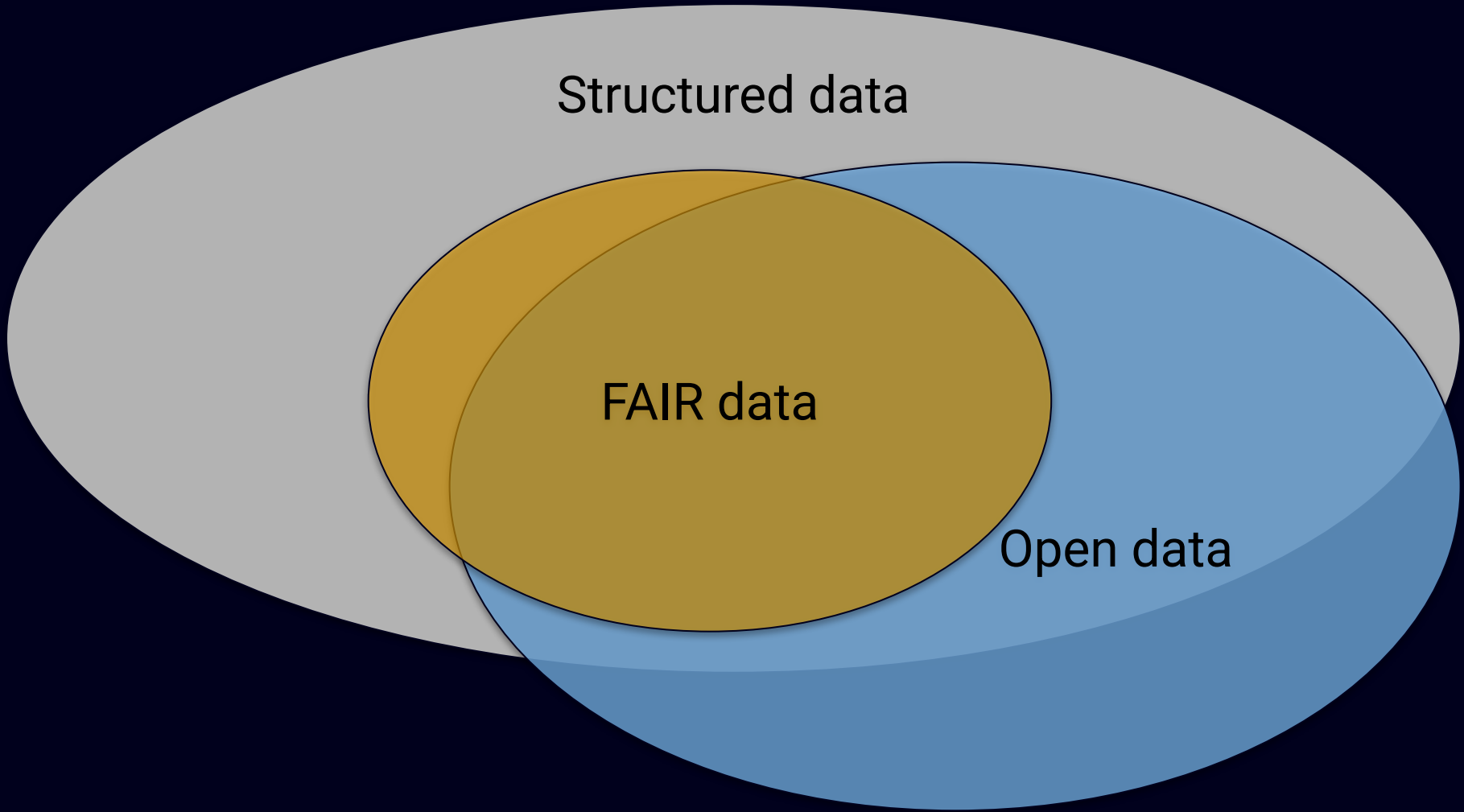
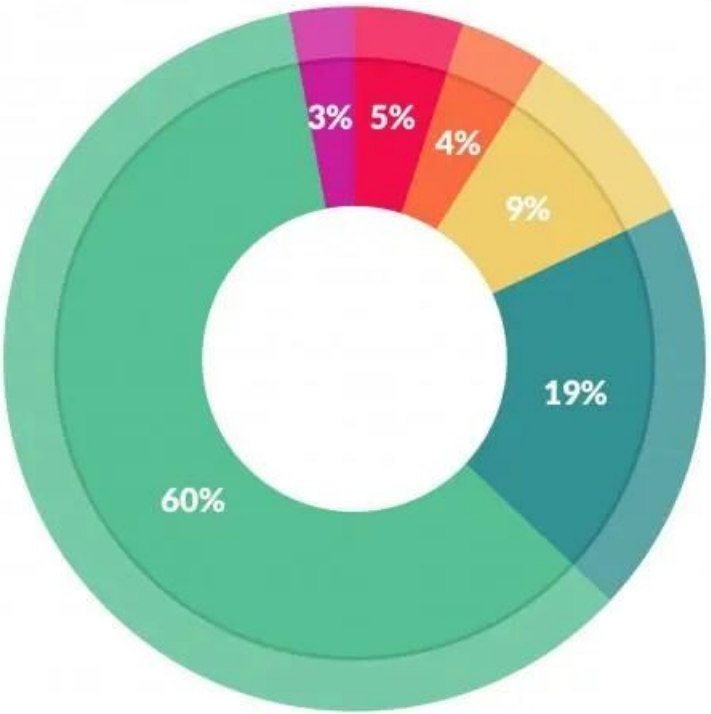
The often understated importance of FAIR data

Luke Marsden
lukem@unis.no

UNIS
The University Centre in Svalbard


Structured data
FAIR data
Open data

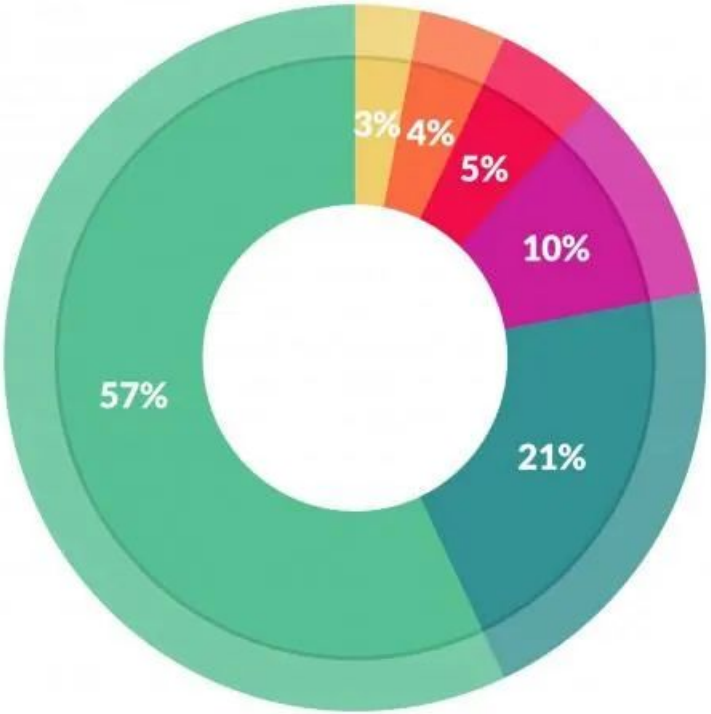
Scientific data should be available to anyone who is curious



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*

G. Press, Cleaning Big Data: Most time-consuming, least enjoyable data science task, survey says, Forbes [Internet], 2016 Mar 23. Available at http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#3643728a7f75.



