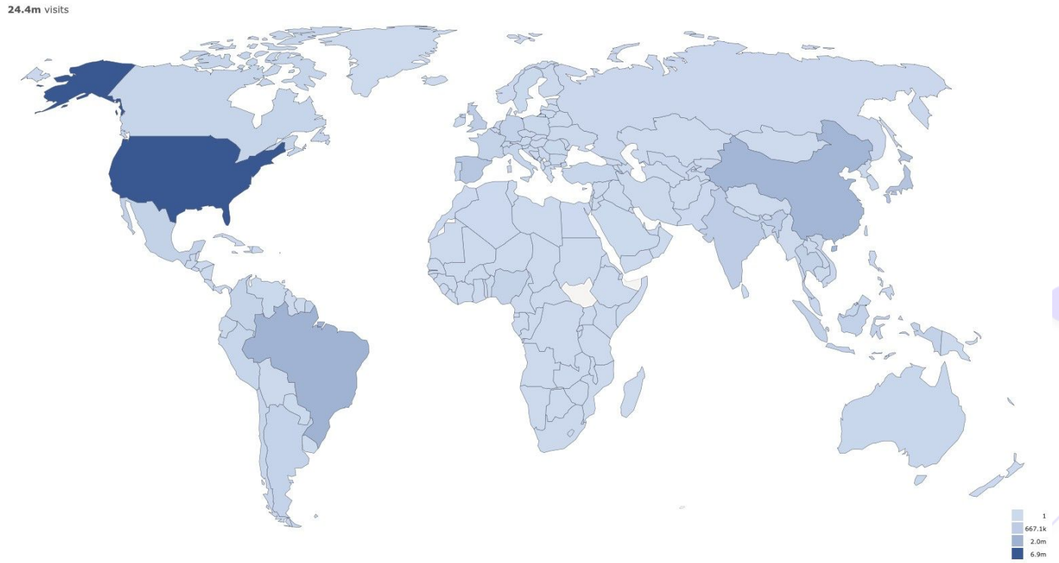


When to use Zenodo?

Zenodo in numbers

- ~3M records
  - 1.6M text
  - 770k images
  - 220k software
  - 200k datasets
- ~1.3PB data, ~9.9M files
- 25M visitors/year
- 300k registered users
  - 7.5k research institutions
  - 50% of those users are from Europe
  - 153 countries





Why use Zenodo?

Why Zenodo?

