CSC

ICT Solutions for
Brilliant Minds

FAIR data. Citation and services.

Introduction to RDM part II

Jessica Parland-von Essen

https://orcid.org/0000-0003-4460-3906

9.5.2019

CSC's Current Computing Services available for research and education

openstack.

RED HAT OPENSHIFT

kubernetes

High Performance

- Sisu
- Massive parallelism
- Fast interconnect

Capacity

- Taito
- General use
- Large memory
- >100 applications

Accelerated

- Taito extension
- Visualization
- Special codes
- Nvidia GPU

Cloud

- IaaS: cPouta + ePouta (OpenStack)
- PaaS: Rahti (OpenShift / Kubernetes)

Hosting

- Kajaani
- Espoo
- Efficient and secure datacenters
- Virtual and physical servers

Storage Services: Fast parallel storage, Object Storage, Archiving

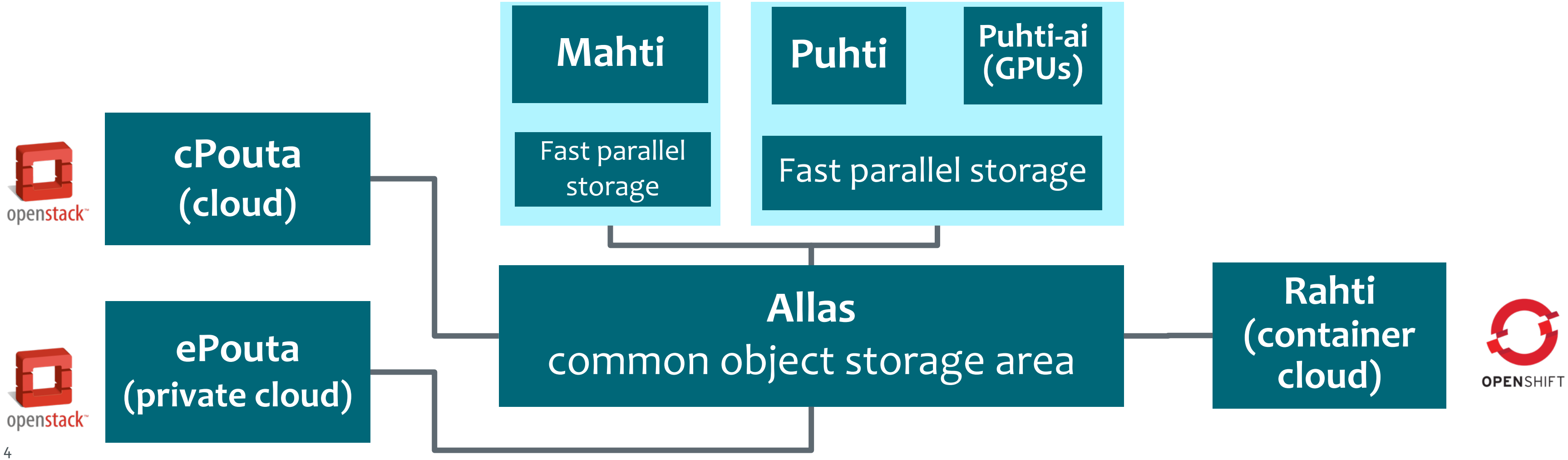
What is CSC upcoming computing and data environment

- Something familiar… but much more powerful
  - **Puhti** - Supercomputer with Intel CPUs
- Something a bit different
  - **Puhti-ai** – Supercomputer with GPUs
  - **Mahti** - Supercomputer with AMD CPUs
- Something new
  - **Allas** – Large storage system with access and usage possibilities beyond traditional filesystem.

The new Finnish research infrastructure for data management and computing

Balanced HPC ecosystem for supporting the six drivers

Heterogeneous, workload-optimized node architecture, support for complex workflows, datasets-as-a-service and containerization



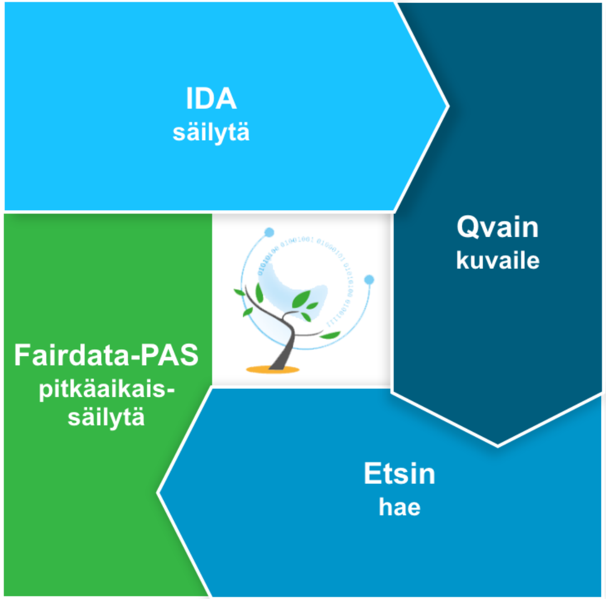
Allas – object storage: what it is for?

