SCIENTIFIC BENEFITS ON OPEN DATA

Timo Vesala
Department of Physics
Department of Forest Sciences

Thanks: Ari Asmi, Eija Juurola, Pasi Kolari, Markku Kulmala, Werner Kutsch, Alex Vermeulen

- Background and motivation
- SMEAR
- ICOS

“We” have opened our data since it :

- is produced by tax-payers money

- fosters collaboration

Some data may not be open:

- technical limitations (raw data)

- special data (intensive campaigns)

- Dario Papale (University of Tuscia) quoted someone:
  "Sharing data is like sharing wife…"

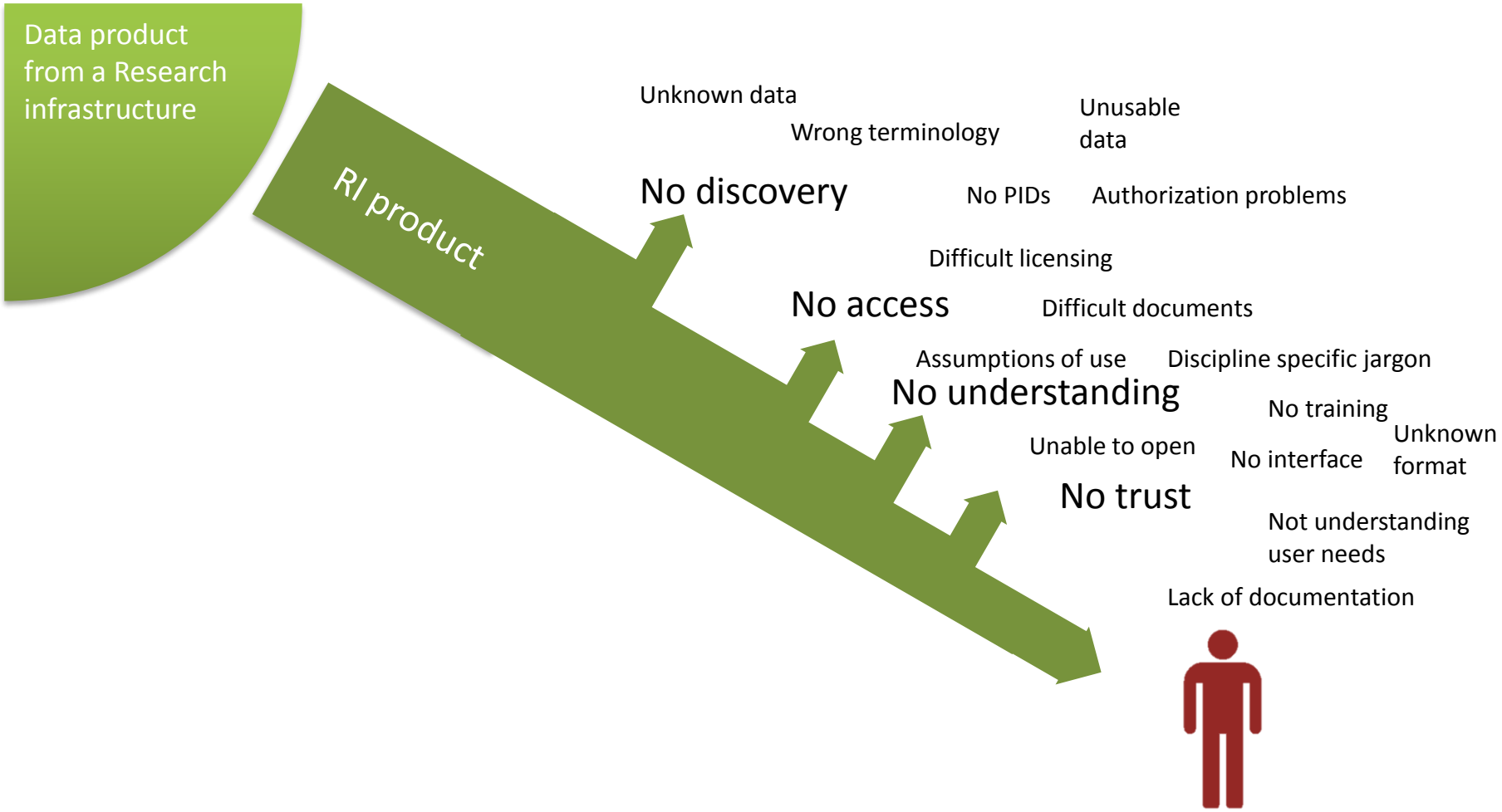
- Russia: just difficult to get data, although exceptions exist