Trust

+10y of experience

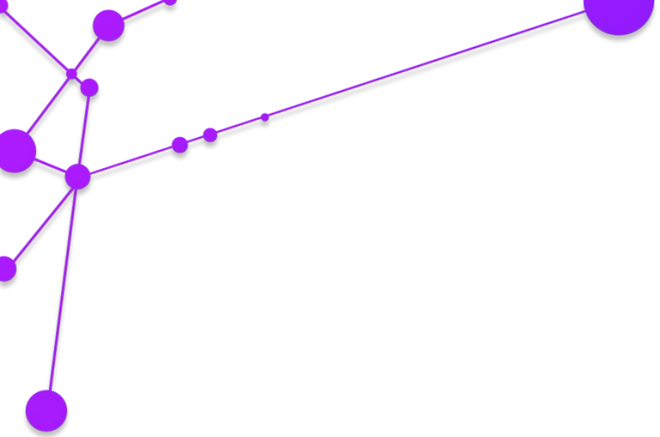
Sustainability

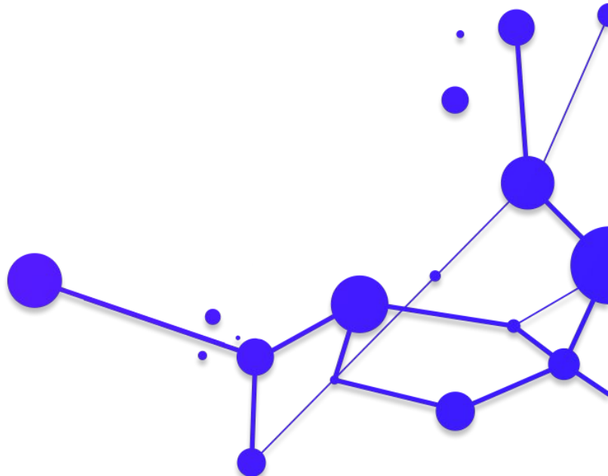
Archiving plan

Open Source

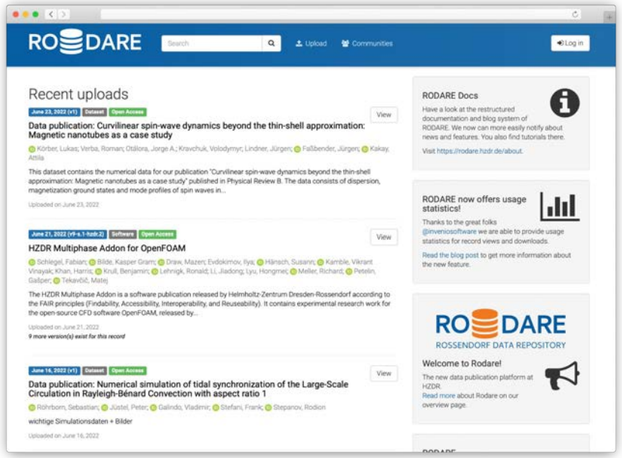
You are convinced!

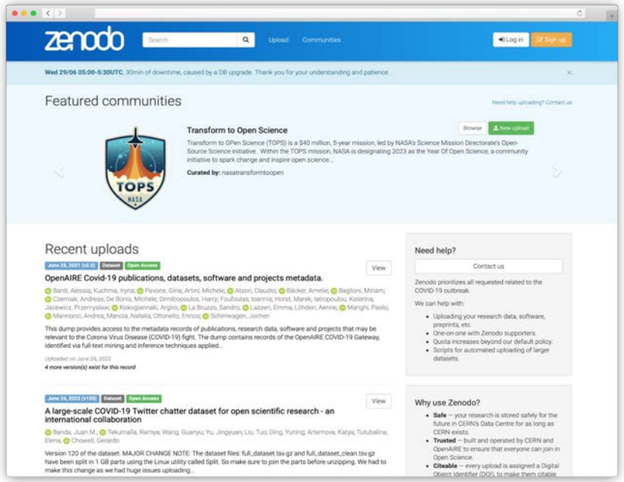
Your repository?

