

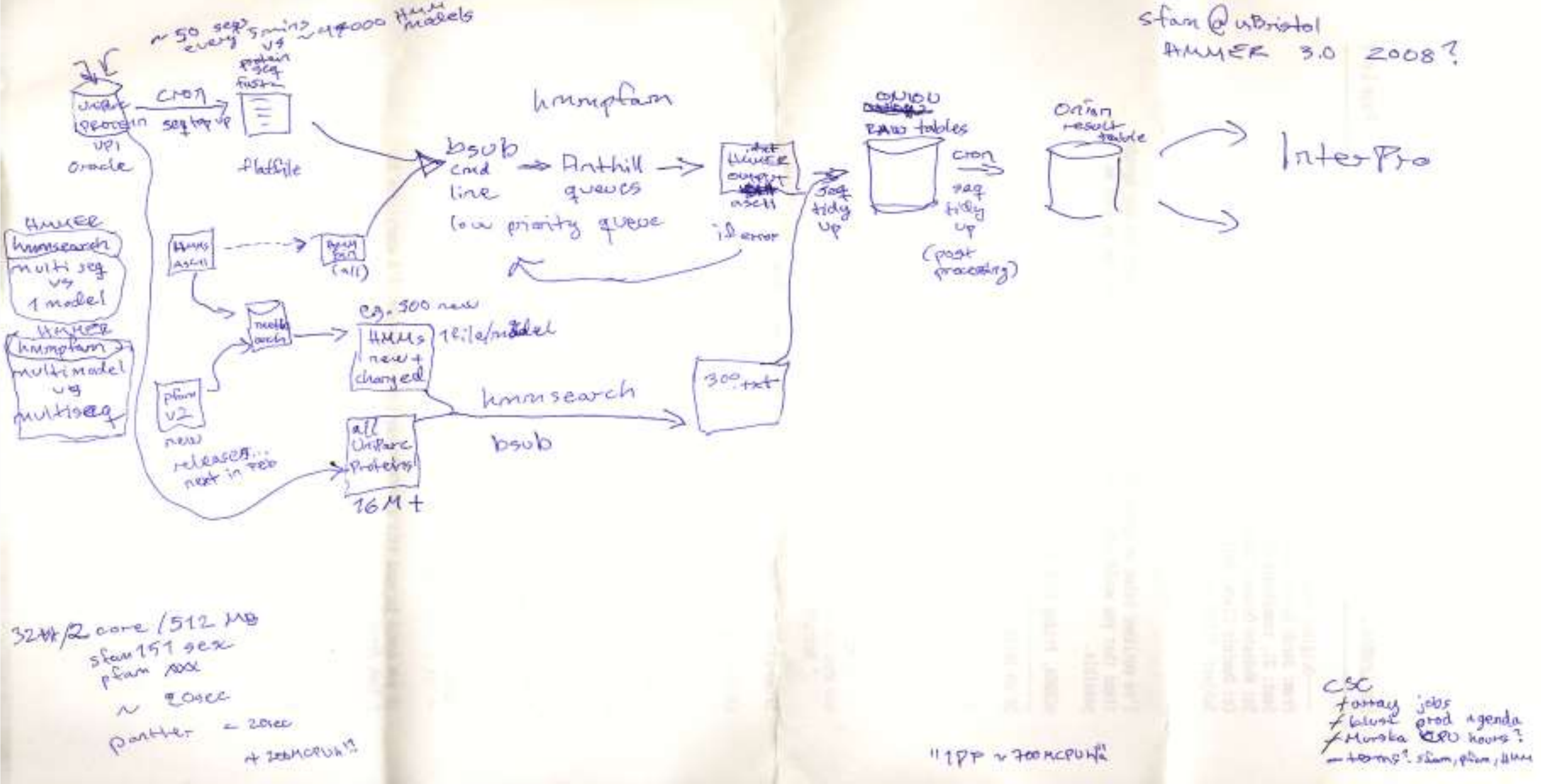
Towards FAIRness in research data

Per Öster, 3 October 2018

CSC – Suomalainen tutkimuksen, koulutuksen, kulttuurin ja julkishallinnon ICT-osaamiskeskus

Towards FAIR data!

Anyone expecting that an advertising company or a retailer of merchandise and media will take us there?



sfam @ uBristol
HMMER 3.0 2008?
~50 seqs every 5 mins vs ~10000 HMM models
protein seq fasta
cron
seq top up
UniParc protein
UPI Oracle
flatfile
hmmpfam
bsub cmd line
Anthill queues
low priority queue
HMMER output ascii
id error
seq tidy up
RAW tables
cron
seq tidy up
(post processing)
Onion result table
InterPro
HMMER
hmmsearch
multi seq vs 1 model
HMMER
hmmpfam
multimodel vs multiseq
HMMs ASCII
(all)
eg. 300 new
HMMs new + changed
1 file/model
pfam V2
new releases... next in Feb
all UniParc Proteins
16M+
hmmsearch
bsub
300.txt
32 bit/2 core/512 MB
sfam 151 sec
pfam xxx
~ 20sec
panther ~ 20sec
+ 200MCPUh!!
"1PP ~ 700 MCPUh"
CSC
farray jobs
blast prod agenda
Murska CPU hours?
terms? sfam, pfam, HMM

Data Generation

Sequencing centers

Data Archiving

EMBL-EBI

European Genome-phenome Archive

ENA

European Nucleotide Archive

EGA at CRG

Secure data access remote API
(GA4GH)

High speed encrypted data transfer

GridFTP/Globus/Aspera

Supporting sample logistics

BBMRI

EMBL-EBI BioSamples

Services and Coordination

- Federated Authentication
- Authorization
- Dataset registry
- Data transfer hub
- Policy and Legal Framework

Managing Access

Data Owner
Data Access Agreement
Data Access Committee
Data Request

Authorization Management Tools
(EGA and CSC REMS)