- China: difficult to get data because of competition,
  although exceptions exist

- I have never suffered on any misusage of data
  (publications without co-authorship offers,
  overlapping what we were doing…), but gained a lot

- In fact, often the problem is that we have too much data
  and no time to analyze and write all articles!

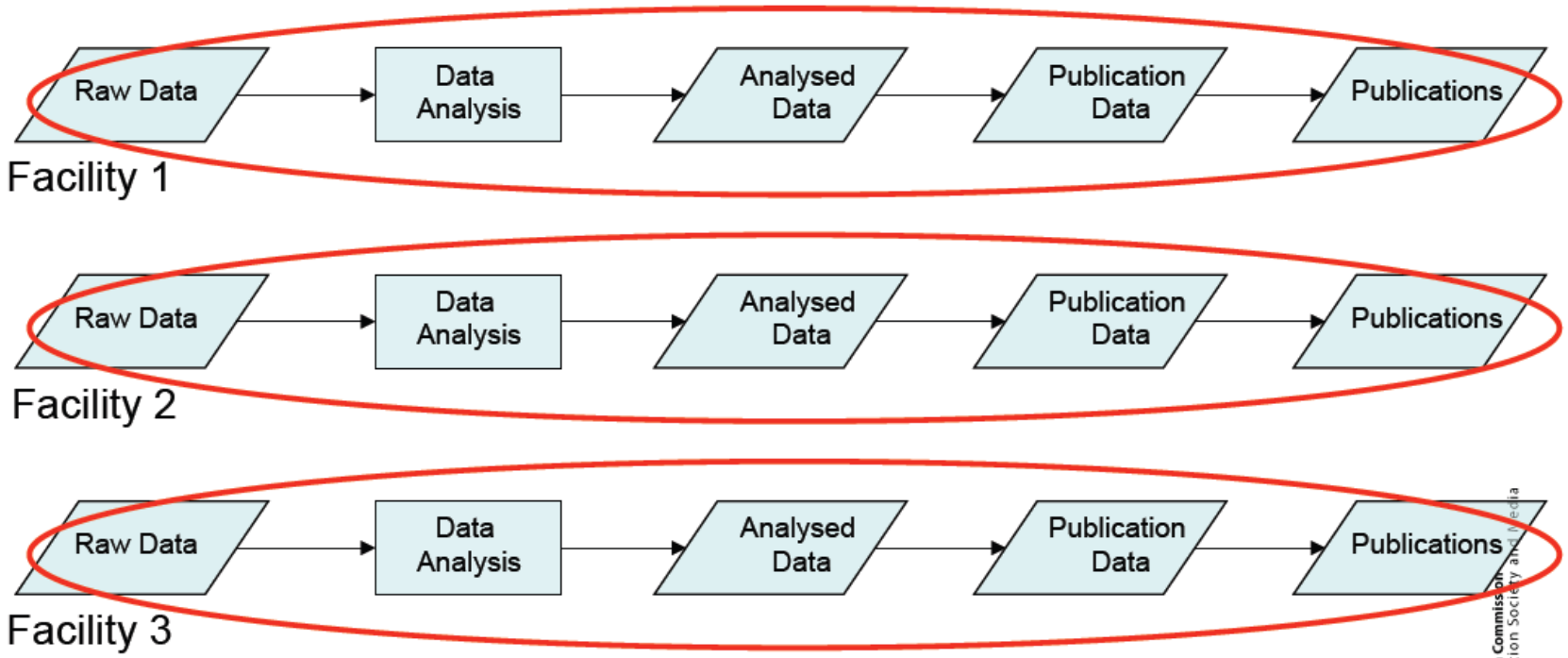