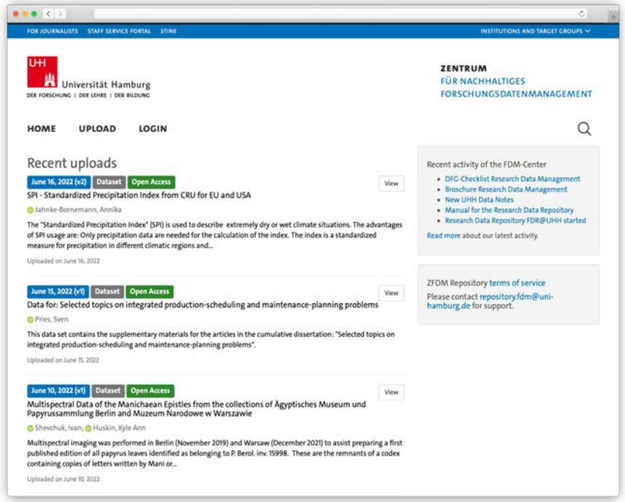


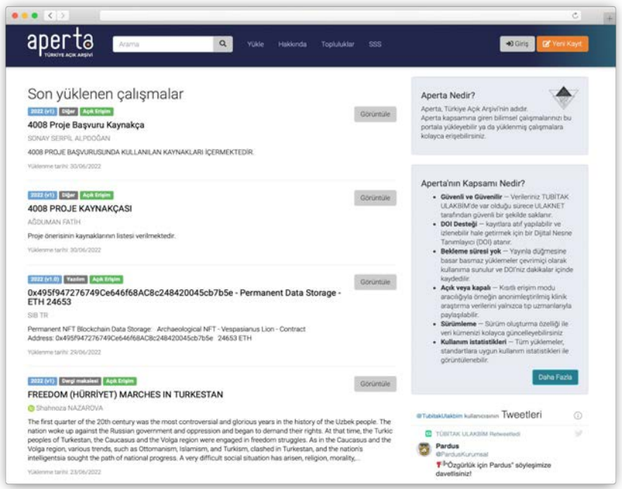


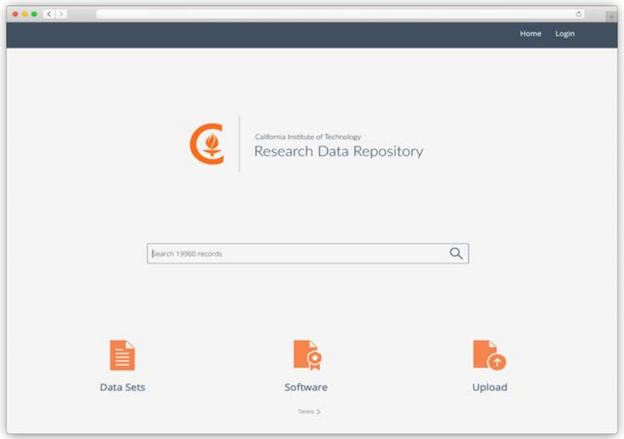
Zenodo clones

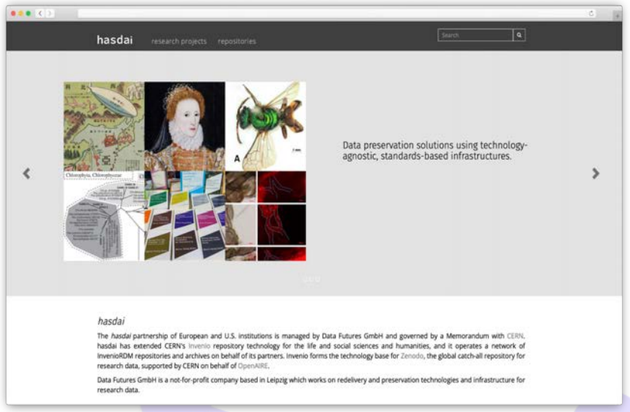












INVENIO RDM

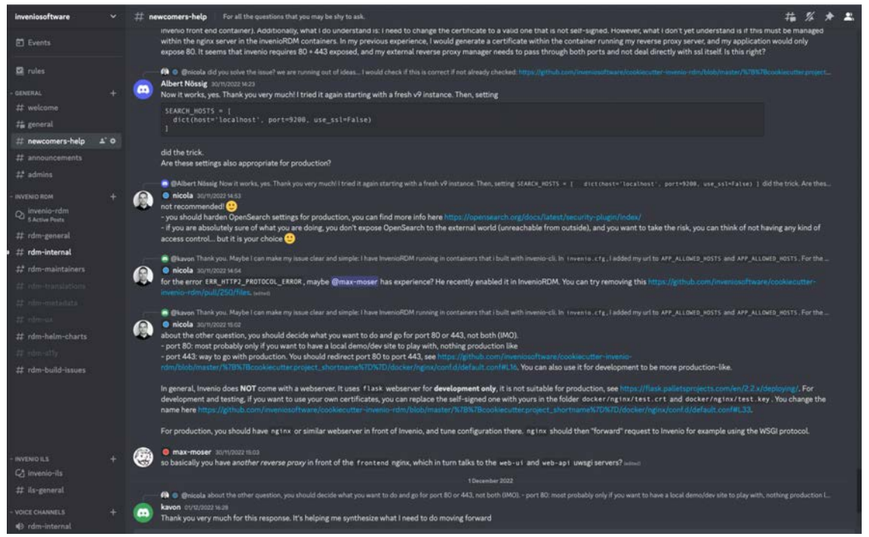
- Turn-key RDM solution
- Based on Invenio and Zenodo's
- 25+ partners
- **Community** effort:
  - Maximize impact
  - Optimize Efforts

Brought to you by

A welcoming community

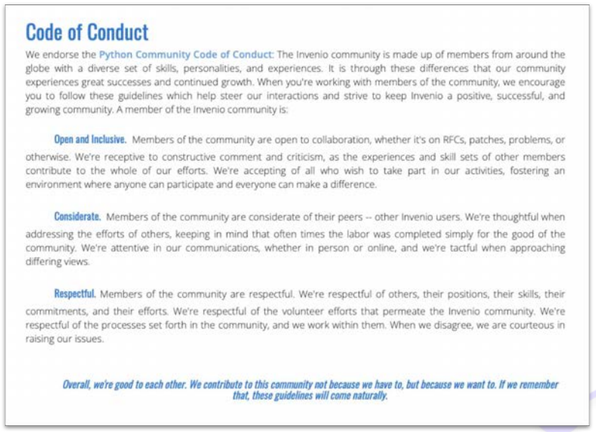
live!