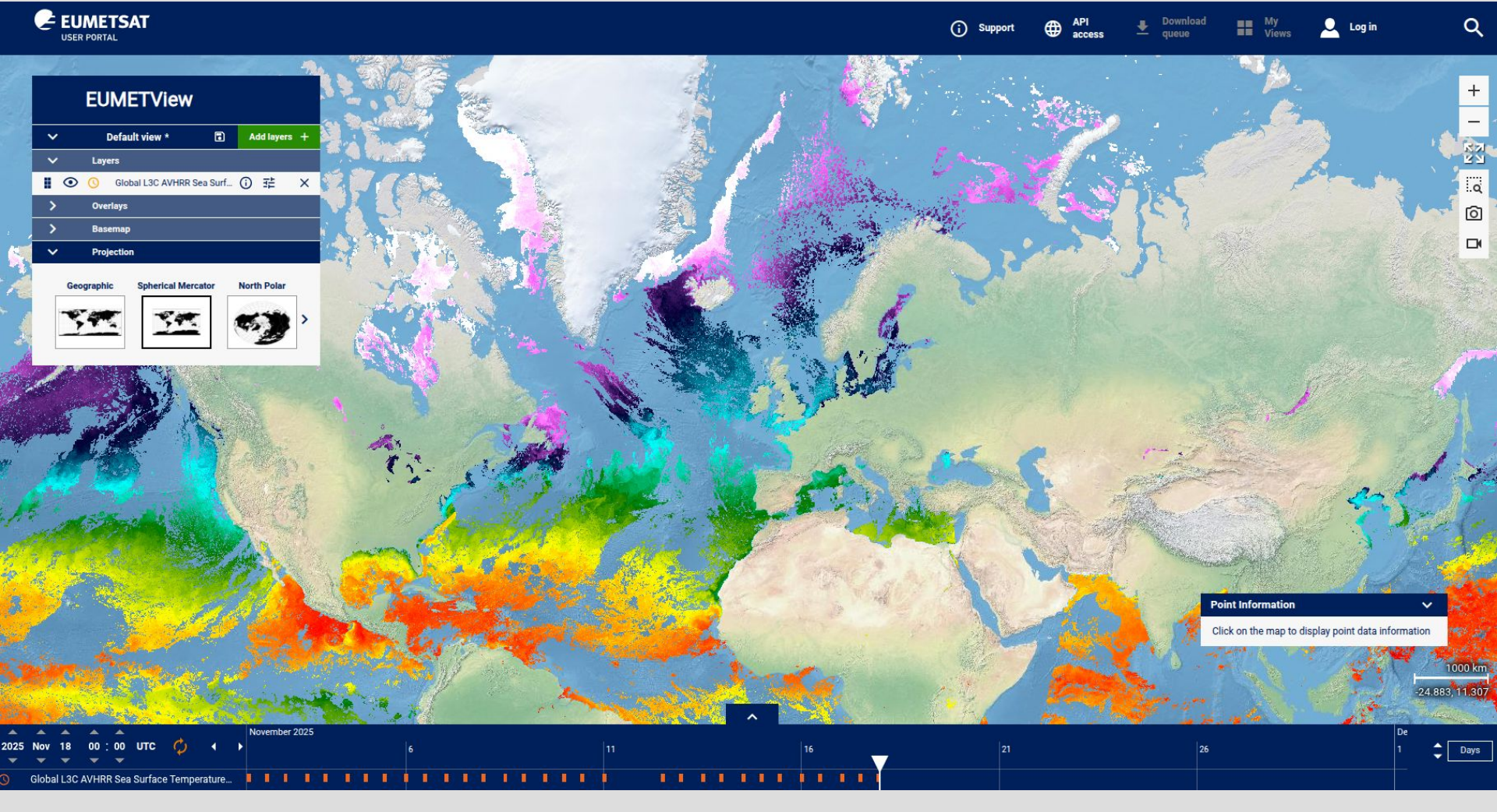
What's the least enjoyable part of data science?

- *Building training sets: 10%*
- *Cleaning and organizing data: 57%*
- *Collecting data sets: 21%*
- *Mining data for patterns: 3%*
- *Refining algorithms: 4%*
- *Other: 5%*

G. Press, Cleaning Big Data: Most time-consuming, least enjoyable data science task, survey says, Forbes [Internet], 2016 Mar 23. Available at http://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#3643728a7f75.

To make data available to anyone who is curious we need….

- Aggregation done for you


EUMETSAT
USER PORTAL
Support
API access
Download queue
My Views
Log in
EUMETView
Default view *
Add layers
Layers
Global L3C AVHRR Sea Surf...
Overlays
Basemap
Projection
Geographic
Spherical Mercator
North Polar
Point Information
Click on the map to display point data information
1000 km
-24.883, 11.307
2025 Nov 18 00 : 00 UTC
November 2025
6
11
16
21
26
1
Days
Global L3C AVHRR Sea Surface Temperature...

Occurrence search

gbif.org/occurrence/map?has_coordinate=true&has_geospatial_issue=false&taxon_key=4261391

Get data | How-to | Tools | Community | About | Login

Occurrences

Licence

Scientific name

- [x] Grewia occidentalis L.

Basis of record

Year

Month

Location

- [x] Including coordinates

Administrative areas (gadm.org)

Country or area

Continent

Dataset

Search

- [ ] iNaturalist Research-grade Observations 1,985
- [ ] Botanical Database of Southern Africa (BODA... 766
- [ ] Tree planting occurrences 71
- [ ] occurence_data_for_wild_food_species_in_zi... 31
- [ ] O armário: Fruiting phenology data for 4,462 pl... 21
- [ ] Pl@ntNet automatically identified occurrences 18
- [ ] National Herbarium of Victoria (MEL) AVH data 17
- [ ] IICT Herbário LISC 14
- [ ] African Plants - a photo guide 11
- [ ] Pl@ntNet observations 8

