

# Zenodo migration

Zacharodimos Zacharias (**Zach**)  
CERN

OR2023 - InvenioRDM Workshop , 2023-06-12

# What is Zenodo?

Cross-domain digital repository for the long tail of research.

Computer science,  
Biodiversity, Humanities,  
Chemistry, ...

.pdf, .zip, .gz, .h5,  
.avi, .tiff, .png,  
.ipynb, .r, ...

- Launched in May 2013, by OpenAIRE & CERN
- Hosted at the CERN data center
- Free

## References

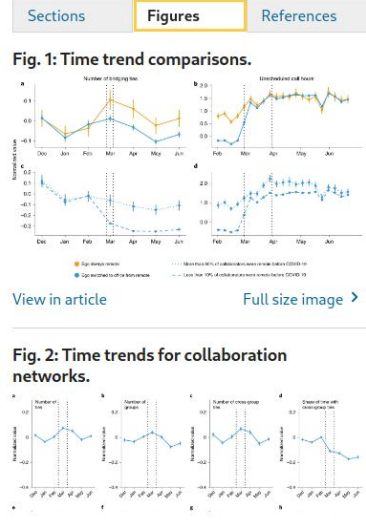
1. Bloom, N. A. *Working From Home and the Future of U.S. Economic Growth Under COVID* (2020); <https://www.yourpaper.com/v=jtdFIZx3hyk>
2. Brynjolfsson, E. et al. *COVID-19 and Remote Work Data*. Technical Report (National Bureau of Economic Research, 2020).
3. Barrero, J. M., Bloom, N. & Davis, S. 60 million hours per day: how Americans use time saved at home. Working Paper (Univ. Chicago Becker-DeMunnick Center for Behavioral & Economic Analysis, 2020); [https://bfj.uchicago.edu/content/uploads/2020/09/BFJ\\_WP\\_2020132](https://bfj.uchicago.edu/content/uploads/2020/09/BFJ_WP_2020132)
4. Dingel, J. I. & Neiman, B. How many jobs can be done at home? *Public Econ.* **189**, 104235 (2020).

## Data availability

An anonymized version of the data is publicly available due to an agreement with Microsoft. The data are available from Microsoft on request with permission from Microsoft Corporation.

## Code availability

The code supporting this study is available for academic purposes. The code is available on GitHub with a Creative Commons Attribution-NonCommercial-ShareAlike license. The code is available on GitHub with a Creative Commons Attribution-NonCommercial-ShareAlike license. The code is available on GitHub with a Creative Commons Attribution-NonCommercial-ShareAlike license. The code is available on GitHub with a Creative Commons Attribution-NonCommercial-ShareAlike license.



# InvenioRDM - A better & customizable Zenodo

- Turn-key research management repository
- Better & customizable Zenodo

The screenshot shows the 'New upload' page in InvenioRDM. At the top, there is a search bar and a user profile for 'lars.holm.nielsen@cern.ch'. The main section is titled 'New upload' and includes a 'Files' dropdown menu. Below this, there is a 'Metadatum-only record' section with a 'Storage available' indicator showing '0 out of 100 files' and '0 bytes out of 10.00 Gb'. A dashed box contains a 'Drag and drop file(s)' area and an 'Upload files' button. A warning message states: 'File addition, removal or modification are not allowed after you have published your upload.' To the right, there are buttons for 'Save draft', 'Preview', 'Publish', and 'Delete'. Below the upload area, there is a 'Basic information' section with a 'DOI' field and a 'Resource type' dropdown. The 'Visibility' section is set to 'Public' for both 'Full record' and 'Files only'. A note at the bottom right says 'Public: The record and files are publicly accessible.'