- Allas is new storage service for all computing and cloud services
  - Meant for data during project lifetime
- Data can be stored and retrived directly from anywhere in Internet
  - CSC supercomputers
  - Local workstation
  - Measurement devices
  - …
  - At simplest, the web browser is enough
- Easy sharing of data outside project
  - Selected data can be shared publicly to Internet
- Arbitrary metadata can be added to data

In customer use end of Q2

Fairdata.fi

- National integrated services for storing, describing and sharing and preserving research data
- Provided by MinEdu
- Produced by CSC and National Library of Finland
- **Make your data safe , documented and citable**
  - **IDA** – Research data storage service
  - **ETSIN** – Research data finder
  - **QVAIN** – Research dataset metadata tool
  - **FAIRDATA-PAS** – Digital preservation for research data



Services for finding research data

- With ETSIN and B2FIND SERVICE you can easily find data for your research
- AVAA is a publication platform where you can seek and download open research data.
- THE LANGUAGE BANK OF FINLAND is a service entity for researchers who use text and speech corpora. The basic use is free of charge to researchers and students.
- PAITULI is a download service of Finnish spatial data. The service contains spatial data that is important in research and education. It deviates from many other similar services as it also contains historical year versions.
- ELIXIR offers services for medical and bioinformatics research.
- VIPUNEN is a statistics service provided by the Finnish National Board of Education where you can find information on, for example, education in a number of educational sectors, research conducted in higher education institutions and the educational structure of the population.
- Coming: the Research Information hub: research.fi

F INDABLE
A CCESSIBLE
I NTEROPERABLE
R EUSABLE

CITABLE

HOW TO GET THERE?

Roadmap

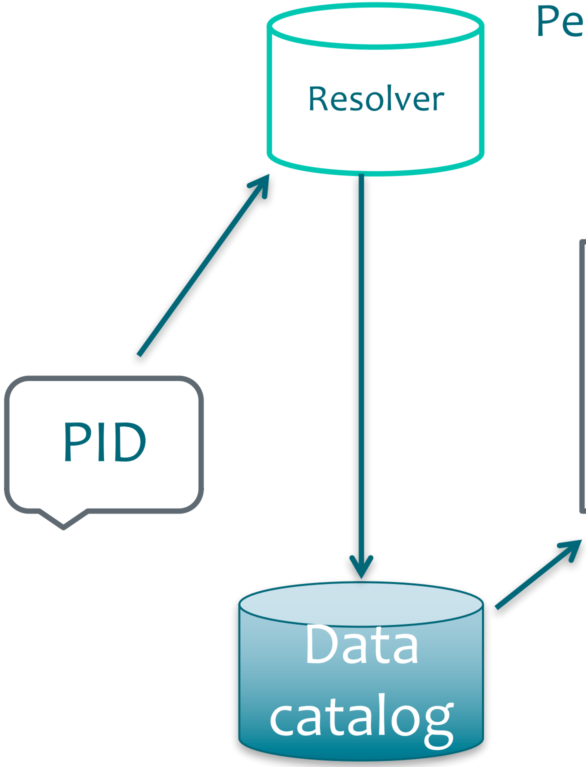
Stakeholders

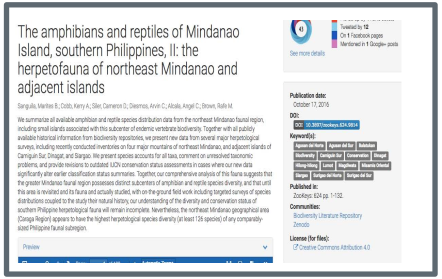
FCRD
https://www.fcrd.fi/data-citation/

Data citation

- Recommendation by Finnish Committee for Research Data
- Research data should be FAIR
- Data should have: ***Creator, title, publisher, publication time, identifier***
- Recommended additional information: *Version, resource type, copyright status*

Persistent identifiers are necessary for science





The amphibians and reptiles of Mindanao Island, southern Philippines, II: the herpetofauna of northeast Mindanao and adjacent islands

Sanguila, Marites B.; Cobb, Kerry A.; Siler, Cameron D.; Diesmos, Arvin C.; Alcala, Angel C.; Brown, Rafe M.

We summarize all available amphibian and reptile species distribution data from the northeast Mindanao faunal region, including small islands associated with this subcenter of endemic vertebrate biodiversity. Together with all publicly available historical information from biodiversity repositories, we present new data from several major herpetological surveys, including recently conducted inventories on four major mountains of northeast Mindanao, and adjacent islands of Camiguin Sur, Dinagat, and Siargao. We present species accounts for all taxa, comment on unresolved taxonomic problems, and provide revisions to outdated IUCN conservation status assessments in cases where our new data significantly alter earlier classification status summaries. Together, our comprehensive analysis of this fauna suggests that the greater Mindanao faunal region possesses distinct subcenters of amphibian and reptile species diversity, and that until this area is revisited and its fauna and actually studied, with on-the-ground field work including targeted surveys of species distributions coupled to the study their natural history, our understanding of the diversity and conservation status of southern Philippine herpetological fauna will remain incomplete. Nevertheless, the northeast Mindanao geographical area (Caraga Region) appears to have the highest herpetological species diversity (at least 126 species) of any comparably-sized Philippine faunal subregion.

Tweeted by **12**
On **1** Facebook pages
Mentioned in **1** Google+ posts
See more details