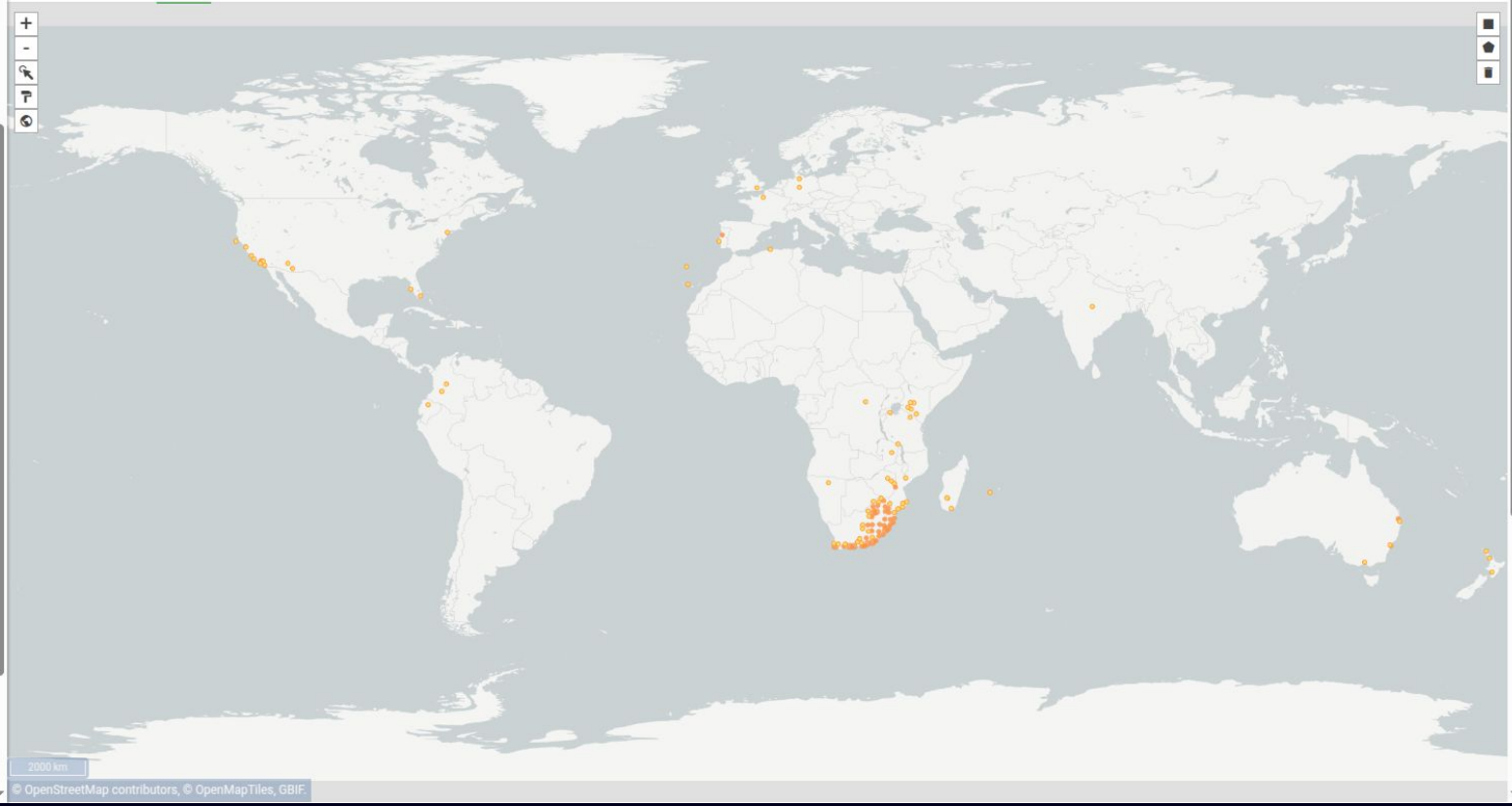
Publisher

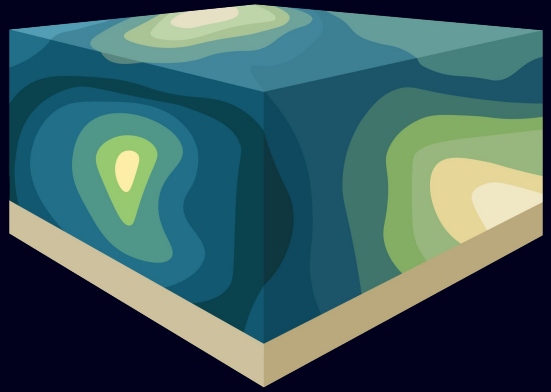
IUCN Global Red List Category

Issues and flags

SEARCH OCCURRENCES | 3,038 WITH COORDINATES

TABLE | GALLERY | MAP | TAXONOMY | METRICS | DOWNLOAD



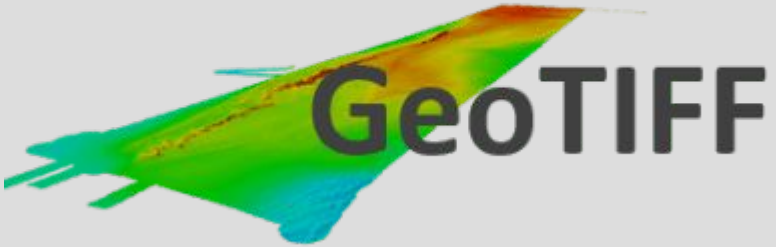


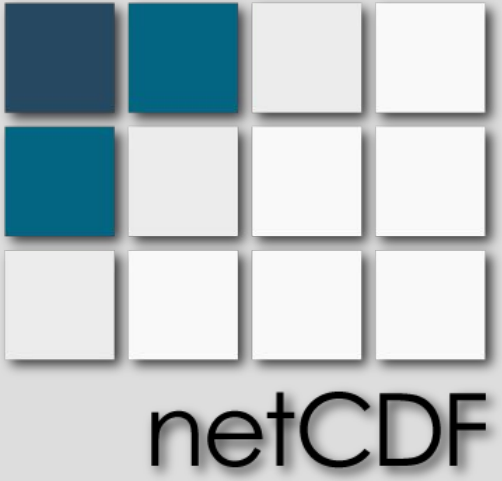
3D visualisation of the data

Collection of similar datasets

Illustrations by Frida Cnossen, Nansen Legacy


GeoJSON


GeoTIFF


netCDF


HDF


CSV

To make data available to anyone who is curious we need….

- Aggregation done for you
- Download in choice of format

To make data available to anyone who is curious we need….

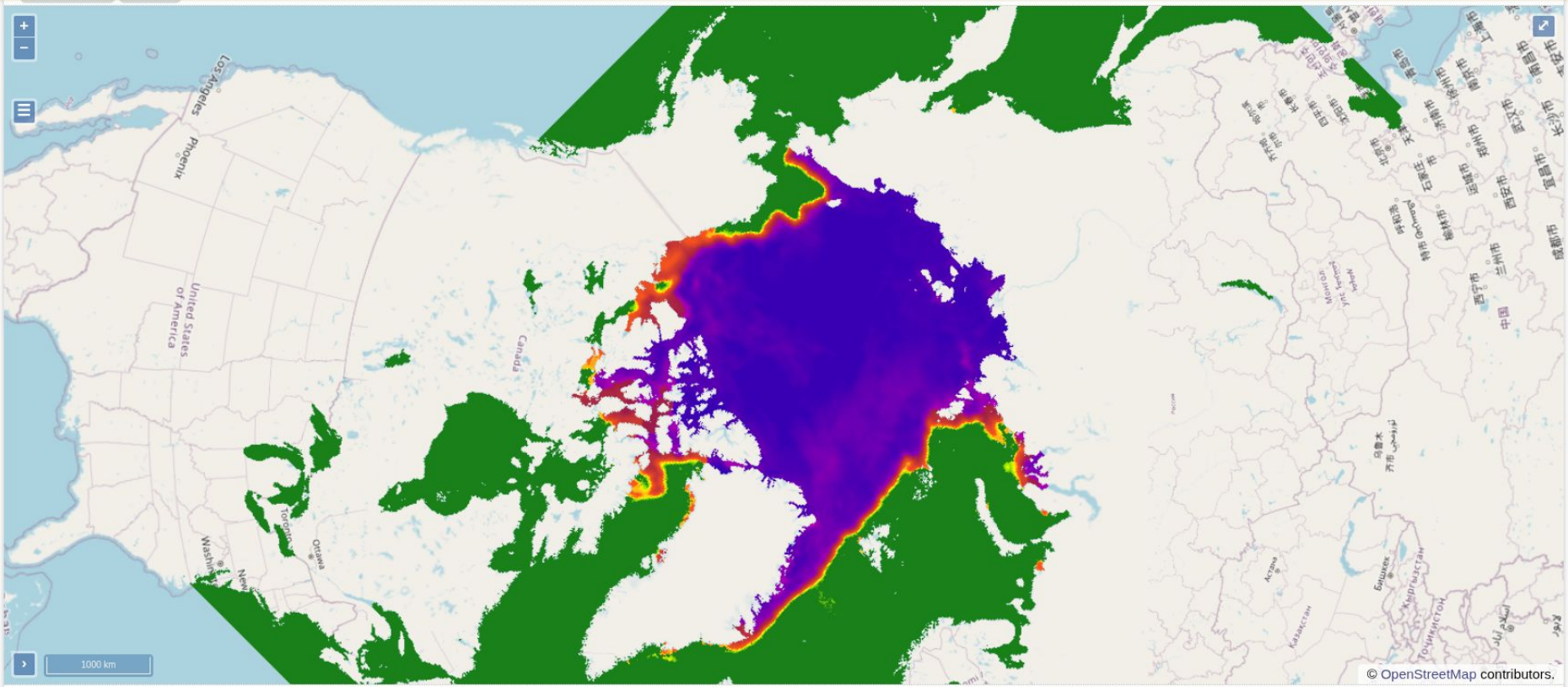
- Aggregation done for you
- Download in choice of format
- Visualisation made easy