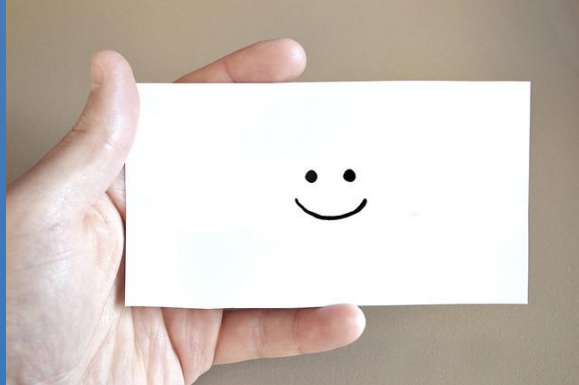
The 'Add creator' dialog box is overlaid on the main interface. It features the ROR logo in the top right corner. The 'Person' radio button is selected. The 'Family name' field contains 'Nielsen' and the 'Given name(s)' field contains 'Lars Holm'. The 'Name identifiers' field contains '0000-0001-8135-3489'. The 'Affiliations' dropdown menu is open, showing 'CERN' as the selected option, with 'Add CERN' and 'European Organization for Nuclear Research (CERN)' as other options. At the bottom, there are buttons for 'Cancel', 'Save and add another', and 'Save'.

Zenodo: 1 petabyte of data!

How???



We didn't have to!



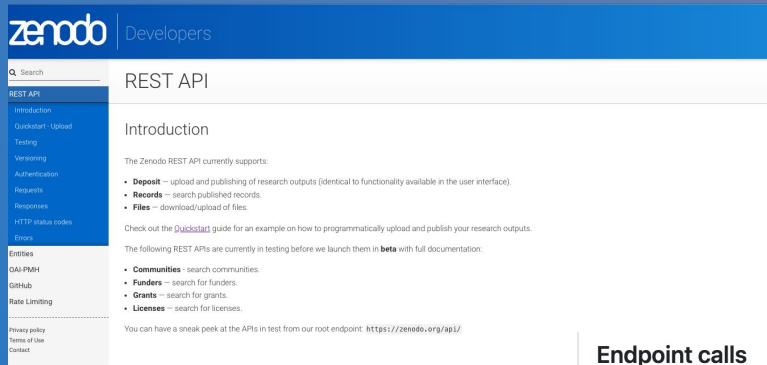
# Why is this important?

- Save data for long-term
- Honour the trust of our users
- Ensure no data loss!
- Minimal downtime

# Solution

Start as early as possible!

# Feature gap analysis



The screenshot shows the Zenodo REST API documentation page. The header includes the Zenodo logo and 'Developers'. A search bar is present. The main content area is titled 'REST API' and 'Introduction'. It lists supported features: Deposit (upload/publishing), Records (search), and Files (download/upload). It also mentions a Duplicate guide and lists REST APIs currently in testing: Communities, Funders, Grants, and Licenses. A URL for testing the APIs is provided: <https://zenodo.org/api/>. A sidebar on the left contains navigation links for REST API, Introduction, Quickstart - Upload, Testing, Versioning, Authentication, Requests, Responses, HTTP status codes, Errors, Entities, OAI-PMH, GitHub, Rate Limiting, Privacy policy, Terms of Use, and Contact.

## Endpoint calls

I processed all access logs from [redacted] user in 2022 and got the following results for:

- total api calls: 29135
- total old files api calls: 519
- total new files api calls: 136
- `oa1d` endpoint calls: 2

New files API endpoint : `/api/files/<bucker_id>`

Old files API endpoint : `/api/deposit/depositions/<depid>/files`

### Conclusion

In 2022 they still used the old files API's endpoint more than the new one.