Publication date:
October 17, 2016

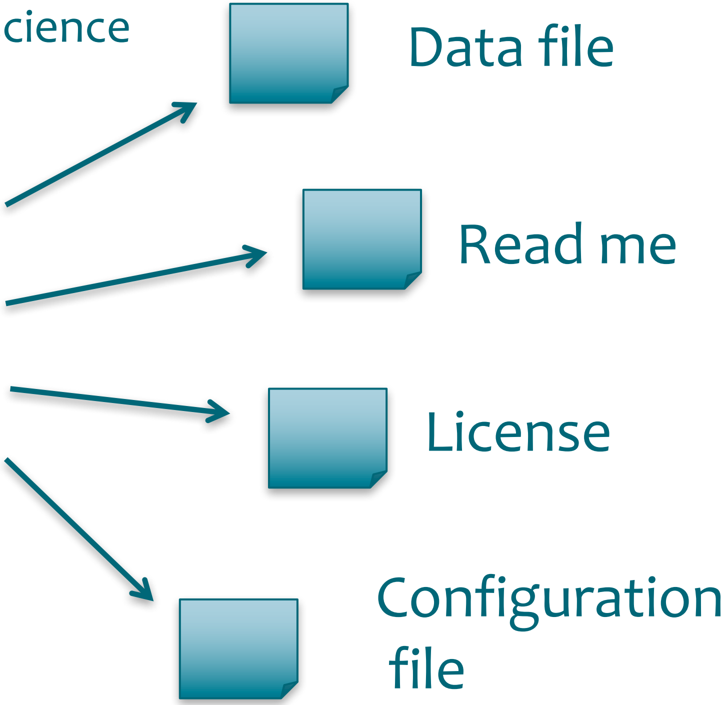
DOI:
DOI 10.3897/zookeys.624.9814

Keyword(s):
Agusan del Norte, Agusan del Sur, Balatukan, Biodiversity, Camiguin Sur, Conservation, Dinagat, Hilong-hilong, Lumot, Magdiwata, Misamis Oriental, Siargao, Surigao del Norte, Surigao del Sur

Published in:
ZooKeys: 624 pp. 1-132.

Communities:
Biodiversity Literature Repository
Zenodo

License (for files):
Creative Commons Attribution 4.0



Metadata describes your data so that it is

- discoverable
- be verifiable
- can be re used
- can be accessed for years to come
- help others understand the data
- can be preserved for the future

Jørgen Stamp. Published on www.digitalbevaring.dk
Licensed under CC BY 2.5 DK

Metadata describes your data, its

- purpose (WHY)
- origin (HOW)
- time references (WHEN)
- geographic location (WHERE)
- creators (WHO)

AND

- access conditions and terms of use (USE)

Jørgen Stamp. Published on www.digitalbevaring.dk
Licensed under CC BY 2.5 DK

Research data and metadata

- The FAIRsharing.org contains more than 1200 standards
- Different fields of science and different kinds of data need different kinds of metadata
- Always try to use or conform to existing standards like schemas or vocabularies as much as possible
- Keep track of master metadata and keep it separate from aggregated metadata or self-reported information
- Choose the right format for each context
- Schema.org helps web users

CC-BY Jer Thorp
Flickr: blprnt

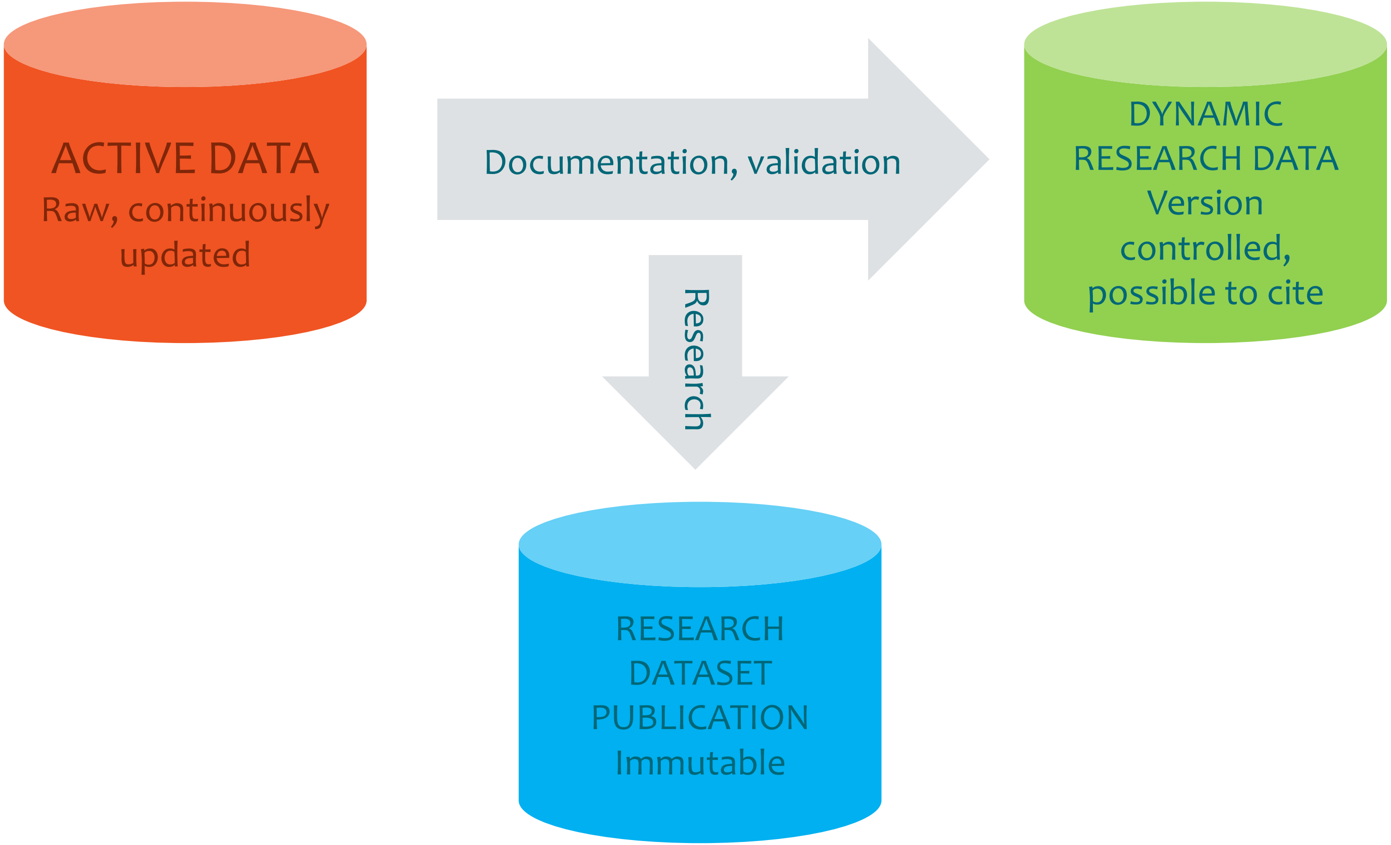
FAIRsharing.org
standards, databases, policies