Secure Compute Clouds

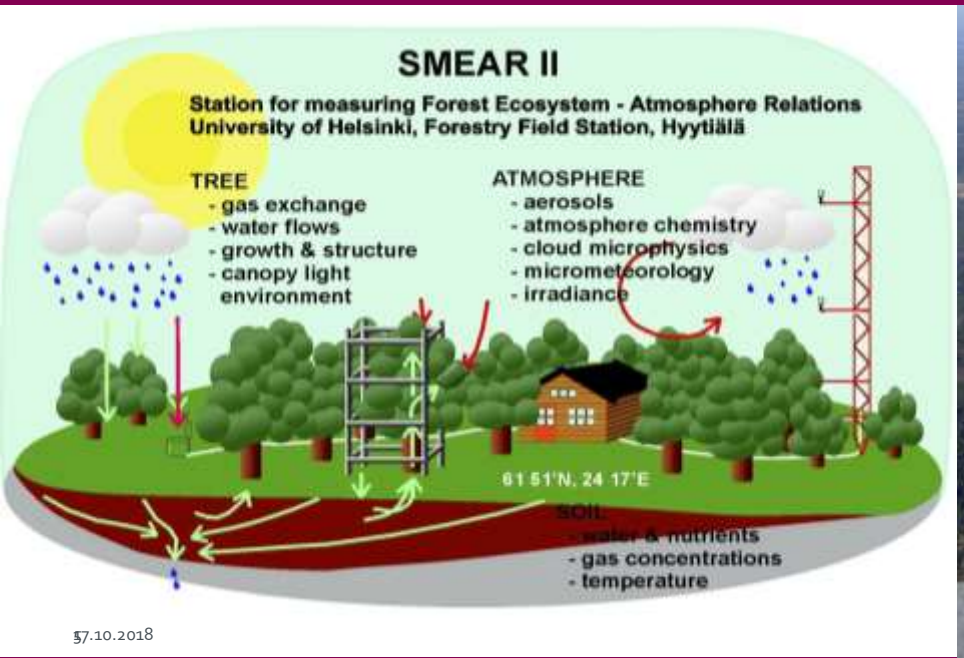
BSC

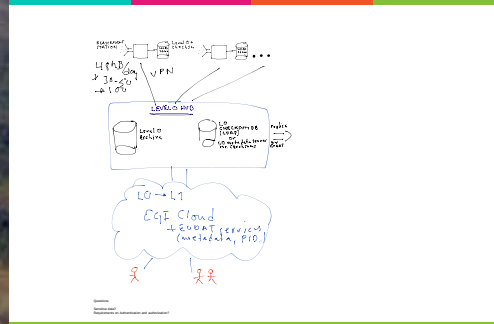
CSC

EMBL-EBI

Bringing users to data

From field measurements to open data





SMEAR data flow

Instrument

A/D conversion
unit conversion

Measuring PCs:
Raw data
at the stations

Field
documentation

File servers at
stations:
Raw data and
field diaries,
cal documents

Routine data processing =
(- unit conversion)
- calibration correction
- quality check, gapfilling
- averaging over space or time

Researchers,
Data processing
server

File servers
in Helsinki:
Raw and
intermediate
data,
documents,
scripts

Feedback on
data quality

ICOS, EBAS,...
databases:
Near real time
and processed data
outside UH

Metadata

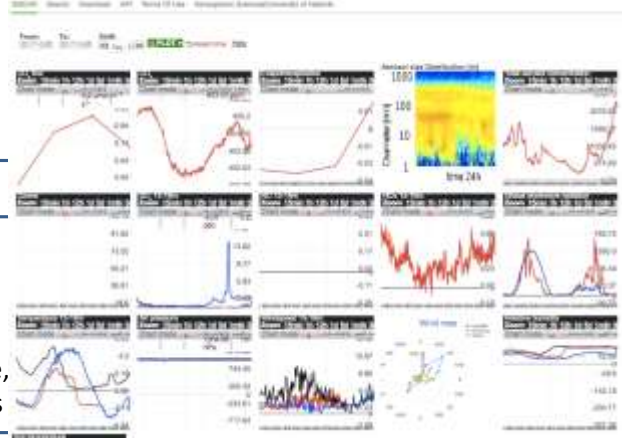
SMEAR database:
Processed data
in Helsinki

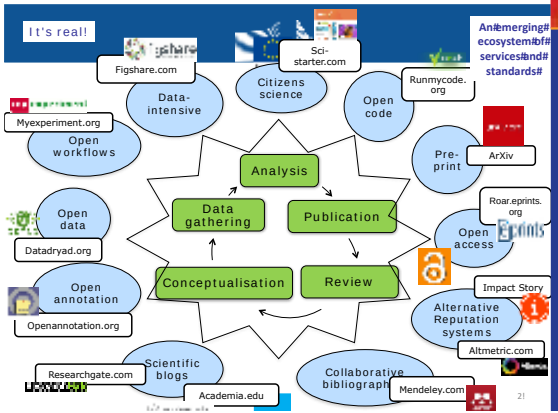
Metadata

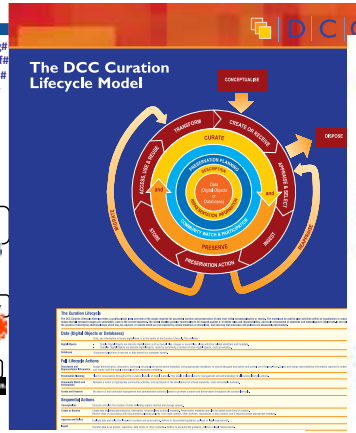
IDA (CSC data
service):
Raw data &
document archive,
database datasets