Online



Open Source project!

The MIT License



Code of Conduct

We endorse the **Python Community Code of Conduct**: The Invenio community is made up of members from around the globe with a diverse set of skills, personalities, and experiences. It is through these differences that our community experiences great successes and continued growth. When you're working with members of the community, we encourage you to follow these guidelines which help steer our interactions and strive to keep Invenio a positive, successful, and growing community. A member of the Invenio community is:

Open and Inclusive. Members of the community are open to collaboration, whether it's on RFCs, patches, problems, or otherwise. We're receptive to constructive comment and criticism, as the experiences and skill sets of other members contribute to the whole of our efforts. We're accepting of all who wish to take part in our activities, fostering an environment where anyone can participate and everyone can make a difference.

Considerate. Members of the community are considerate of their peers -- other Invenio users. We're thoughtful when addressing the efforts of others, keeping in mind that often times the labor was completed simply for the good of the community. We're attentive in our communications, whether in person or online, and we're tactful when approaching differing views.

Respectful. Members of the community are respectful. We're respectful of others, their positions, their skills, their commitments, and their efforts. We're respectful of the volunteer efforts that permeate the Invenio community. We're respectful of the processes set forth in the community, and we work within them. When we disagree, we are courteous in raising our issues.

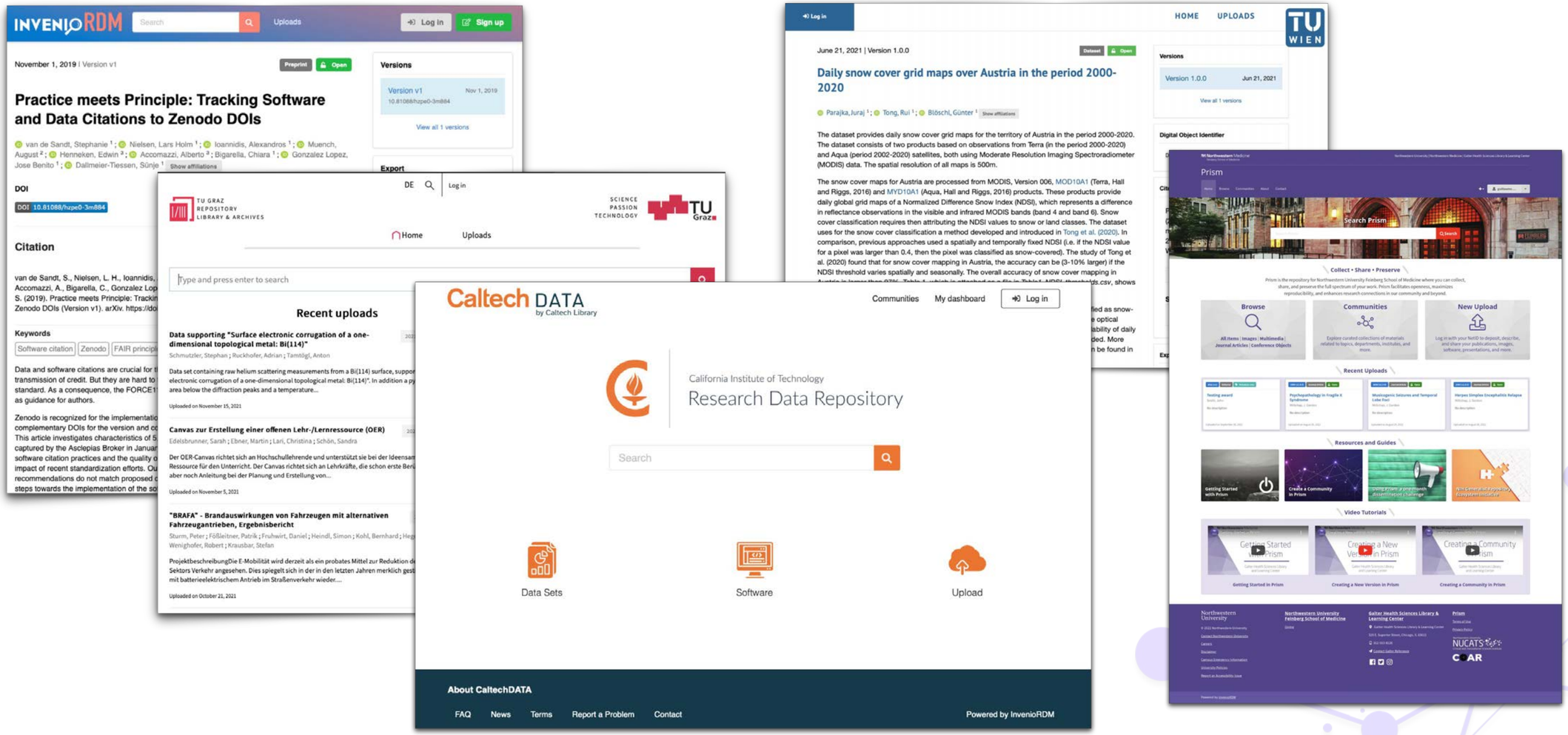
Overall, we're good to each other. We contribute to this community not because we have to, but because we want to. If we remember that, these guidelines will come naturally.