Recommended and optional properties in DataCite metadata

- Subject (with sub-properties)
- Contributor (with sub-properties)
- Date (with sub-property)
- Language
- Alternate identifier (with sub-property)
- Related identifier (with sub-properties)
- Size

- Format
- Version
- Rights
- Language
- Description (with sub-property)
- Geo location (with sub-properties)
- Funding reference (with sub-properties)

Research Data Types



ACTIVE DATA
Raw, continuously updated

Documentation, validation

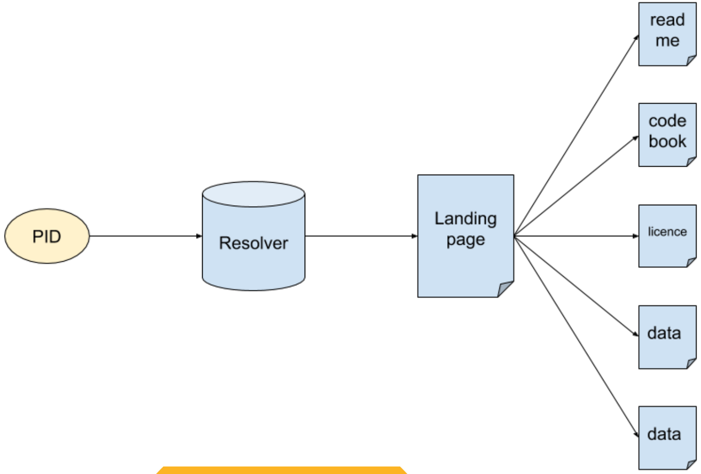
DYNAMIC RESEARCH DATA
Version controlled, possible to cite

Research

RESEARCH DATASET PUBLICATION
Immutable

Persistent identifiers

IMMUTABLE DATASETS



DYNAMIC DATASETS

a) Cite a specific slice or subset (the set of updates to the dataset made during a particular period of time or to a particular area of the dataset).

b) Cite a specific snapshot (a copy of the entire dataset made at a specific time).

c) Cite the continuously updated dataset, but add Access Date and Time to the citation. (Does not necessarily ensure reproducibility.)

d) Cite a query, time-stamped for re-execution against a versioned database.

FAIR

Findable

Accessible

Interoperable

Reusable

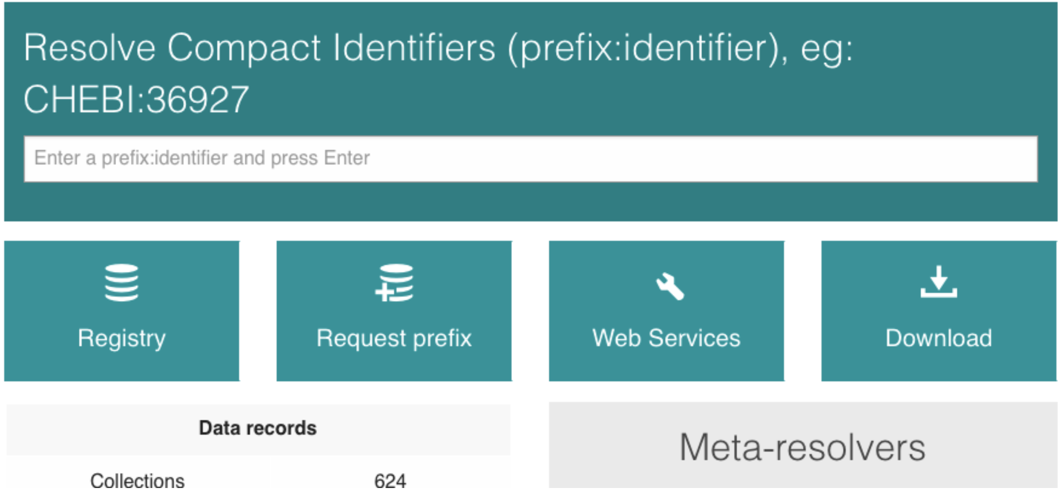
By Martin Grandjean - Own work : http://www.martingrandjean.ch/wp-content/uploads/2013/10/Graphe3.png, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=29364647

Persistent identifiers

- A good identifier is linked to
  - A landing page
  - Access information
  - Metadata
  - Tombstone page