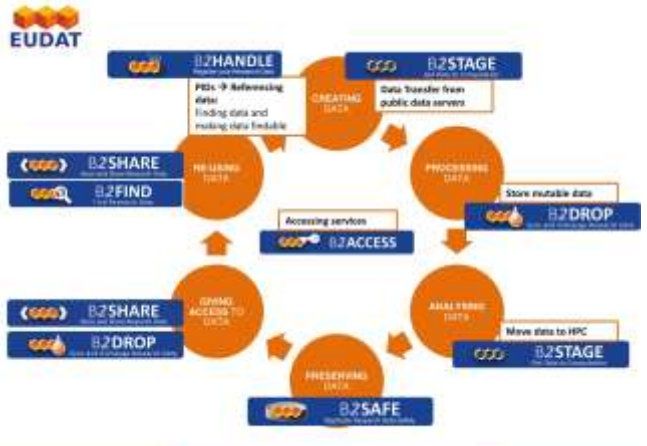
Metadata

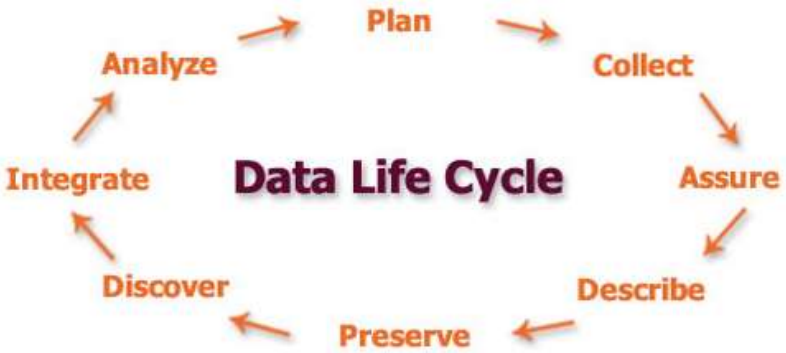
https://avaa.tdata.fi/web/smart/smear




It's real!
An#emerging#ecosystem#of#services#and#standards#
Figshare.com
Sci-starter.com
Citizens science
Runmycode.org
Open code
Myexperiment.org
Open workflows
Data-intensive
Pre-print
ArXiv
Analysis
Data gathering
Publication
Roar.eprints.org
Open data
Datadryad.org
Open access
Conceptualisation
Review
Impact Story
Open annotation
Openannotation.org
Alternative Reputation systems
Altmetrics.com
Researchgate.com
Scientific blogs
Academia.edu
Collaborative bibliography
Mendeley.com