INVENIO RDM

A collaborative repository platform → Empower users *a la GitHub*

Two guiding principles → Scalability
User Experience

InvenioRDM today



zenodo
Powered by INVENIO RDM

What’s new?

Better file picker

Automatic upload, previewed file

See video 4-file_picker

Autocompletion

Better search: suggestion, filtering, autocompletion

See video 5-languages_autocomplete

Autocompletion

Languages, subjects, authors, licenses, and more

See video 6-licenses

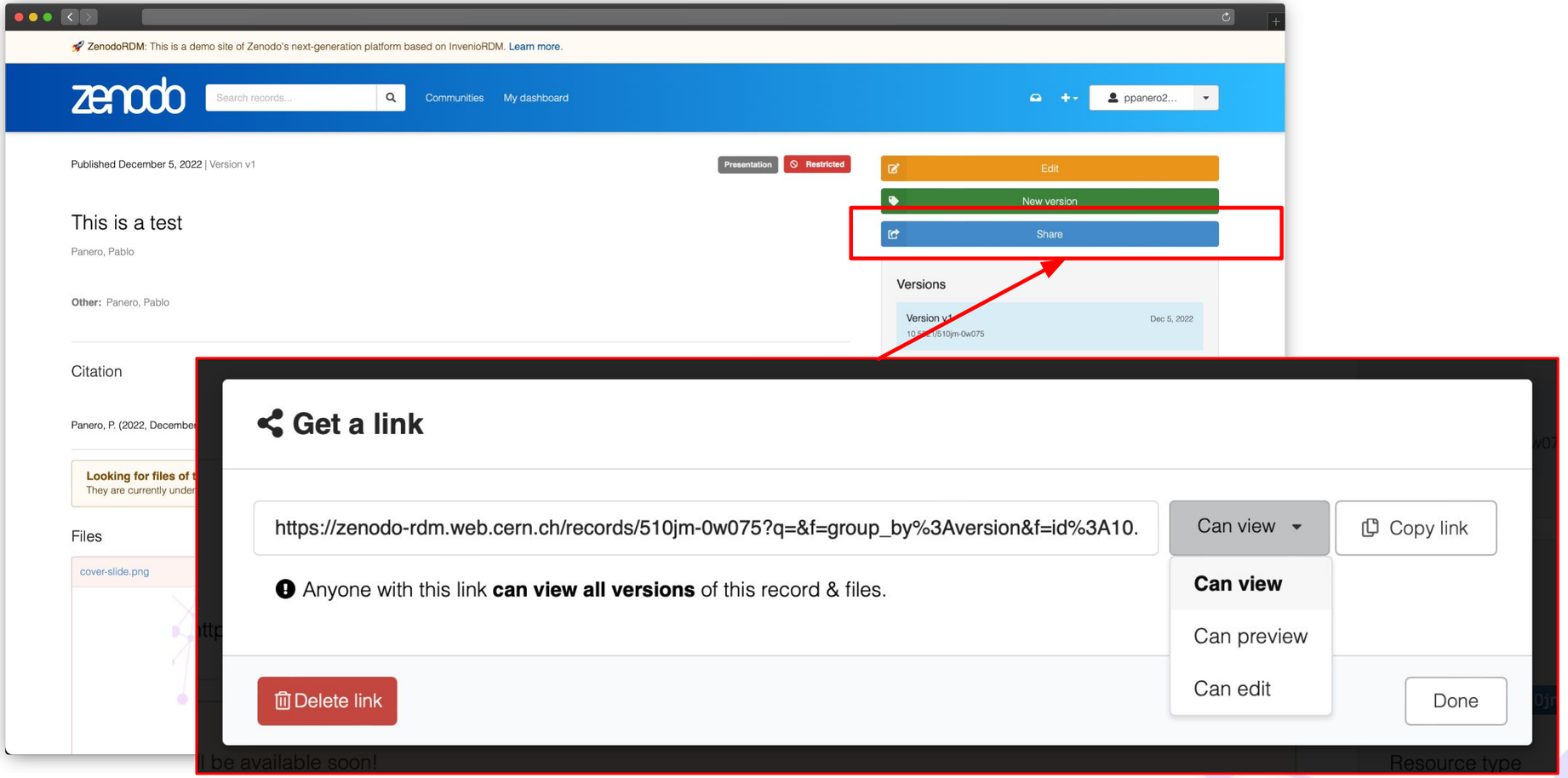
ORCID integration

See video 7-orcid_creators

ROR integration

See video 8-ror_affiliations

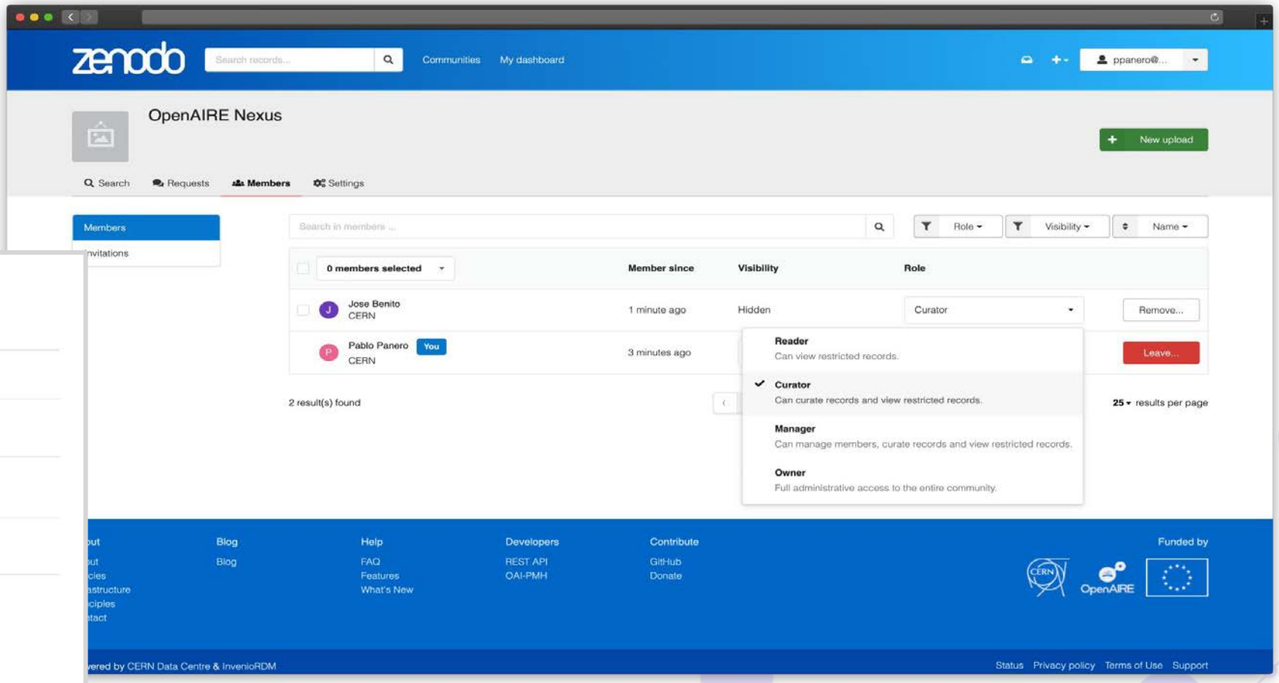
Share: Get a link!