# Feature gap analysis

# The following routes can be redirected in this task

Zenodo /record/<id>

RDM /records/<id>

```

'/deposit/<pid_value>',
'/schemas/<string:schema_path>',
'/record/<pid_value>',
'/record/<pid_value>?record_id_class="zenodo.modules.records.api:ZenodoRecord":pid_value',
'/communities/<string:community_id>/curate',
'/communities/<string:community_id>/about',
'/communities/<string:community_id>/edit',
'/record/<pid_value>/preview/<path:filename>',
'/record/<pid_value>/formats',
'/record/<pid_value>/export/<any(csl, hx, xn, schemaorg_jsonld, xm, dcite, '
'ef, json, xd, xe, dcite3, geojson, dcite4, hm, dcat, cp, xw):format>',
'/record/<pid_value>/thumb<thumbnail_size>',
'/record/<pid_value>/files/<path:filename>',
'/communities/<string:community_id>/search',
'/deposit',
'/favicon.ico',
'/deposit/new',

```

## Successful migrations

The following terms are migrated and fully tested.

```

{
  "resource_type.test_subtype": "metadata.resource_type.props.subtype",
  "resource_type.type": "metadata.resource_type.props.type",
  "access_right": "access.status",
  "alternate_identifier": "metadata.identifiers.identifier",
  "alternate_scheme": "metadata.identifiers.scheme",
  "communities": "parent.communities.ids",
  "conceptdoi": "parent.id",
  "created": "created",
  "creators.affiliation": "metadata.creators.affiliations.name",
  "description": "metadata.description",
  "doi": "pids.doi.identifier",
  "embargo": "access.embargo.until",
  "grants.code": "metadata.funding.award.number",
  "grants.title": "metadata.funding.award.title.en",
  "grants.funder.doi": "metadata.funding.id",
  "grants.funder.name": "metadata.funding.funder.name",
  "keywords": "metadata.subjects.subject",
  "language": "metadata.languages.id",
  "license.identifier": "metadata.rights.id",
  "license.url": "metadata.rights.props.url",
  "publicationdate": "metadata.publication_date",
  "publication_date": "metadata.publication_date",
  "recid": "id",
  "related_identifier": "metadata.related_identifiers.identifier",
  "related_scheme": "metadata.related_identifiers.scheme",
  "related_relation": "metadata.related_identifiers.relation_type.id",
  "title": "metadata.title",
  "type": "metadata.resource_type.props.type",
  "version": "versions.index",
}

```

Zenodo doi search  
../api/records?q=doi:<doi>

RDM doi search  
../api/records?  
q=pids.identifier.doi:<doi>

# Data analysis

```
zenodo=> SELECT pg_size_pretty( pg_database_size('zenodo') );
pg_size_pretty
-----
264 GB
(1 row)
```

Amount of data

rdm\_records\_metadata

- > Columns
- ▼ Constraints (3)
  - fk\_rdm\_records\_metadata\_bucket\_id\_files\_bucket
  - fk\_rdm\_records\_metadata\_parent\_id\_rdm\_parents\_metadata
  - pk\_rdm\_records\_metadata

Records -> Files

Data dependencies

# Tooling



# Tooling - existing code

Method	Time 50k	Time 3M	Setup	Comment
services	7m30s	7h30m	Running from one celery worker with 10 threads. Disabled indexing.	Creating a draft takes only one third of the time.

# Tooling - ETL

*Extract, Transform, Load is a process where data is extracted, transformed (cleaned, sanitized, normalized) from a source and loaded in a target destination.*