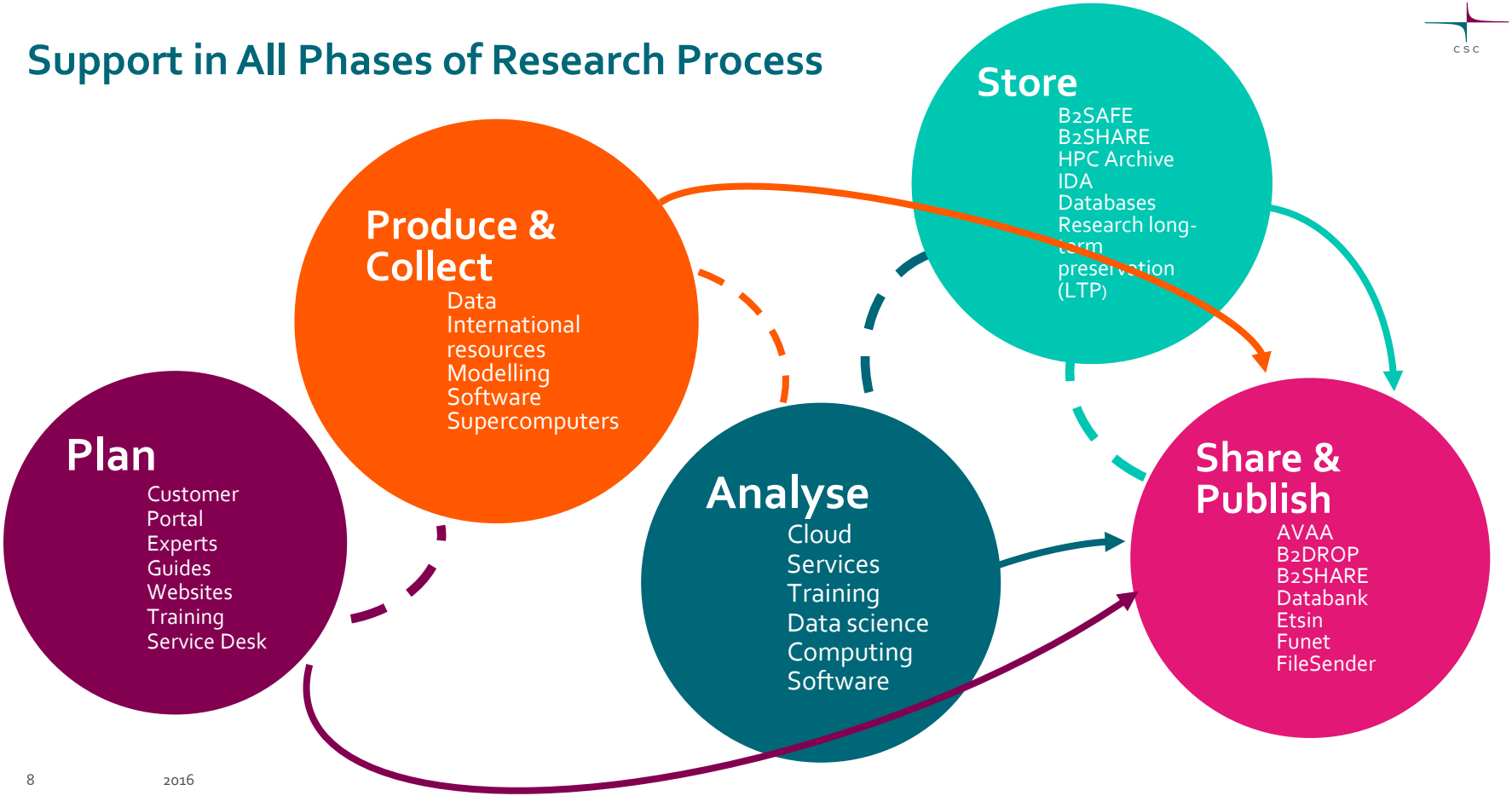

DCC
The DCC Curation Lifecycle Model

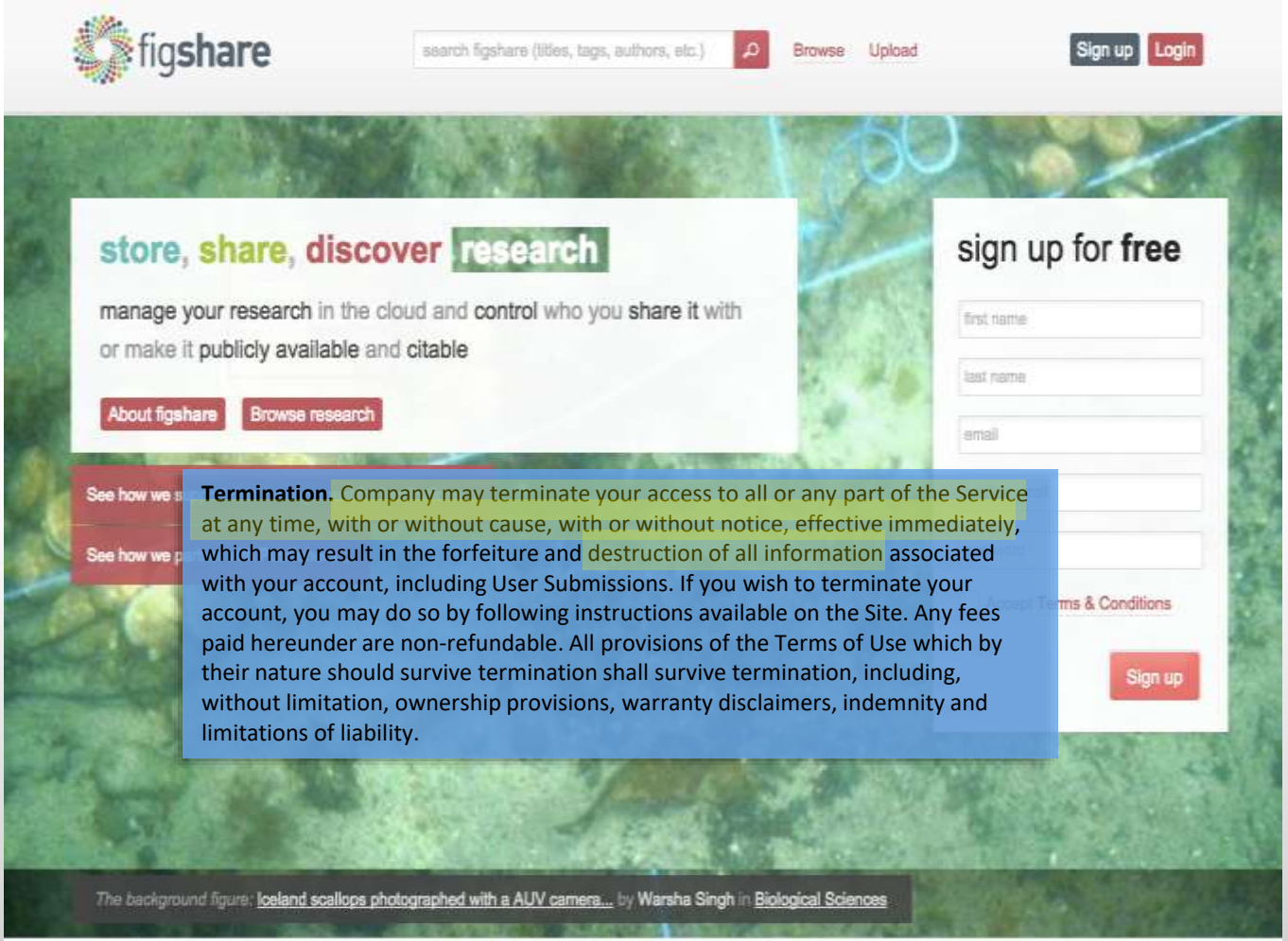

EUDAT
B2HANDLE
B2STAGE
B2SHARE
B2FIND
B2ACCESS
B2DROP
B2SHARE
B2DROP
B2STAGE
B2SAFE


Plan
Collect
Assure
Describe
Preserve
Discover
Integrate
Analyze
Data Life Cycle


Support in All Phases of Research Process
Produce & Collect
Store
Plan
Analyse
Share & Publish

Support in All Phases of Research Process




figshare
search figshare (titles, tags, authors, etc.)
Browse
Upload
Sign up
Login
store, share, discover research
manage your research in the cloud and control who you share it with or make it publicly available and citable
About figshare
Browse research
sign up for free
first name
last name
email
Sign up
Termination. Company may terminate your access to all or any part of the Service at any time, with or without cause, with or without notice, effective immediately, which may result in the forfeiture and destruction of all information associated with your account, including User Submissions. If you wish to terminate your account, you may do so by following instructions available on the Site. Any fees paid hereunder are non-refundable. All provisions of the Terms of Use which by their nature should survive termination shall survive termination, including, without limitation, ownership provisions, warranty disclaimers, indemnity and limitations of liability.
The background figure: Iceland scallops photographed with a AUV camera... by Warsha Singh in Biological Sciences

ResearchGate

Recruit researchers | Join for free | Log in

Advance your research

Discover scientific knowledge, and make your research visible.

ARTICLE 2: DISCLAIMER

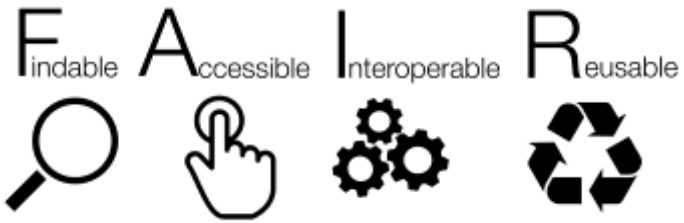
1. The Service is provided "as is" and the Provider disclaims any and all representations and warranties, whether express or implied, including;- but not limited to;- implied warranties of title, merchantability, fitness for any particular purpose or non-infringement. The Provider does not promise any specific results, effects or outcome from the use of the Service.
2. …
3. The Provider reserves the right to change, reduce, interrupt or discontinue the Service or parts of it at any time.
4. No one has a right to use the Service; the Provider reserves the right to exclude certain Users.

"Revolutionizing how research is conducted and disseminated in the digital age."

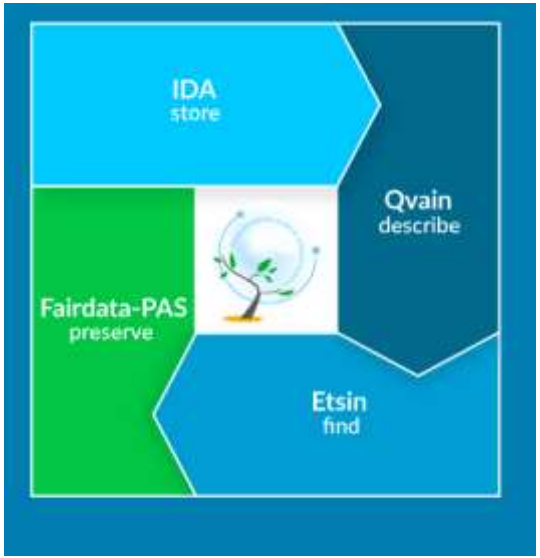
Los Angeles Times

Are the commercial services sufficient?