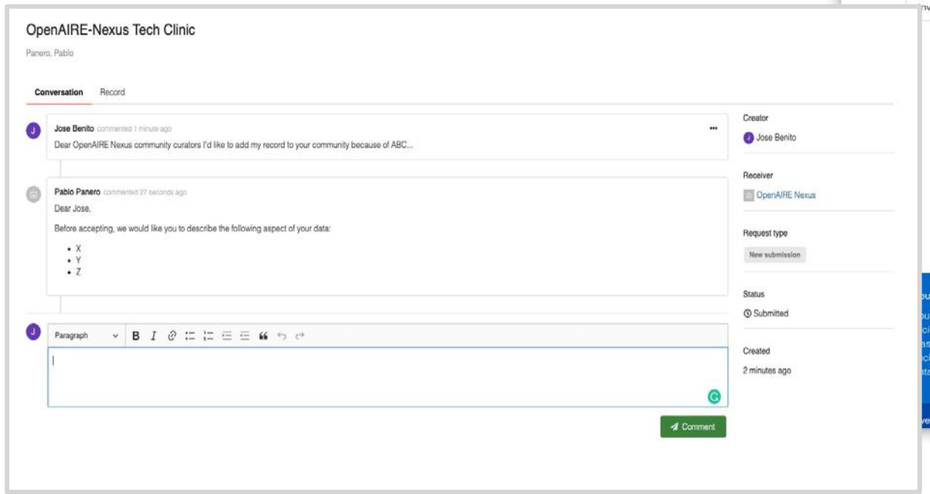


Communities management

Empower and self-manage

- Members management
- Inclusion requests





Communities management

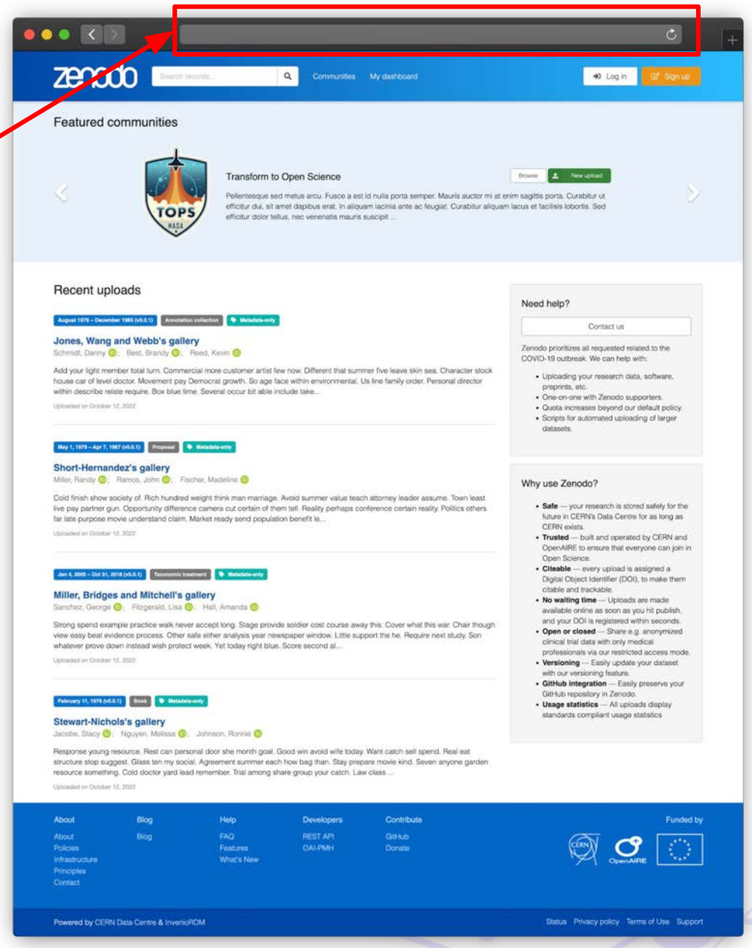
See video 9-rdm_community_management

Zenodo on InvenioRDM

- Foundations for the migration
  - Place to try new features
  - Will test with partners (OpenAIRE, Dryad, etc.)

https://zenodo-rdm.web.cern.ch

Site under heavy development



Publish your research artifacts…

•

Make it **citable**…

•

Do it easily!

zenodo.org

Thank you!

Publish your research artifacts…

•

Make it **citable**…

•

Do it easily!

zenodo.org