Nano publications, linking data and compact identifiers

- https://www.go-fair.org/
- http://identifiers.org/

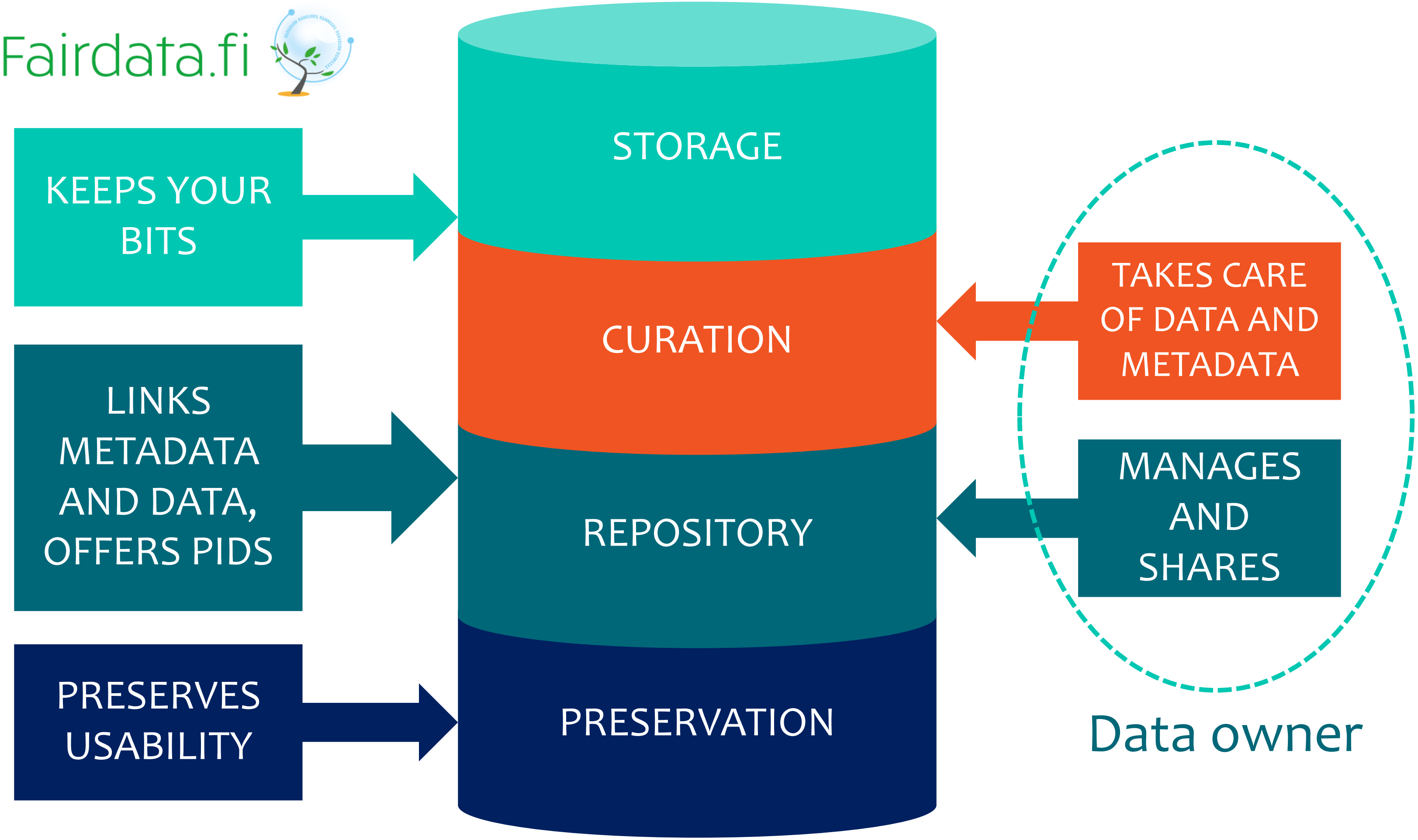


Resolve Compact Identifiers (prefix:identifier), eg: CHEBI:36927

Enter a prefix:identifier and press Enter

Data records	
Collections	624

Fairdata.fi



KEEPS YOUR BITS

STORAGE

LINKS METADATA AND DATA, OFFERS PIDS

CURATION

TAKES CARE OF DATA AND METADATA

REPOSITORY

MANAGES AND SHARES

PRESERVES USABILITY

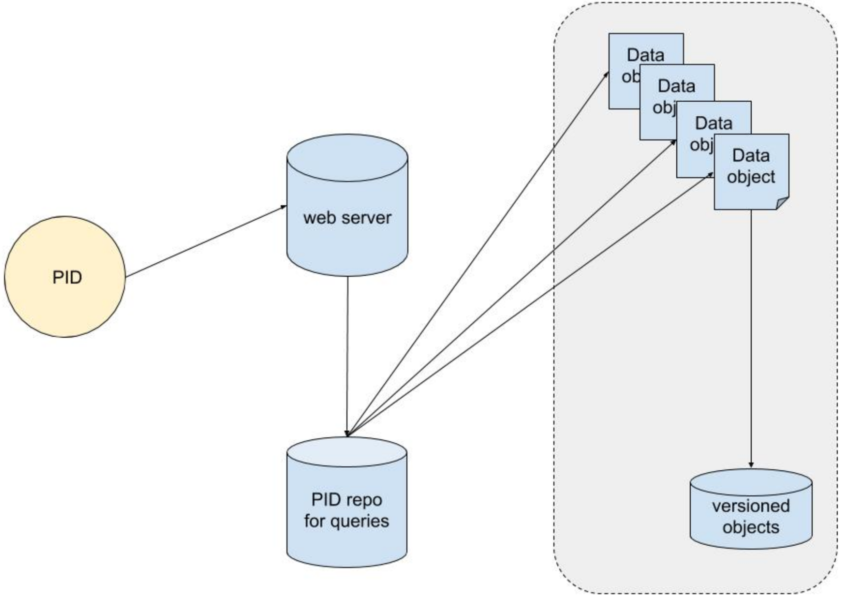
PRESERVATION

Data owner

CEOS Cumulative Research Data

- Suitable for on going campaigns
  1. Consistent format
  2. Coherent datasets, configurations
  3. For research
  4. Data might be added
  5. If data is retracted or changed a new version with new PID is created
- Best practice http://bit.ly/2Gjrknj