sea ice

Visualisation on SIOS

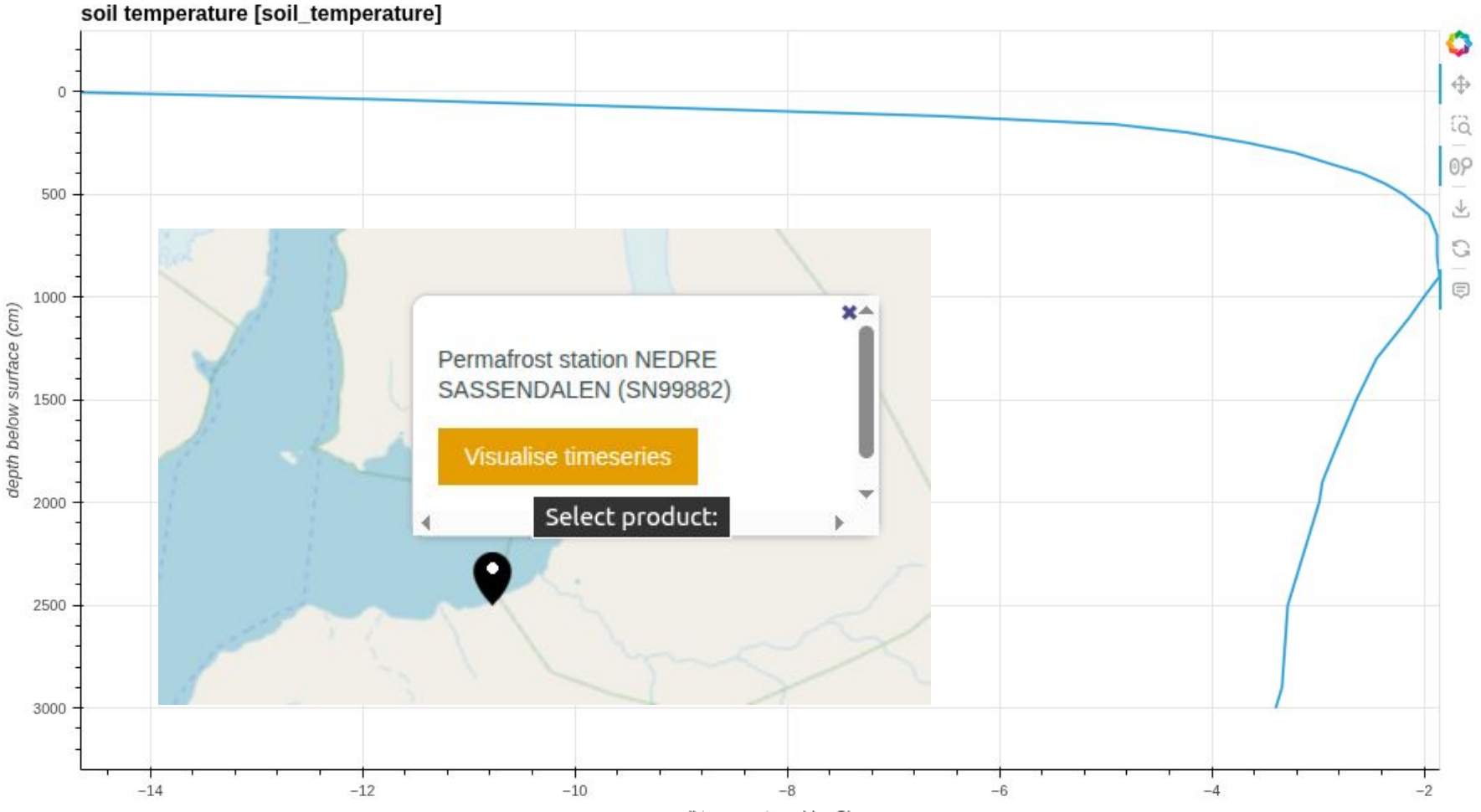
sea ice

Home | About SIOS | Services | Data | Research Infrastructure (RI) | Intranet

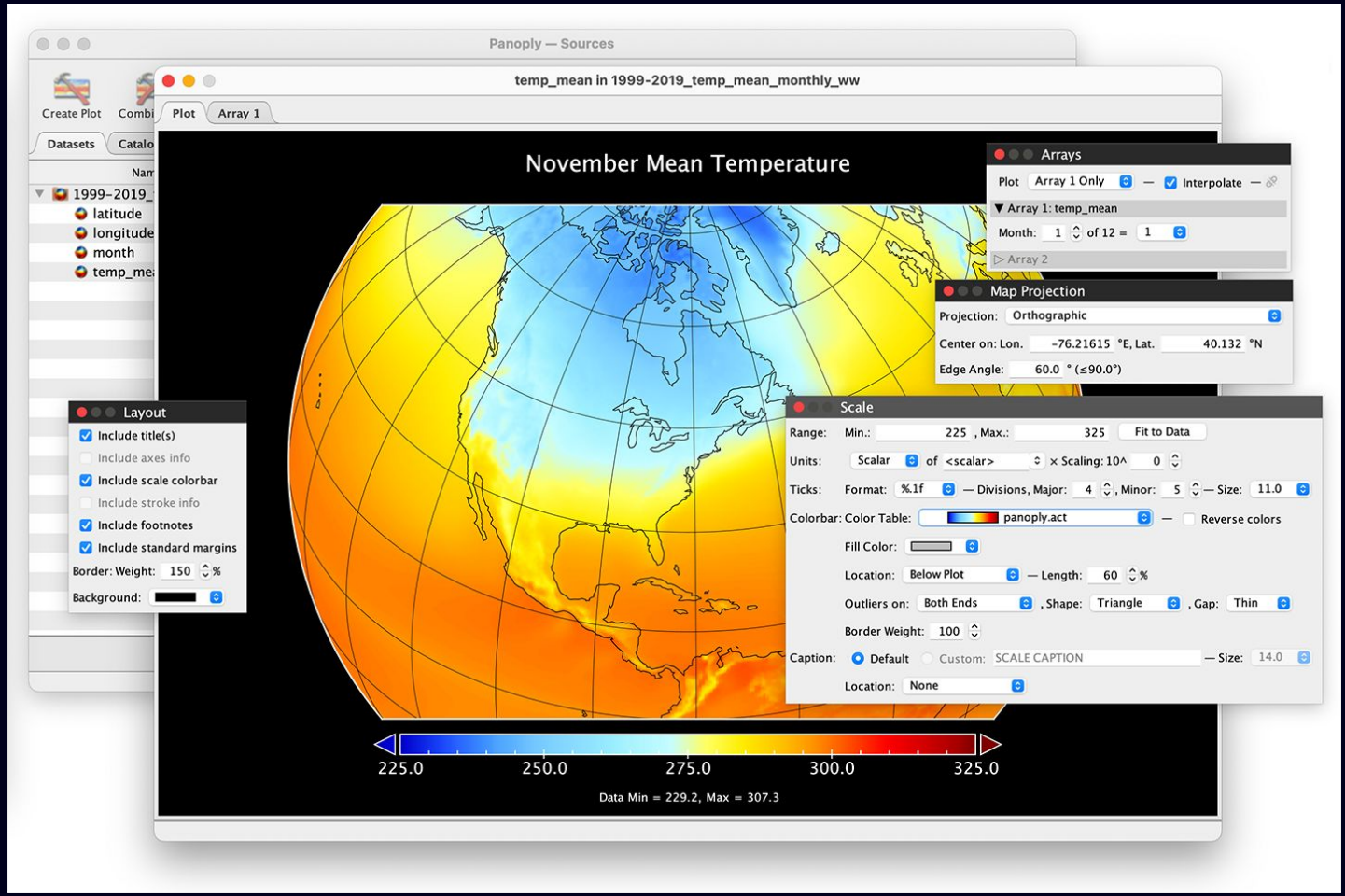


Time dimensions: 2025-11-08T12:00:00.000Z

time with frequency of 1 hour: **2024-03-25 18:00:00**



Data Visualisation In Panoply



To make data available to anyone who is curious we need….

- Aggregation done for you
- Download in choice of format
- Visualisation made easy

Machine Actionable Data

FAIR is a set of guiding principles

How to make data machine actionable

So tedious and boring tasks can be automated.