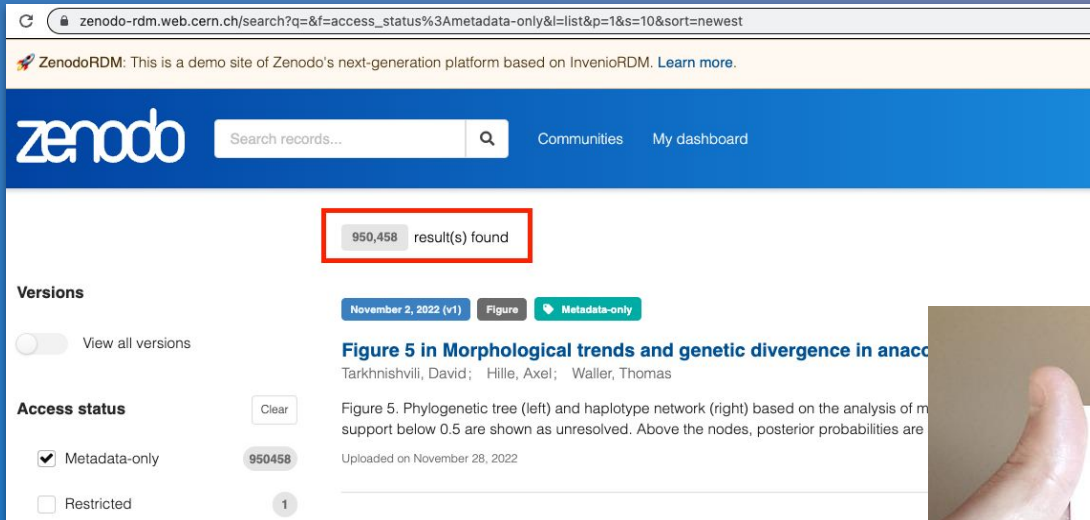
# Tooling - Invenio-RDM-Migrator

```
! streams.yaml x
site > zenodo_rdm > migrator > ! streams.yaml
1  data_dir: /path/to/data
2  tmp_dir: /path/to/tmp
3  cache_dir: /path/to/cache
4  log_dir: /path/to/log
5  db_uri: postgresql://zenodo:zenodo@localhost:5432/zenodo
6  users:
7    - extract:
8      - filepath: /path/to/users.jsonl
9  files:
10   - existing_data: True
11  communities:
12   - extract:
13     - filepath: /path/to/communities.jsonl
14  records:
15   - extract:
16     - filepath: /path/to/records.jsonl
17   - load:
18     - versioning: false
19  drafts:
20   - extract:
21     - filepath: /path/to/deposits.jsonl
22  requests:
23   - extract:
```

```
RecordStreamDefinition = StreamDefinition(
    - - - name="records",
    - - - extract_cls=JSONLExtract,
    - - - transform_cls=ZenodoRecordTransform,
    - - - load_cls=RDMRecordCopyLoad,
)
""""ETL stream for Zenodo to RDM records.""""
```

Invenio-rdm-migrator: <https://github.com/inveniosoftware/invenio-rdm-migrator>

# Milestone 1 - migrate 1m records



The screenshot shows the Zenodo RDM website interface. The URL in the browser is `zenodo-rdm.web.cern.ch/search?q=&f=access_status%3Ametadata-only&l=list&p=1&s=10&sort=newest`. The page header includes the Zenodo logo, a search bar, and navigation links for "Communities" and "My dashboard". A search result summary shows "950,458 result(s) found", with this text highlighted by a red box. Below the search results, there are filters for "Versions" (November 2, 2022 (v1), Figure, Metadata-only) and "Access status" (Metadata-only, Restricted). The main content area displays the title "Figure 5 in Morphological trends and genetic divergence in anaco" and the authors "Tarkhnishvili, David; Hille, Axel; Waller, Thomas".