Picture by Tumi-1983 (Own work) [Public domain], via Wikimedia Commons



Dynamic and growing datasets

URN allows use of fragments

Avoid PID inflation

Consider costs and sustainability

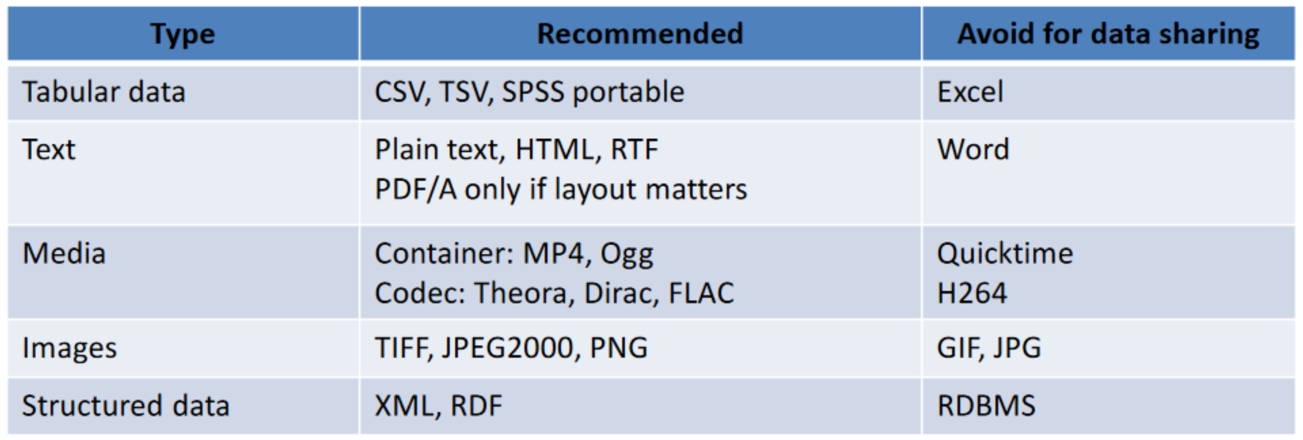
Ad hoc creation rather than automatic minting and allocation?

What defines a good repository?