FAIR data

Findability

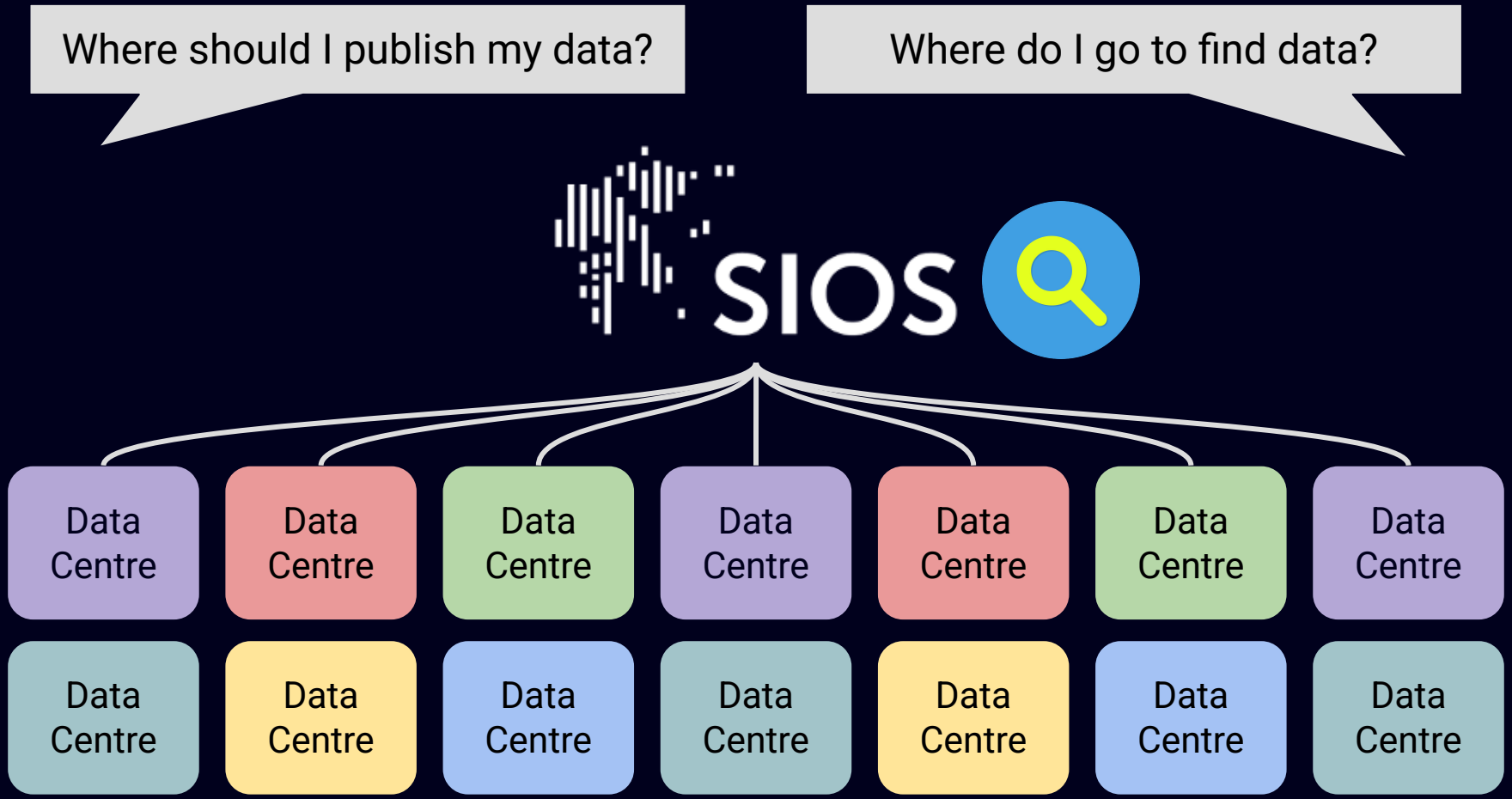
Accessibility

Interoperability

Reusability

Data formats

Data centres


Where should I publish my data?
Where do I go to find data?
SIOS
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre
Data Centre

Search

Contains all of these words

Enter your search here

Start Date mm/dd/yyyy End Date mm/dd/yyyy

Has children
Select whether datasets are parents with children (i.e. records of the same type)

▶ Advanced options

Search Reset

Dataset Level

Parent (505945)

Iso Topic Category

- climatologyMeteorologyAtmosphere (503992)
- oceans (503304)
- imageryBaseMapsEarthCover (502573)
- biota (155)
- geoscientificInformation (145)
- environment (131)
- elevation (25)
- society (16)
- boundaries (10)
- inlandWaters (9)

Show more

Keywords

- Land cover (502513)
- Orthoimagery (502509)
- Earth Science > Atmosphere > Atmospheric

- About the Data Access Portal
- Citation of data and service
- SIOS Core Data
- Nansen Legacy Data
- Ny-Ålesund research data
- Brief user guide

https://sios-svalbard.org/metsis/search

Select projection: EPSG:4326 UPS North UPS South **Select spatial filter:** Within Intersects

Create bounding box filter | Reset search | Reset map

505945 datasets found. Showing datasets 1 - 15 on page 1 of 33730 pages.

Project

- NBS (502509)
- Nansen Legacy (835)
- APPLICATE (450)
- SIOS (450)
- YOPP (450)

Show more

Collection

- SIOS (505909)
- NBS (502509)
- ADC (3415)
- NSDN (2495)
- GCW (852)

Show more

Personnel

- NBS Helpdesk (502509)
- NBS team (462350)
- Arnfinn Morvik (1047)
- Øystein Godøy (475)
- Norwegian Polar Institute (329)

Show more

Organisation

- Norwegian Meteorological Institute (503007)
- NO01L, Norwegian Institute for Air Research, NILU, Instituttveien 18, 2007, Kjeller, Norway (432)
- NO01L, Norwegian Institute for Air Research, NILU, Atmosphere and Climate Department, Instituttveien 18, 2007, Kjeller, Norway (89)
- DK01L, National Environmental Research Institute, NERI, P.O.Box 358, Fredriksborgvei 399, DK-4000 Roskilde, Denmark (68)
- Norwegian Meteorological Institute / Arctic Data Centre (62)

Show more

Data Center

- METNO (502509)
- NO/NMDC/IMR (605)
- Norwegian Polar Institute (506)
- WMO (450)