- 1m records - 1:30 hours!
- **Disclaimer:** Only required record metadata



# Next steps - add more metadata

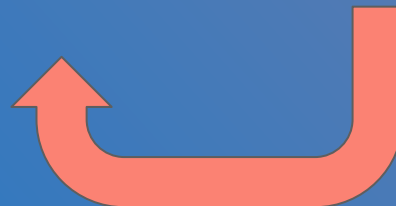




# Try once more



```
744         error => $quote['sort_order'],
745     });
746     }
747     }
748     }
749     $sort_order = array();
750     }
751     foreach ($quotes as $key => $value) {
752         $sort_order[$key] = $value['sort_order'];
753     }
754     array_multisort($sort_order, SORT_ASC, $quotes);
755     $this->session->data['lpa']['shipping_methods'] = $quotes;
756     $this->session->data['lpa']['address'] = $address;
757     if (empty($quotes)) {
758         $icon['error'] = $this->language->get('
759             error_no_shipping_methods');
760     } else {
761         $icon['quotes'] = $quotes;
762     }
763     }
764     }
765     if (isset($this->session->data['lpa']['shipping_method'])) &&
766         empty($this->session->data['lpa']['shipping_method']) &&
767         isset($this->session->data['lpa']['shipping_method']['code'])
768     ) {
769         $icon['selected_shipping_method'] = $this->session->data['lpa']['shipping_method']['code'];
770     }
771     }
772     }
773     }
774     }
775     }
776     }
777     }
778     }
779     }
780     }
781     }
782     }
783     }
784     }
785     }
786     }
787     }
788     }
789     }
790     }
791     }
792     }
793     }
794     }
795     }
796     }
797     }
798     }
799     }
800     }
801     }
802     }
803     }
804     }
805     }
806     }
807     }
808     }
809     }
810     }
811     }
812     }
813     }
814     }
815     }
816     }
817     }
818     }
819     }
820     }
821     }
822     }
823     }
824     }
825     }
826     }
827     }
828     }
829     }
830     }
831     }
832     }
833     }
834     }
835     }
836     }
837     }
838     }
839     }
840     }
841     }
842     }
843     }
844     }
845     }
846     }
847     }
848     }
849     }
850     }
851     }
852     }
853     }
854     }
855     }
856     }
857     }
858     }
859     }
860     }
861     }
862     }
863     }
864     }
865     }
866     }
867     }
868     }
869     }
870     }
871     }
872     }
873     }
874     }
875     }
876     }
877     }
878     }
879     }
880     }
881     }
882     }
883     }
884     }
885     }
886     }
887     }
888     }
889     }
890     }
891     }
892     }
893     }
894     }
895     }
896     }
897     }
898     }
899     }
900     }
901     }
902     }
903     }
904     }
905     }
906     }
907     }
908     }
909     }
910     }
911     }
912     }
913     }
914     }
915     }
916     }
917     }
918     }
919     }
920     }
921     }
922     }
923     }
924     }
925     }
926     }
927     }
928     }
929     }
930     }
931     }
932     }
933     }
934     }
935     }
936     }
937     }
938     }
939     }
940     }
941     }
942     }
943     }
944     }
945     }
946     }
947     }
948     }
949     }
950     }
951     }
952     }
953     }
954     }
955     }
956     }
957     }
958     }
959     }
960     }
961     }
962     }
963     }
964     }
965     }
966     }
967     }
968     }
969     }
970     }
971     }
972     }
973     }
974     }
975     }
976     }
977     }
978     }
979     }
980     }
981     }
982     }
983     }
984     }
985     }
986     }
987     }
988     }
989     }
990     }
991     }
992     }
993     }
994     }
995     }
996     }
997     }
998     }
999     }
1000    }
```



# A new hope

ZenodoRDM: This is a demo site of Zenodo's next-generation platform based on InvenioRDM. [Learn more.](#)

zenodo Search records... Communities My dashboard

2,862,188 result(s) found

**Versions**

View all versions

April 3, 2023 (v1) Journal article Open

**Secure Group Key Management for Group Communication in IoMT Envir**  
PramanandaPerumal S. Bhuvaneswari\*, T.

**Access status**

Internet of Things (IoT) role is inevitable in today's scenario. The secure communication between IoT devices using an encryption scheme (AES and RSA) to ensure secure group communication. The doctors and devices

Migrated latest versions	Total latest versions	Missing	Success rate
2862188	2890136	28889	99.9%

- Disclaimer
  - Records with full metadata
  - Communities
  - Users
  - Files

- 99,9% integrated records
- 1h30 loading time

Present - 3 months from final migration



# What did we learn so far?

- Migration is a continuous process
- Progress is not linear
- **Start as early as possible**