- Nice complement but can not serve as the fundamental infrastructure for research data of national and international interest
- Need for publicly funded and operated infrastructure


Findable
Accessible
Interoperable
Reusable


FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE
CITABLE
HOW TO GET THERE?
Roadmap
Stakeholders


IDA
store
Qvain
describe
Fairdata-PAS
preserve
Etsin
find

Fairdata.fi

THE FAIR DATA PRINCIPLES

- **FINDABLE:**
  - Data are assigned a globally unique and eternally persistent identifier.
  - Data are described with rich metadata.
  - (Meta)data are registered or indexed in a searchable resource.
  - metadata specify the data identifier.
- **ACCESSIBLE:**
  - (Meta)data are retrievable by their identifier using a standardized communications protocol.
  - The protocol is open, free, and universally implementable.
  - The protocol allows for an authentication and authorization procedure, where necessary.
  - Metadata are accessible, even when the data are no longer available.
- **INTEROPERABLE**
  - (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - (Meta)data use vocabularies that follow FAIR principles.
  - (Meta)data include qualified references to other (meta)data.
- **RE-USABLE:**
  - Meta(data) have a plurality of accurate and relevant attributes.
  - Meta(data) are released with a clear and accessible data usage license.
  - Meta(data) are associated with their provenance.
  - Meta(data) meet domain-relevant community standards.

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

Fairdata.fi

RESEARCH DATA

F FINDABLE
- Described in relevant catalog with enough detail
- Landing page with globally unique identifier

A ACCESSIBLE
- Can be retrieved over the internet
- Versioning and lifecycle documented
- Tombstone page if data is deleted

I INTEROPERABLE
- Common, documented, and open formats

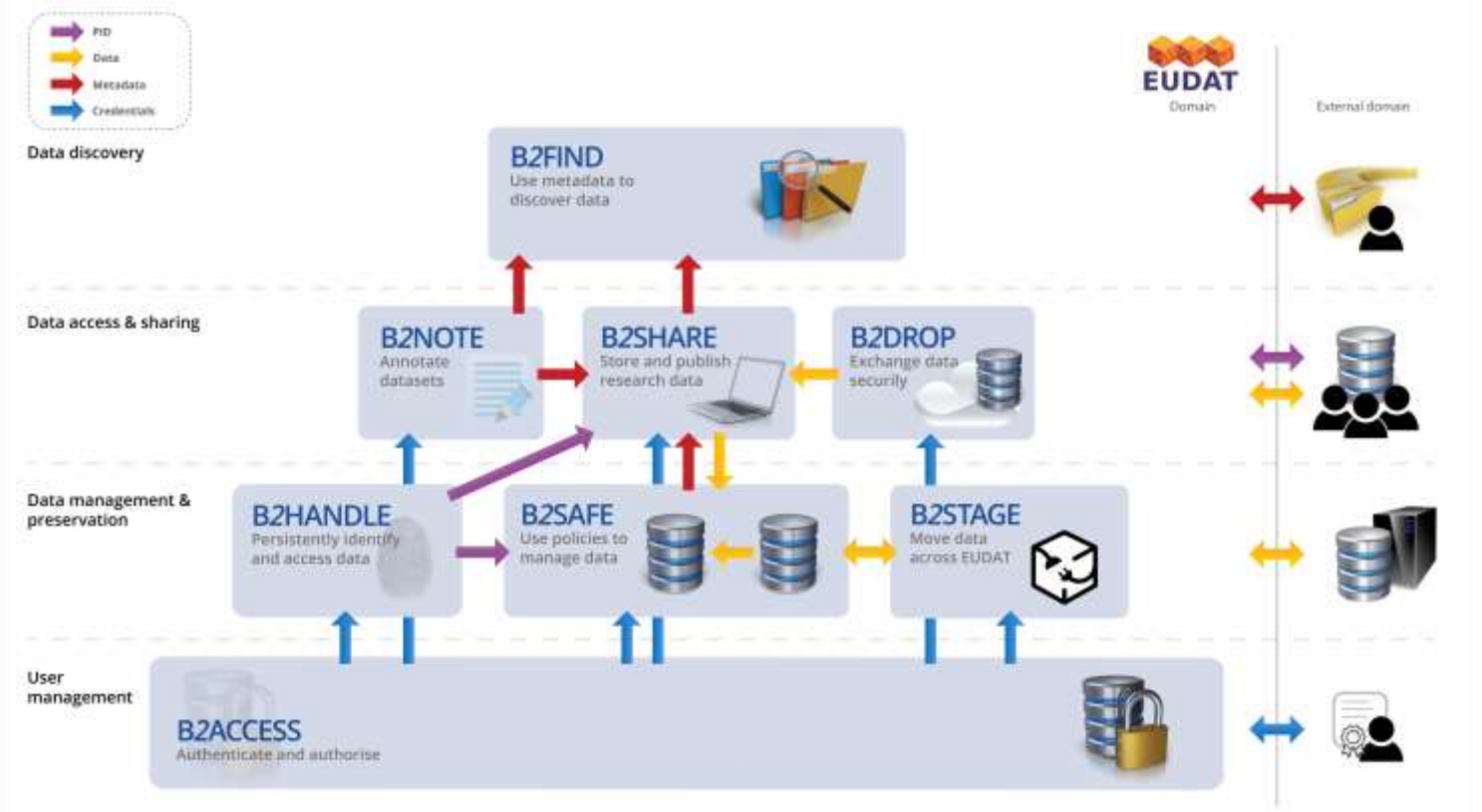
R RE-USABLE
- Well documented and intelligible
- Rights clearly stated

Services to support the data lifecycle

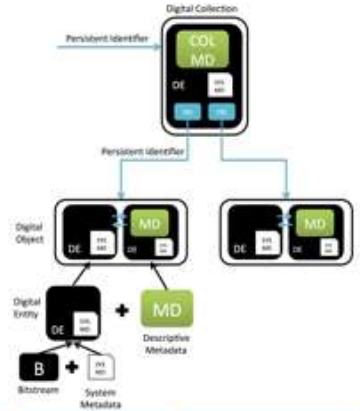
- Plan data management with **DMPTuuli**
- Discover research datasets via **Etsin** and **B2Find** and make sure your data is discoverable, too
- Store data needed in analysis in CSC's **user directories** or within your cloud environment
- Store stable data in **IDA, B2SHARE, HPC Archive** or in dedicated **databases** (Kaivos etc.)
- Share stored data via **B2SHARE** or **IDA**, send large files with **FUNET FileSender**
- Reuse open research datasets: geo-informatics data (**PaiTuli**), speech and text corpora (**Language Bank**) and from various fields of science (**AVAA**)