Other data portals

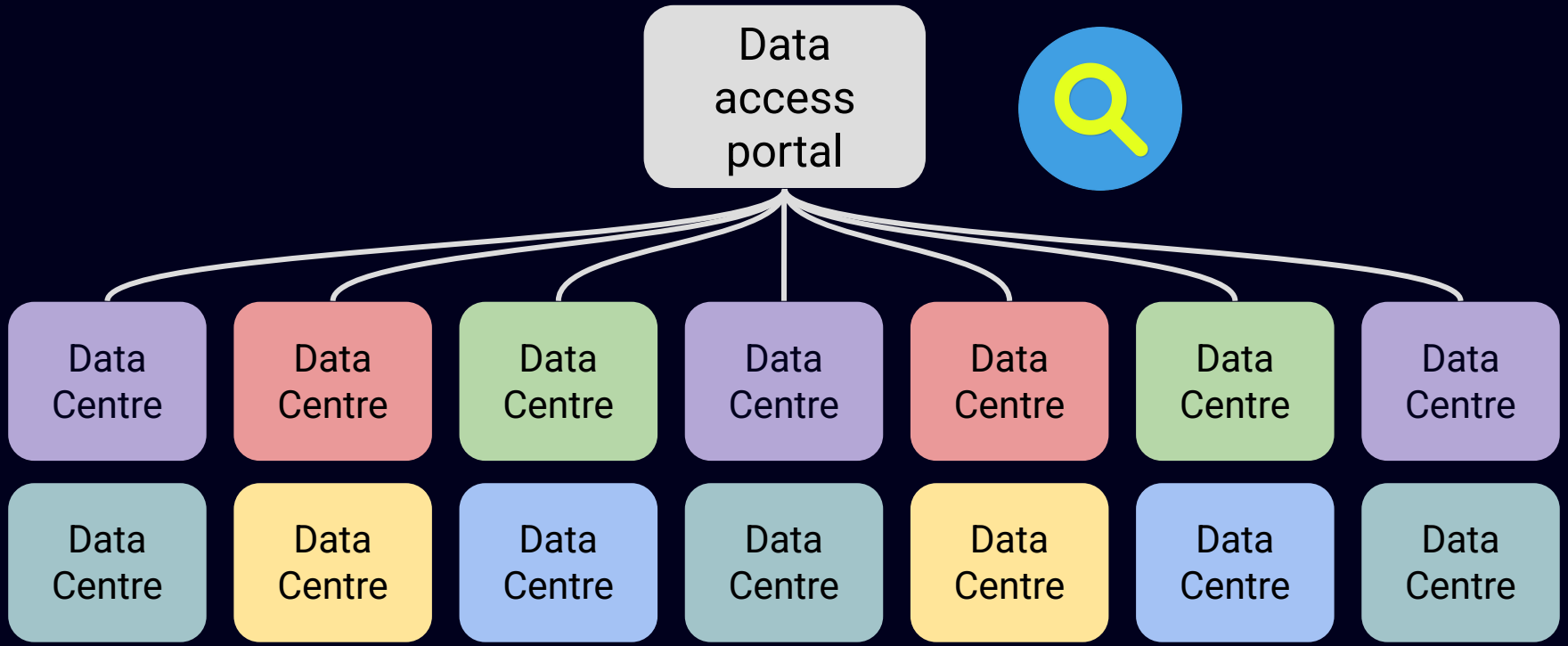
https://sios-svalbard.org/metsis/search

https://search.polder.info/

https://www.geoportal.org/

Other platforms yet to be developed…

Not plausible to connect all data centres to all portals



Discovery metadata conventions

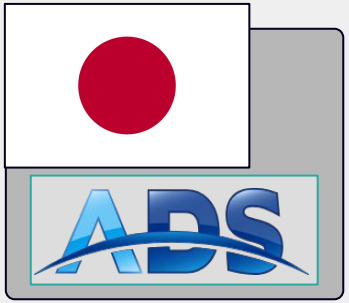
e.g. ISO 19115, GCMD DIF

Metadata cataloguing and harvesting protocols

e.g. OAI-PMH, OGC-CSW

Method for serving the metadata

Data centres that contribute to SIOS



More to come…

Where to publish (most) UNIS data

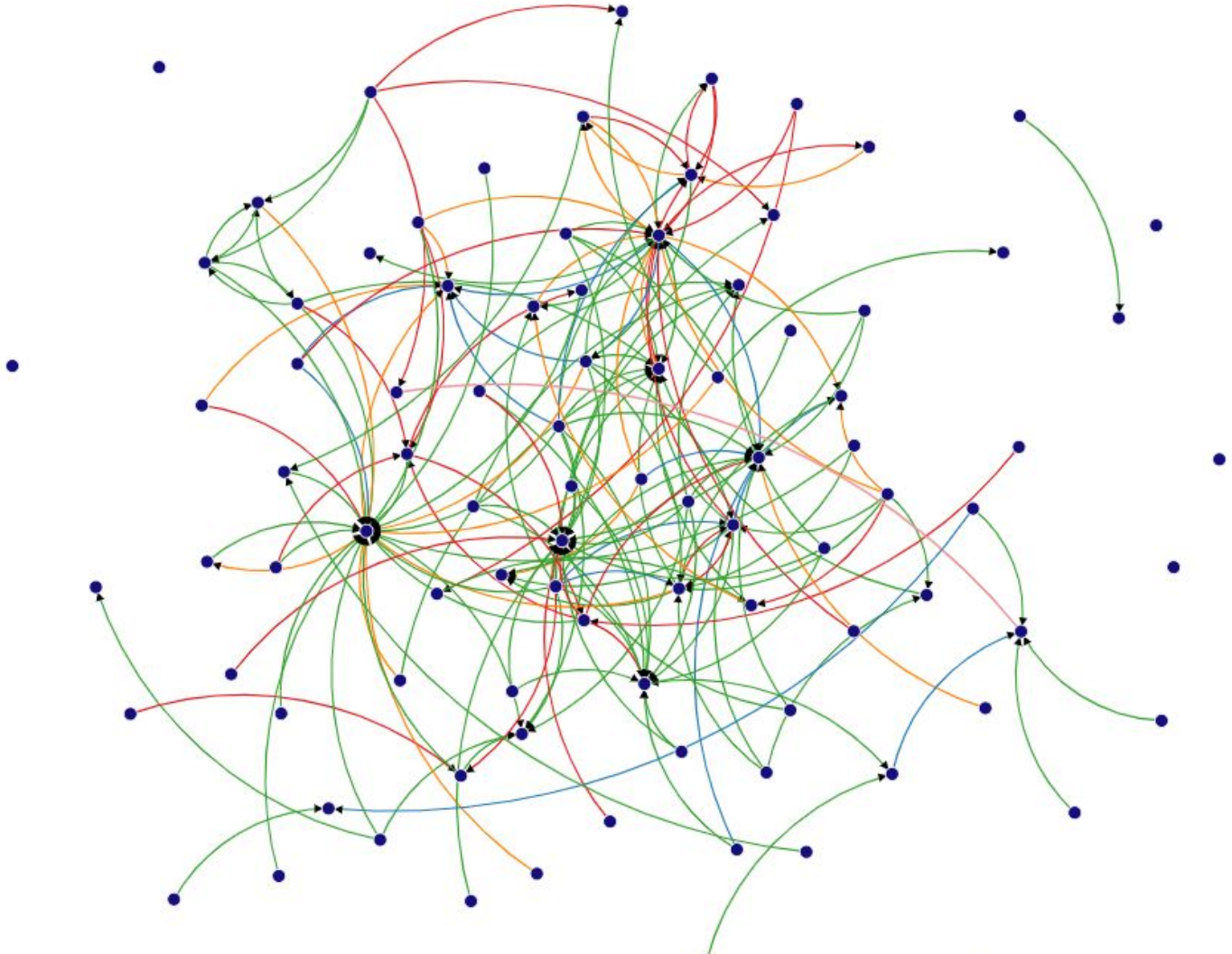
Arctic Data Centre (MET)
https://adc.met.no/

Norwegian Marine Data Centre (IMR)
http://prod1.nmdc.no/UserInterface/

Norwegian Polar Data Centre (NPI)
https://data.npolar.no/

NIRD Research Data Archive
https://archive.sigma2.no/

EBAS (NILU)
https://ebas.nilu.no/



FAIR data

Findability

Accessibility

Interoperability

Reusability

Data formats

Data centres

un-FAIR data

Date	Time	Temperature	Salinity	X

Machine actionable data formats

Standardised structure

Include metadata

Controlled vocabularies

Software independent

Well documented & maintained

Machine-actionable data formats

CF-NetCDF

- Measurements of physical environment
- Time series, vertical profiles, multiple dimensions
- Outputs from models
- Point clouds
- Satellite data
- Flow cytometry
- Sedimentary logs
- …

Darwin Core Archive

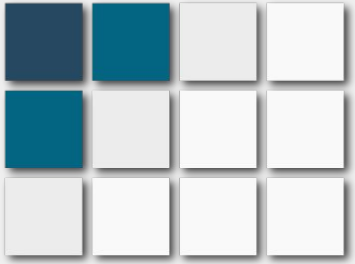
- Biodiversity data
- Associated measurements/facts
- Species lists from DNA data
- Fossil data
- …

Other formats

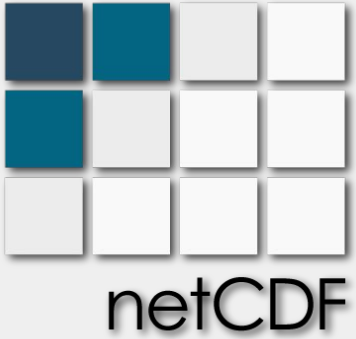
- Open is not machine-actionable. Need standardised structure
  - e.g. JSON, XML
- Consistent format but few standards and insufficient metadata
  - e.g. PLY, LAS, outputs from instruments, netCDF without CF
- Metadata conventions not mature enough to be recommended by SIOS, but international efforts to improve this
  - e.g. geoJSON, geoTIFF, geoZARR, CAAML

Advantages of fewer standards

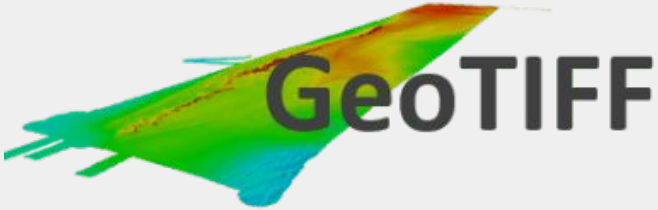
- Fewer standards to learn
- Easier to share data between disciplines
- More efficient to integrate data into services and maintain them



SAFE



Coming soon...