- Established
- Good metadata

- Curation and preservation
- Open and machine readable
- Re3data.org

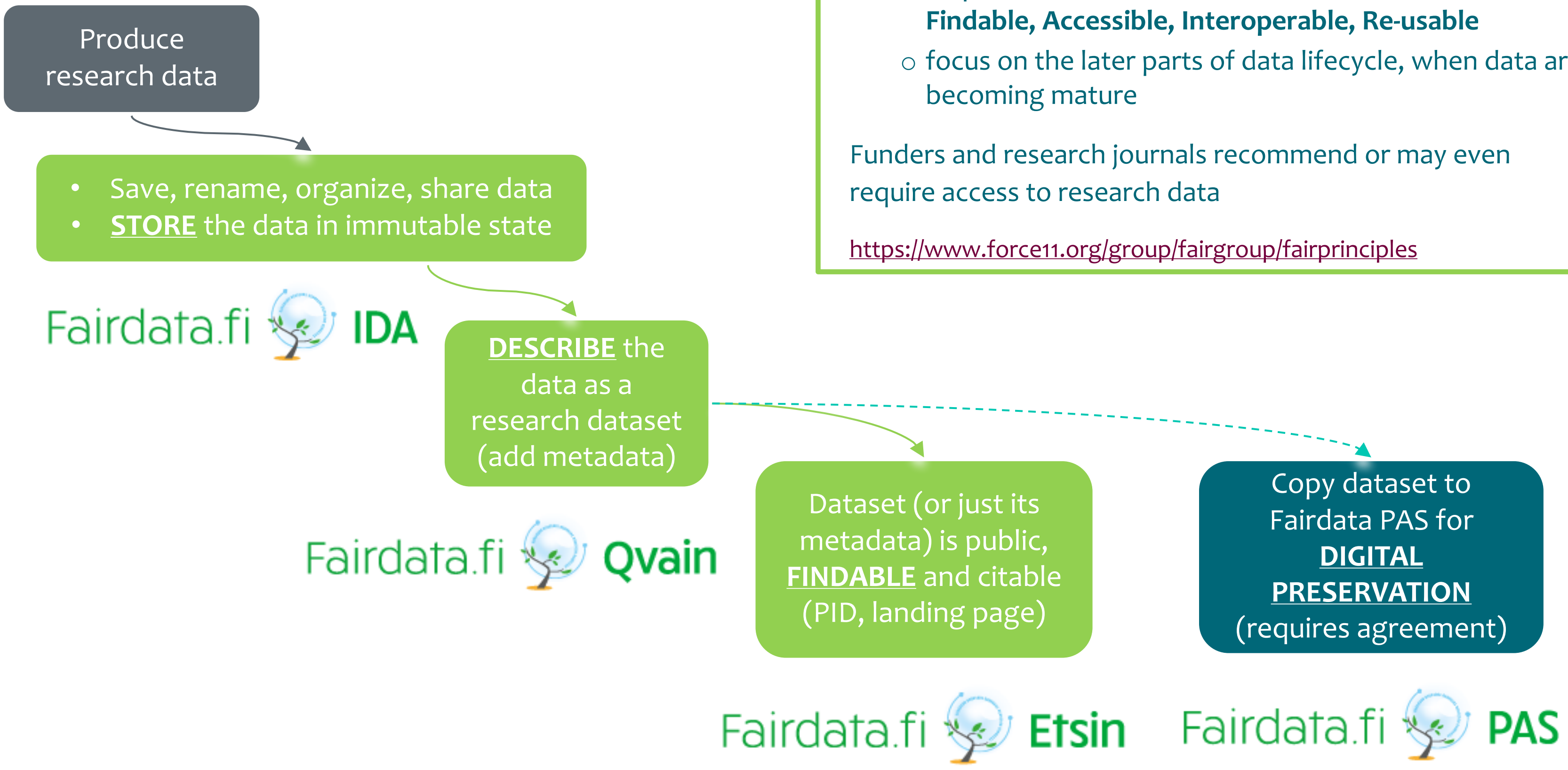
CORE TRUST SEAL



Type	Recommended	Avoid for data sharing
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

A PID is a Promise

1 What do they do? (1/2)



Produce research data

- Save, rename, organize, share data
- **STORE** the data in immutable state

Fairdata.fi IDA

DESCRIBE the data as a research dataset (add metadata)

Fairdata.fi Qvain

Dataset (or just its metadata) is public, **FINDABLE** and citable (PID, landing page)

Fairdata.fi Etsin

Copy dataset to Fairdata PAS for **DIGITAL PRESERVATION** (requires agreement)

Fairdata.fi PAS

Services, which

- help to make research data and related metadata **Findable, Accessible, Interoperable, Re-usable**
- focus on the later parts of data lifecycle, when data are becoming mature

Funders and research journals recommend or may even require access to research data

https://www.force11.org/group/fairgroup/fairprinciples

How to get access?

- **Your Haka user ID is your access to more than 160 services at CSC.**
  - Web based services ready to use
  - Register to get a personal CSC user account
  - If your organization does not have Haka, contact servicedesk@csc.fi
- **For IDA: through the contact persons in your own organization**
  - Project supervisor applies for resources and services, and can invite other people to projects
  - IDA has contact persons in customer organizations
    - https://www.fairdata.fi/en/ida/becoming-an-ida-user/
- **Customer service**
  - Instructions https://research.csc.fi/accounts-and-projects
  - Support and guidance **servicedesk@csc.fi**
  - Weekdays 8.30–16.00.

Fairdata.fi IDA

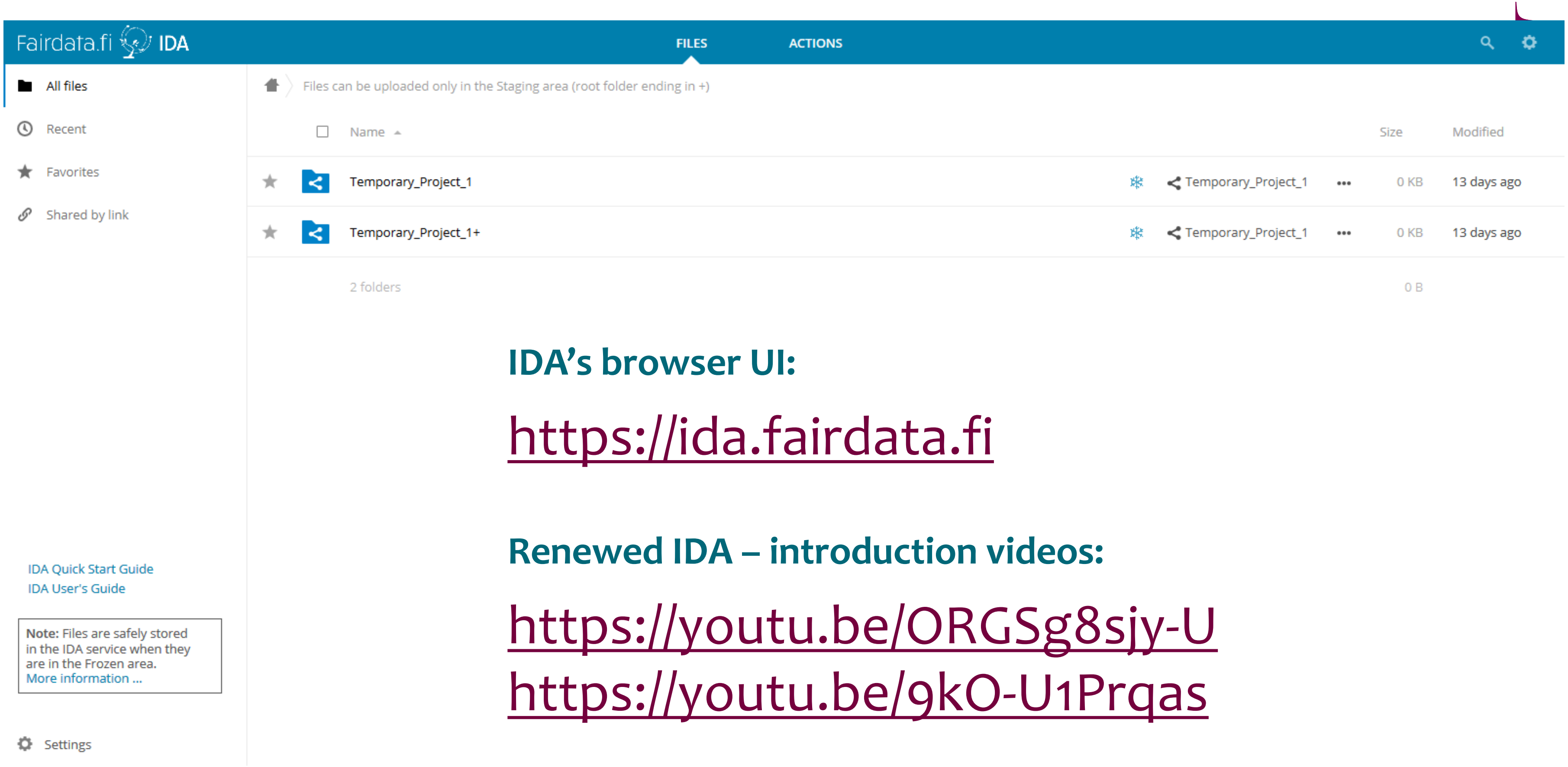
IDA – Research data storage service (renewed in summer 2018)