Fairdata.fi

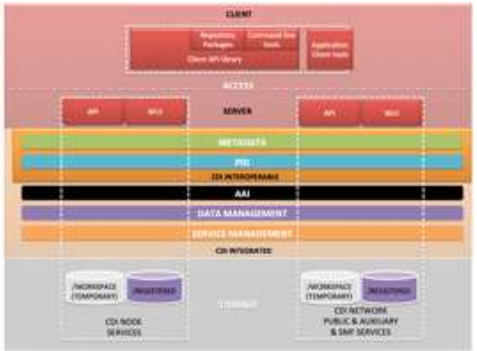
The EUDAT Service Suite



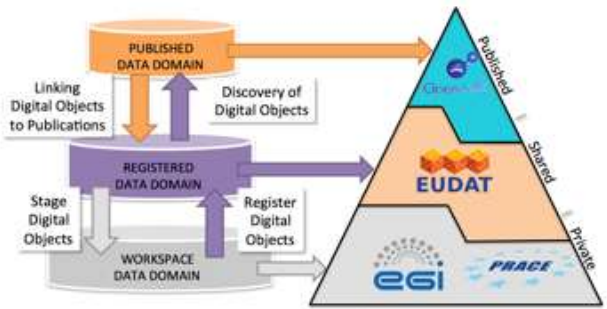
CDI Data Model



CDI Architecture



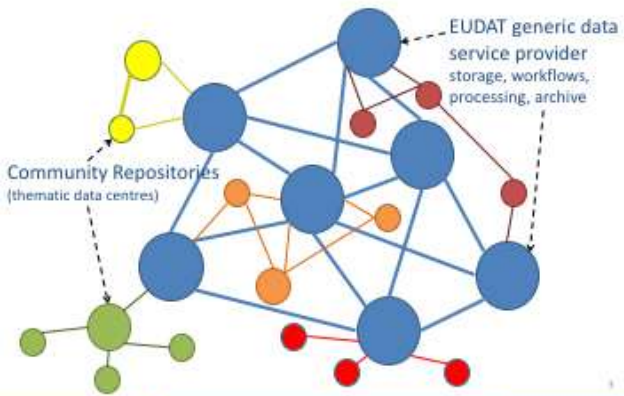
CDI Data Domain



EUDAT Data Domain modeled on the ANDS[1] Data Curation Continiuum

1. Australian National Data Service organization – www.ands.org.au

Collaborative Data Infrastructure (CDI)