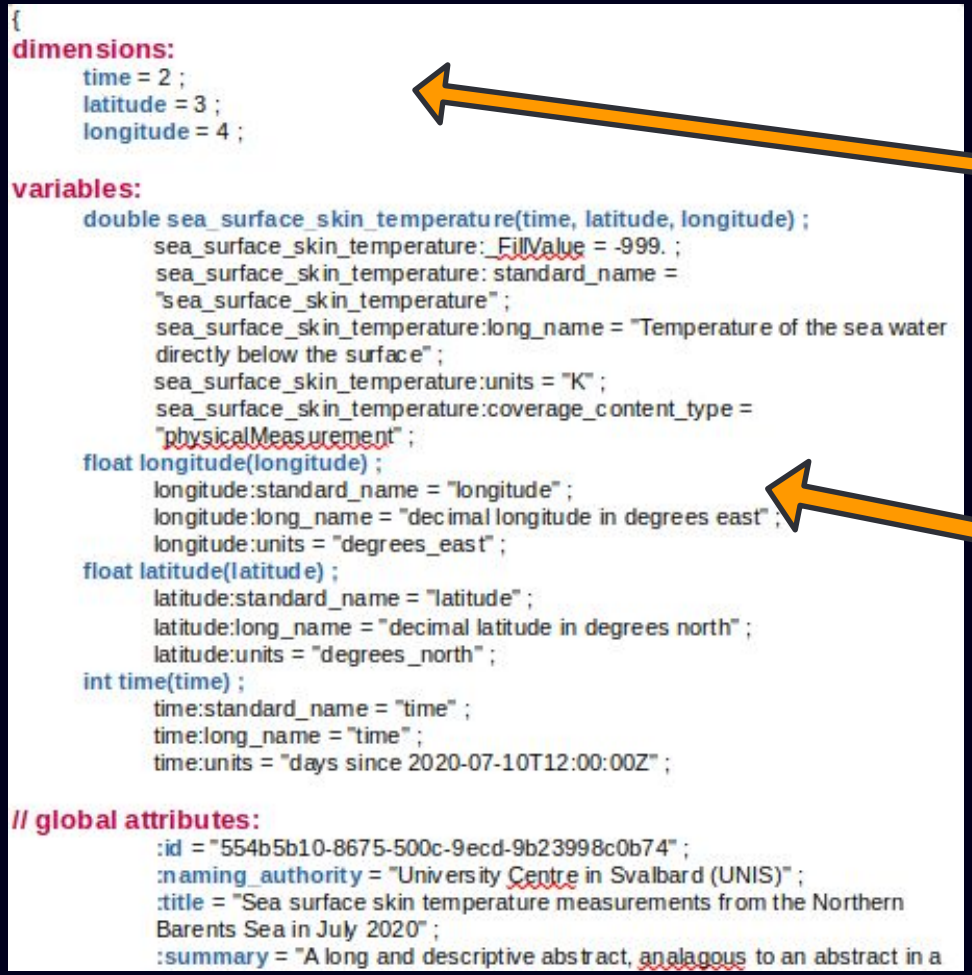
Climate & Forecast (CF) Conventions

Metadata convention

Traditionally only for netCDF (CF-NetCDF)

Investigating how to apply these conventions to other formats

Allow interoperability between formats



Dimensions:

- 3 points of latitude
- 4 points of longitude
- 2 points time

Variables:

- Variable ‘lat’ has dimension lat
- Variable ‘…temperature…’ has 3 dimensions
- Variable attributes

// global attributes:

:id = "554b5b10-8675-500c-9ecd-9b23998c0b74" ;
:naming_authority = "University Centre in Svalbard (UNIS)" ;
:title = "Sea surface skin temperature measurements from the Northern Barents Sea in July 2020" ;
:summary = "A long and descriptive abstract, analagous to an abstract in a journal article, that describes the data" ;
:keywords = "Earth Science > Oceans > Ocean Temperature > Sea Surface Temperature > Sea Surface Skin Temperature" ;
:keywords_vocabulary = "GCMD" ;
:geospatial_lat_min = 76LL ;
:geospatial_lat_max = 78LL ;
:geospatial_lon_min = 28LL ;
:geospatial_lon_max = 31LL ;
:time_coverage_start = "2020-07-10T12:00:00Z" ;
:time_coverage_end = "2020-07-19T12:00:00Z" ;
:Conventions = "ACDD-1.3, , CF-1.8" ;
:history = "File created at 2022-01-10T14:21:58 using the xarray library in Python" ;
:source = "Sea surface temperature measurements recorded by....okay, I made them up!" ;
:processing_level = "raw" ;
:date_created = "2022-01-10T14:21:58Z" ;
:creator_type = "person; person" ;
:creator_institution = "The University Centre in Svalbard, Norway; The University Centre in Svalbard, Norway" ;
:creator_name = "Luke Marsden; John Doe" ;
:creator_email = "lukem@unis.no; jodoe@unis.no" ;
:creator_url = "https://www.unis.no/staff/luke-marsden/; https://www.unis.no/staff/john-doe/" ;
:institution = "The University Centre in Svalbard (UNIS)" ;
:publisher_name = "Norwegian Meteorological Institute - Arctic Data Centre" ;
:publisher_email = "adc-suppoort@met.no" ;
:publisher_url = "https://adc.met.no/" ;
:publisher_type = "institution" ;
:project = "The Nansen Legacy (RCN # 276730)" ;
:license = "https://creativecommons.org/licenses/by/4.0/" ;
:metadata_link = "" ;
:metadata_catalogue = "https://sios-svalbard.org/aen/tools, event ID = a5ec89da-f45e-11eb-b6ae-0f3712703d75" ;
:acknowledgements = "Funded by the Research Council of Norway. John Smith was involved in collecting the data" ;