- Project based storage space for storing and sharing stable research data
- Data can be described as a dataset and published via additional Fairdata services
  - Qvain and Etsin will be in use soon, see https://www.fairdata.fi/
- Accessible from the Internet
- Offered free of charge to researchers from Finnish higher education institutions, state research institutes and those funded by the Academy of Finland
  - Storage space granted based on application by organization (from 1 GB to ~100 TB)
  - Currently ~240 projects, 640 TBs of stored data

User interfaces

- Graphical browser UI (based on Nextcloud): http://ida.fairdata.fi/
- command line tools available in GitHub and on Taito: https://www.fairdata.fi/en/ida/user-guide/#command-line-tools

IDA is one route to digital preservation (Fairdata PAS)



Files can be uploaded only in the Staging area (root folder ending in +)

☐ Name ▲			Size	Modified
★ Temporary_Project_1	❄	Temporary_Project_1	0 KB	13 days ago
★ Temporary_Project_1+	❄	Temporary_Project_1	0 KB	13 days ago
2 folders			0 B	

IDA Quick Start Guide
IDA User's Guide

Note: Files are safely stored in the IDA service when they are in the Frozen area. More information ...

IDA's browser UI:

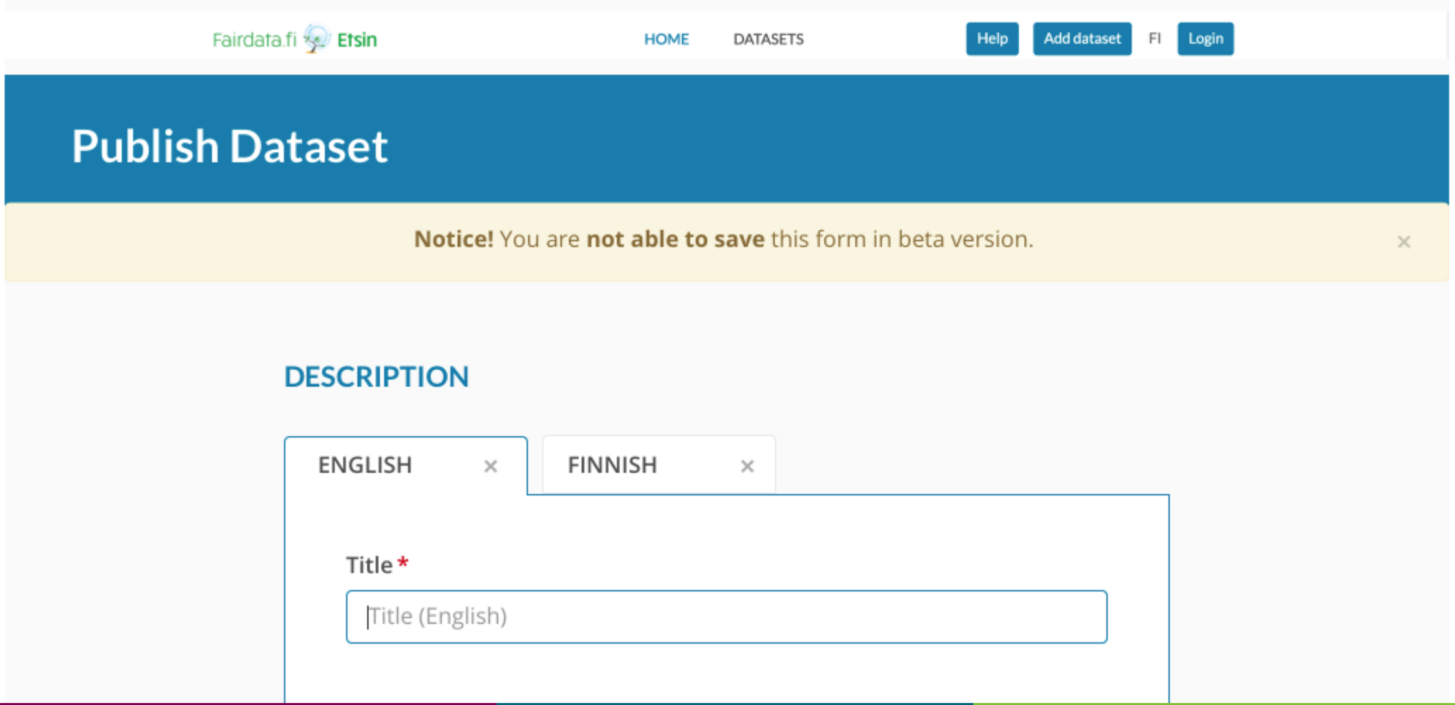
https://ida.fairdata.fi

Renewed IDA – introduction videos:

https://youtu.be/ORGSg8sjy-U
https://youtu.be/9kO-U1Prqas

DEMO “Qvain”

- https://bit.ly/2VrJqi4



Publish Dataset

Notice! You are **not able to save** this form in beta version.

DESCRIPTION

Title *

Title (English)