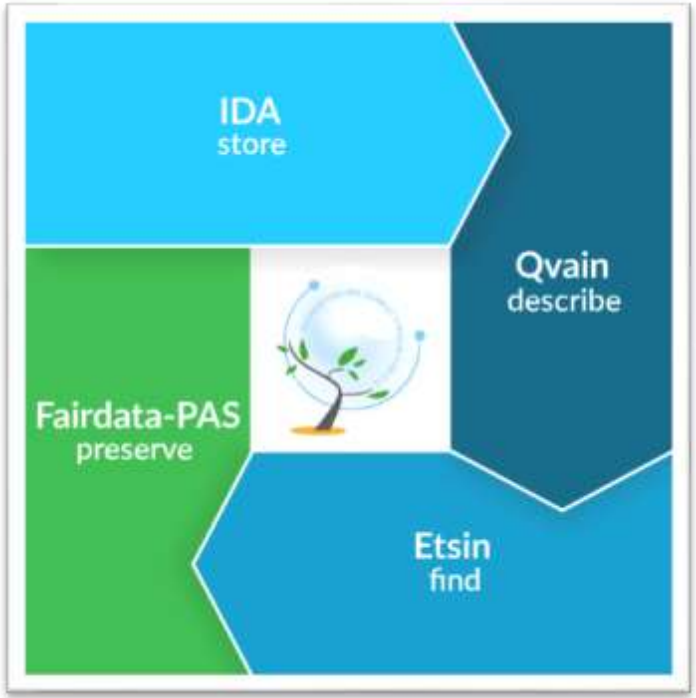
Services to support the data lifecycle

https://research.csc.fi/data-management-and-analytics

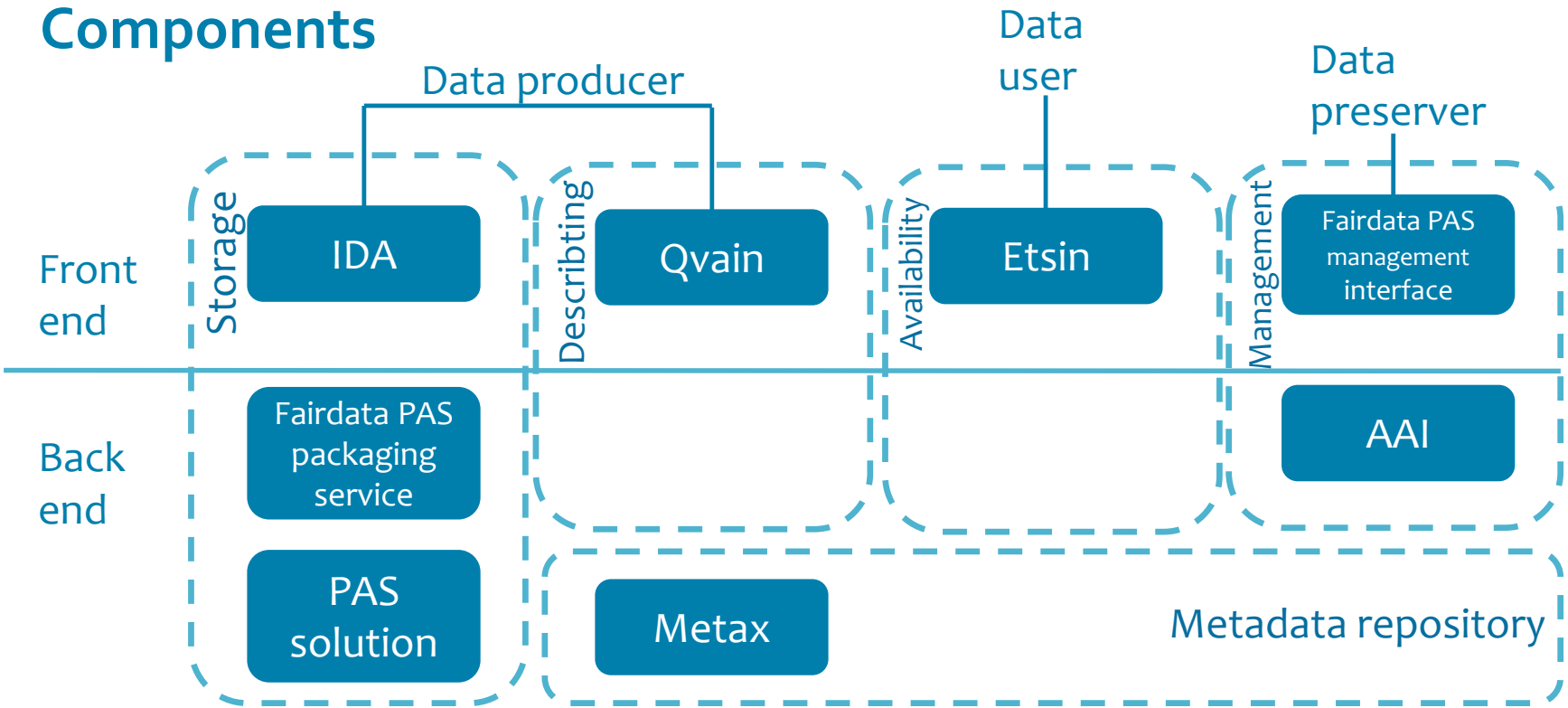
Fairdata.fi

Fairdata services

- **IDA** – store research data
- **ETSIN** – find research data
- **QVAIN** – describe metadata
- **FAIRDATA-PAS** – Preserve research outputs
- In addition (1) metadata repository and (2) authentication solution
- Authentication, ontology, and other concurrent services