Conventions used:
CF
ACDD

Dataset validation form

Select the test you want to run *

- [x] CF
- [x] ACDD

Select the CF convesion version

- ( ) CF-1.6
- ( ) CF-1.7
- (•) CF-1.8

Upload Your File *

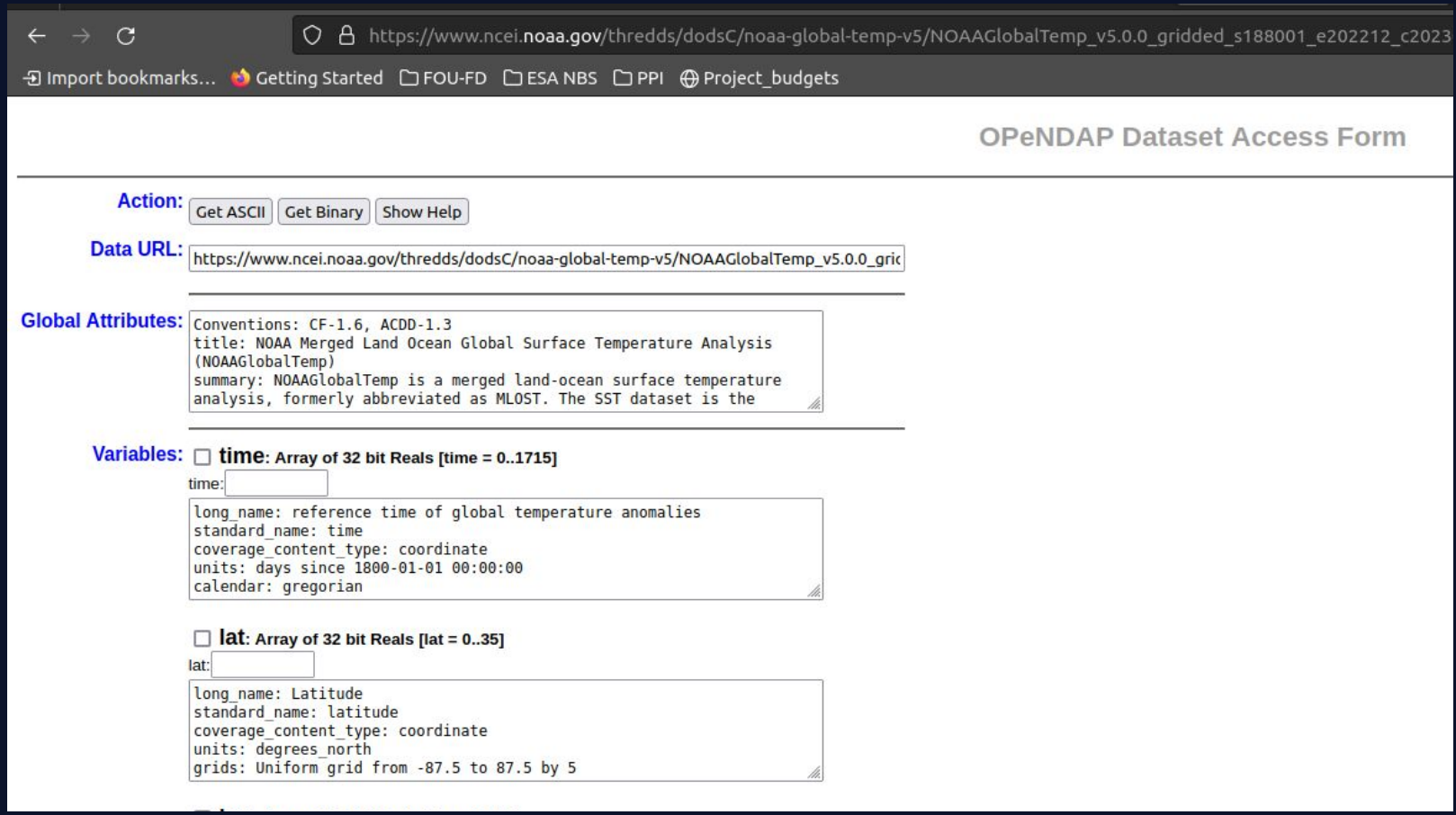
Choose File No file chosen

You can upload a single netCDF (.nc) file, or an archive with multiple netCDF files (zip tar tgz gz bz2) Maximum filesize is 1500M. You need to upload a bigger file, take contact with the website support directly.

Validate

Webform for validation of netCDF files based on the IOOS compliance checker

OPeNDAP


https://www.ncei.noaa.gov/thredds/dodsC/noaa-global-temp-v5/NOAAGlobalTemp_v5.0.0_gridded_s188001_e202212_c2023
Import bookmarks...
Getting Started
FOU-FD
ESA NBS
PPI
Project_budgets
OPeNDAP Dataset Access Form
Action:
Get ASCII
Get Binary
Show Help
Data URL:
https://www.ncei.noaa.gov/thredds/dodsC/noaa-global-temp-v5/NOAAGlobalTemp_v5.0.0_gric
Global Attributes:
Conventions: CF-1.6, ACDD-1.3
title: NOAA Merged Land Ocean Global Surface Temperature Analysis (NOAAGlobalTemp)
summary: NOAAGlobalTemp is a merged land-ocean surface temperature analysis, formerly abbreviated as MLOST. The SST dataset is the
Variables:
time: Array of 32 bit Reals [time = 0..1715]
time:
long_name: reference time of global temperature anomalies
standard_name: time
coverage_content_type: coordinate
units: days since 1800-01-01 00:00:00
calendar: gregorian
lat: Array of 32 bit Reals [lat = 0..35]
lat:
long_name: Latitude
standard_name: latitude
coverage_content_type: coordinate
units: degrees_north
grids: Uniform grid from -87.5 to 87.5 by 5

Are FAIR data available to anyone who is curious today? No

Specialist users 😊

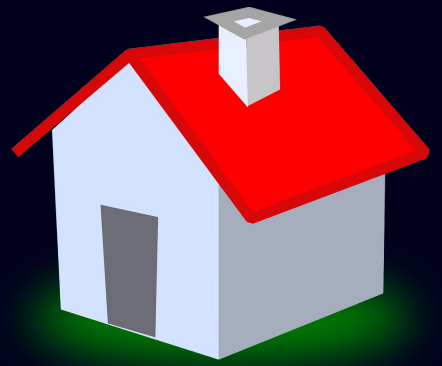
You 😐

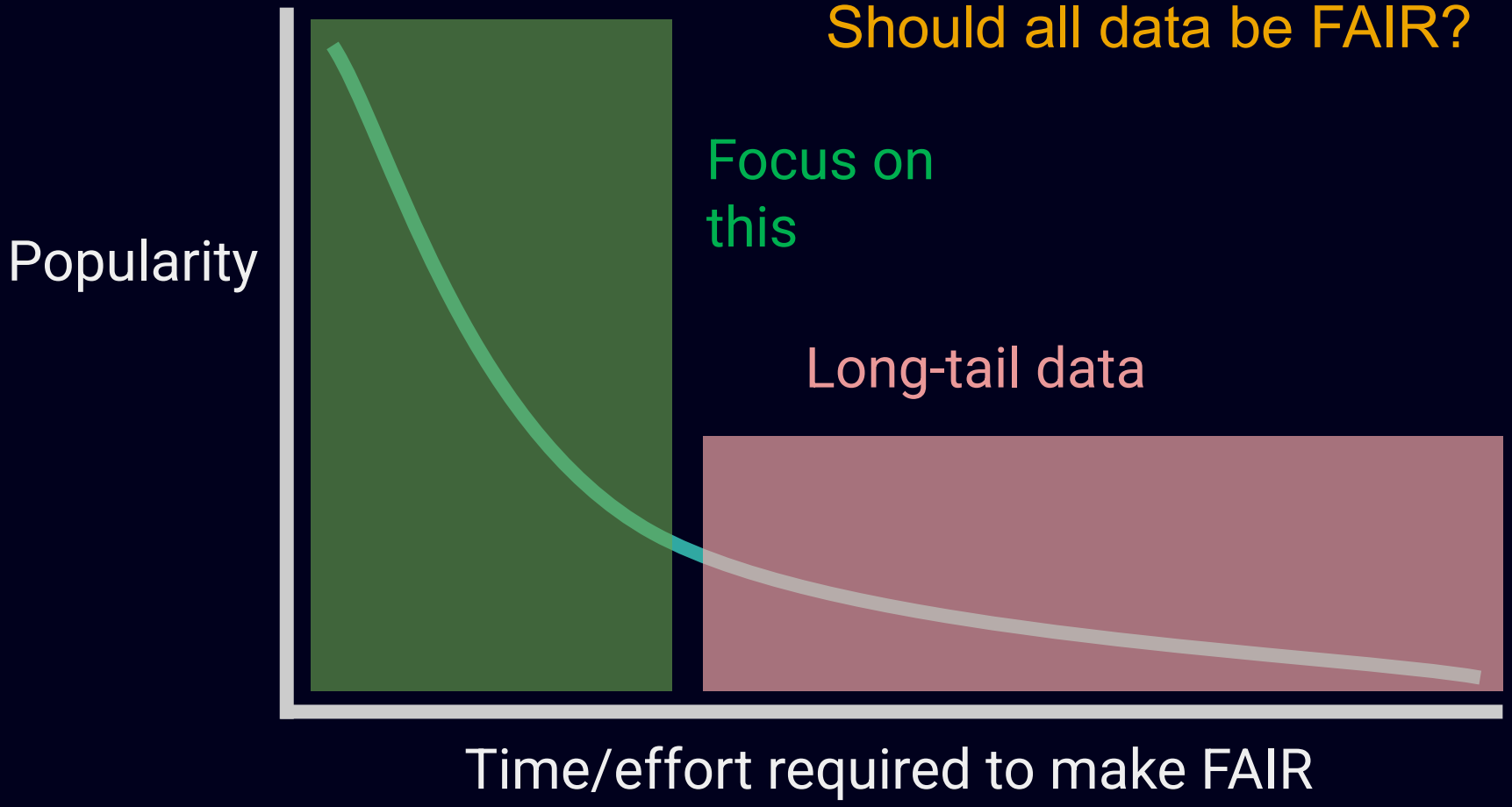
The general public 😞

Data that are not FAIR will never be available to the public

Publish today based on what should be possible in the future

Or in the future we will have no data from today




Should all data be FAIR?
Focus on this
Popularity
Long-tail data
Time/effort required to make FAIR

Human-readable data

- Descriptive terms
  - Controlled vocabularies where possible
- Don’t rely on cell formatting to make your data understandable
- Software-independent
- Templates
- README

https://www.unis.no/how-to-publish-unis-data/

How to publish UNIS data

The most important product of research is knowledge. At UNIS, most of the knowledge is derived from in-situ observations of nature during field work. To preserve this information, UNIS has teamed up with Svalbard Integrated Earth Observation System (SIOS) to make scientific data available and easily accessible according to the FAIR (Findable Accessible Interoperable Reusable) principles.

Here you will find a step-by-step procedure to go from digital measurements to data availability via the SIOS data portal.

- Data Collection, Preparation and Quality Control +
- Suitable Data Formats +
- Verification and Review +
- Selecting a Data Centre +
- Making your data available via SIOS +
- Citing your data (or someone else's) +