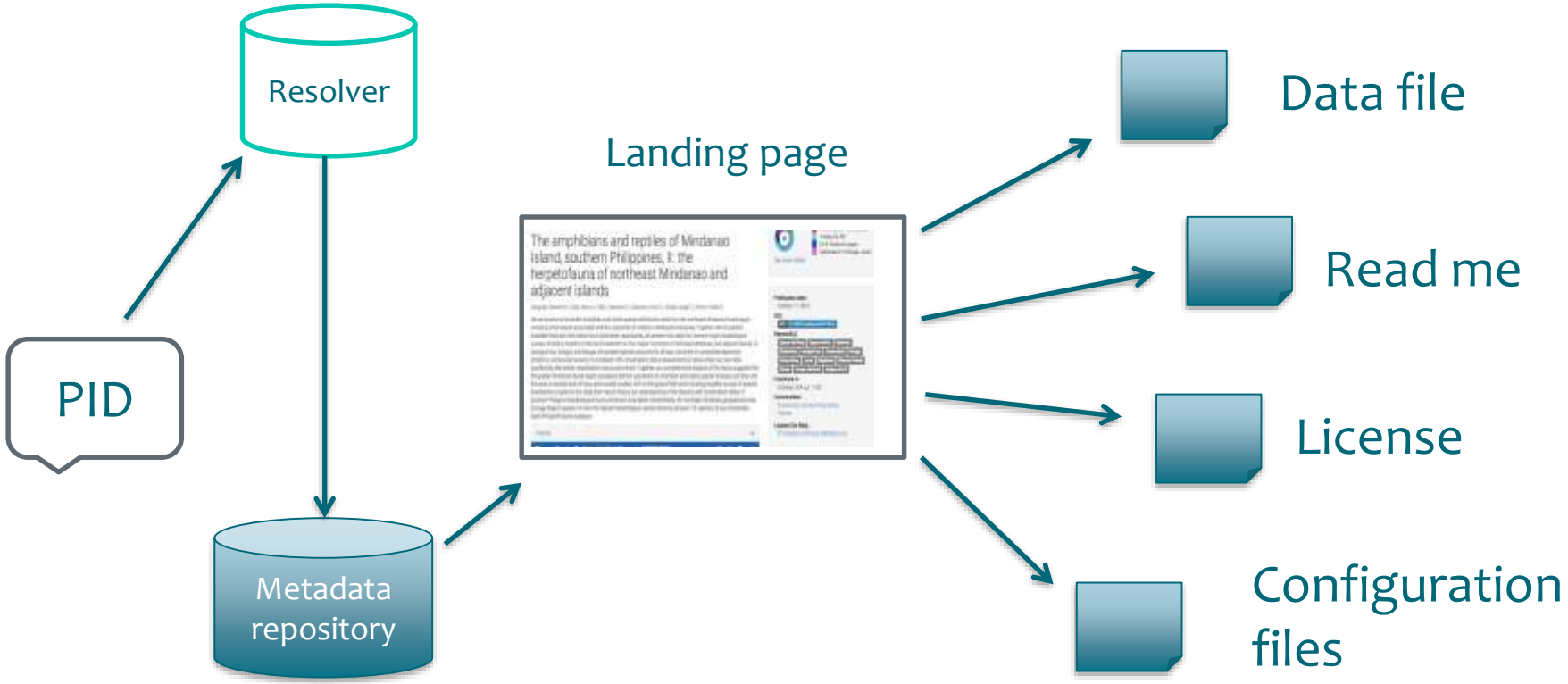


Components



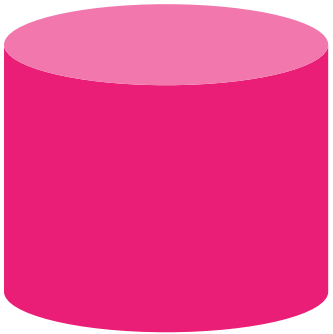
Fairdata.fi

Persistent Identifiers



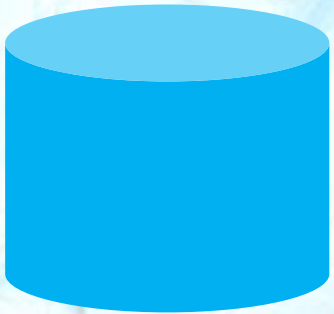
Operational data

- Active data
- Not citable
- Dynamic, volatile



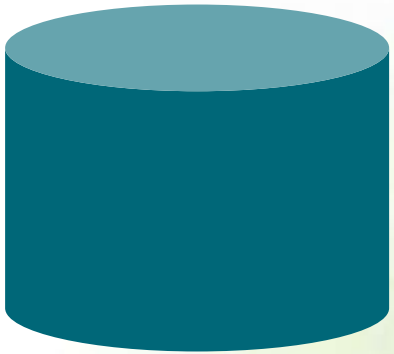
Published data

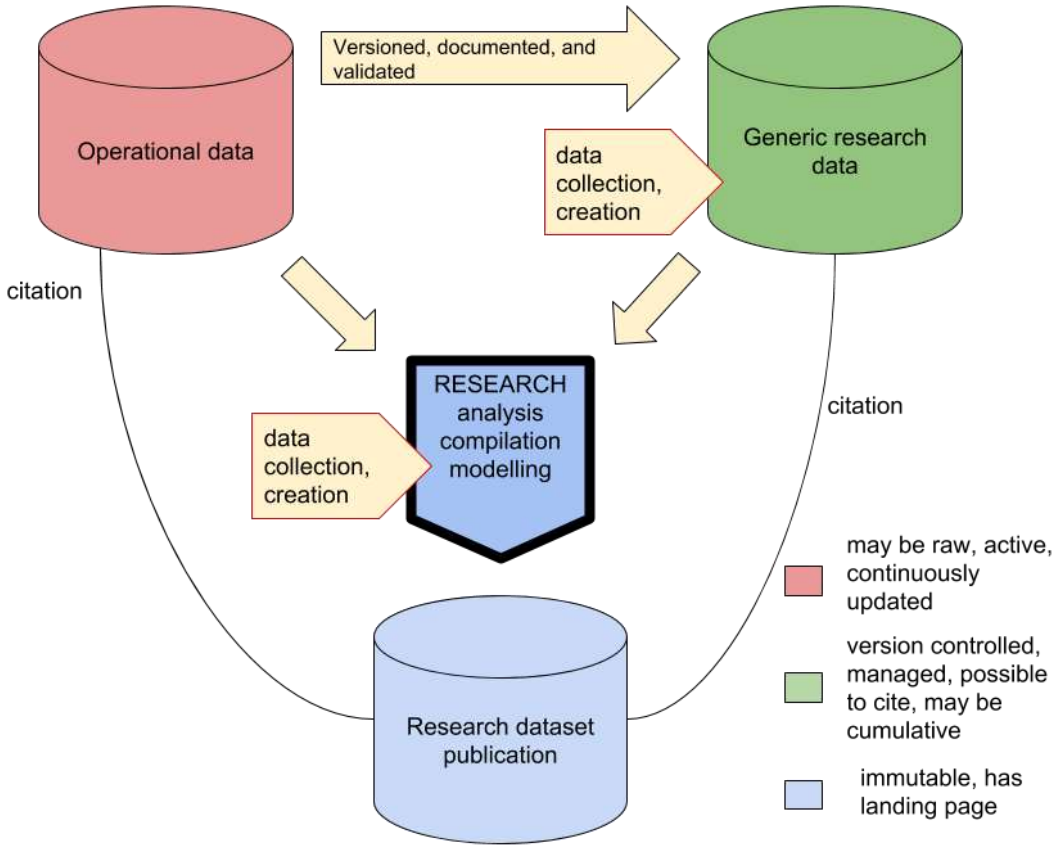
- Static and persistent
- Assigned a PID
- Product of specific research
- Possibly low reusability
- Needed for review and transparence



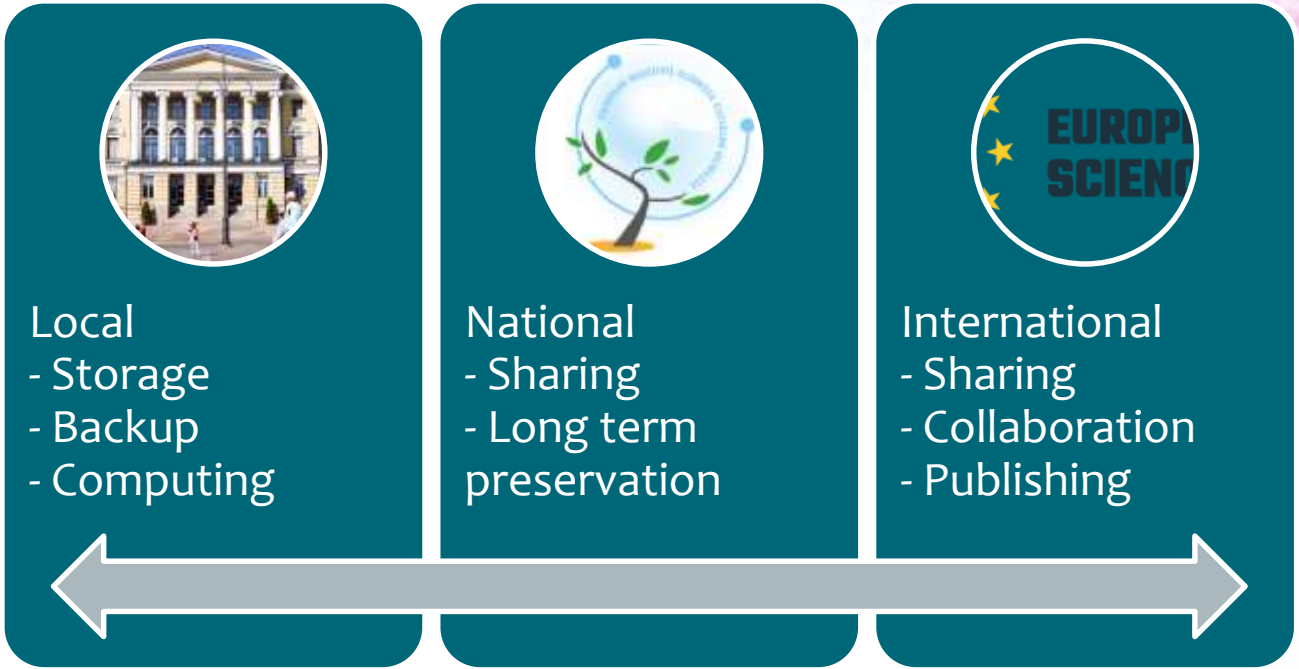
Generic Research data

- Validated by and for researchers
- Possible to cite
- Documented (metadata) and versioned
- Can be dynamic or cumulative




Versioned, documented, and validated
Operational data
Generic research data
data collection, creation
citation
RESEARCH analysis compilation modelling
data collection, creation
citation
Research dataset publication
may be raw, active, continuously updated
version controlled, managed, possible to cite, may be cumulative
immutable, has landing page

The Target


Local
- Storage
- Backup
- Computing
National
- Sharing
- Long term preservation
International
- Sharing
- Collaboration
- Publishing

Acknowledgment

- Timo Vesala, INAR RI, Helsinki University
- Ville Tenhunen, Helsinki Univeristy
- Jessica Parland-von Essen, CSC and Fairdata.fi
- Tommi Nyrönen, ELIXIR-Finland Head of Node, CSC
- Mikael Linden, CSC
- Damien Lecarpentier, EUDAT, CSC

CSC – IT Center for Science Ltd

Per Öster
Director, Research Infrastructures & Policy
per.oster@csc.fi

facebook.com/CSCfi

twitter.com/CSCfi

youtube.com/CSCfi

linkedin.com/company/csc---it-center-for-science

github.com/CSCfi