Barriers of information



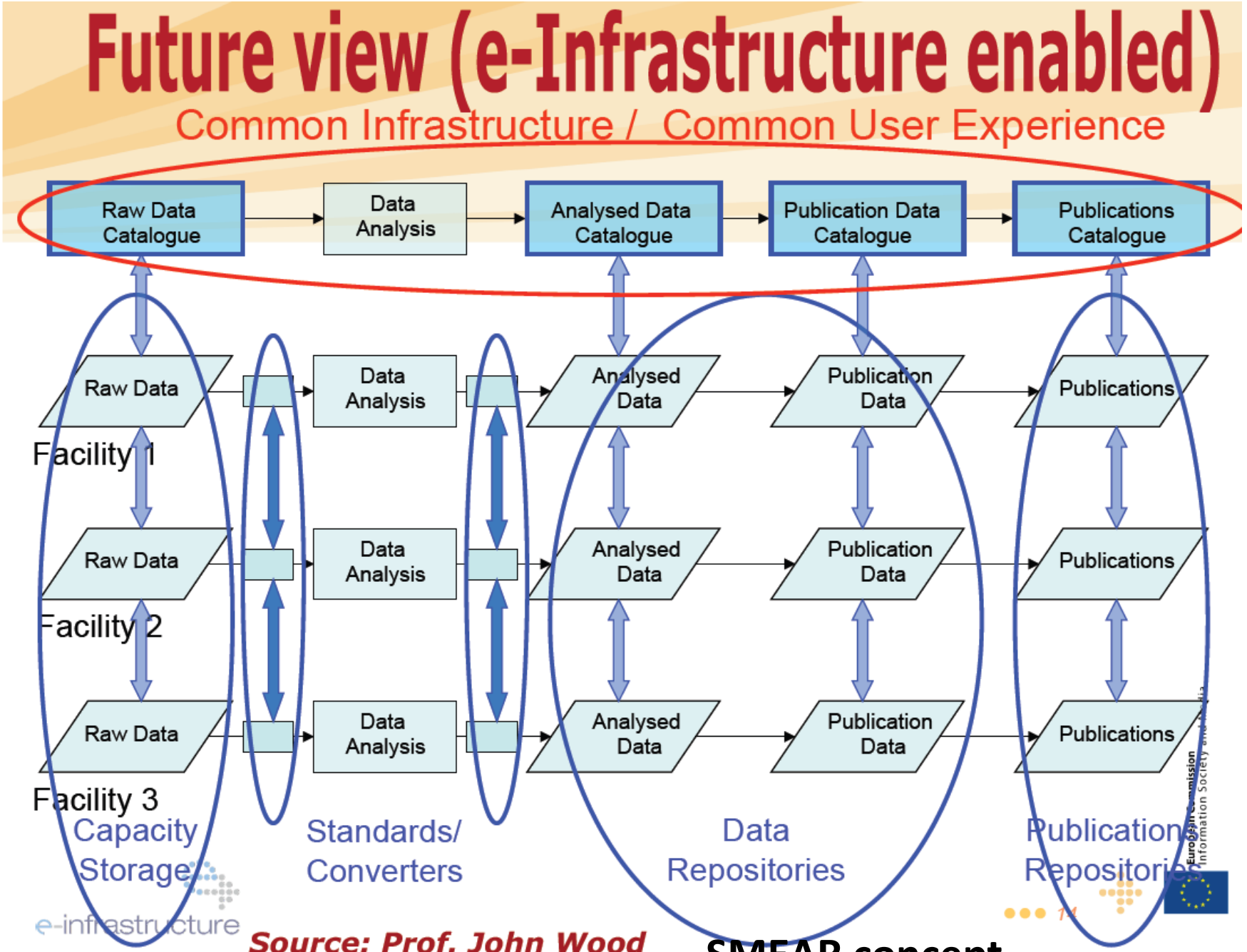
A major challenge for the infrastructures is to find the best compromise between optimized access for users and sufficient visibility and acknowledgement of data providers.

Current view

Distinct Infrastructures / Distinct User Experiences



Source: Prof. John Wood



SMEAR concept

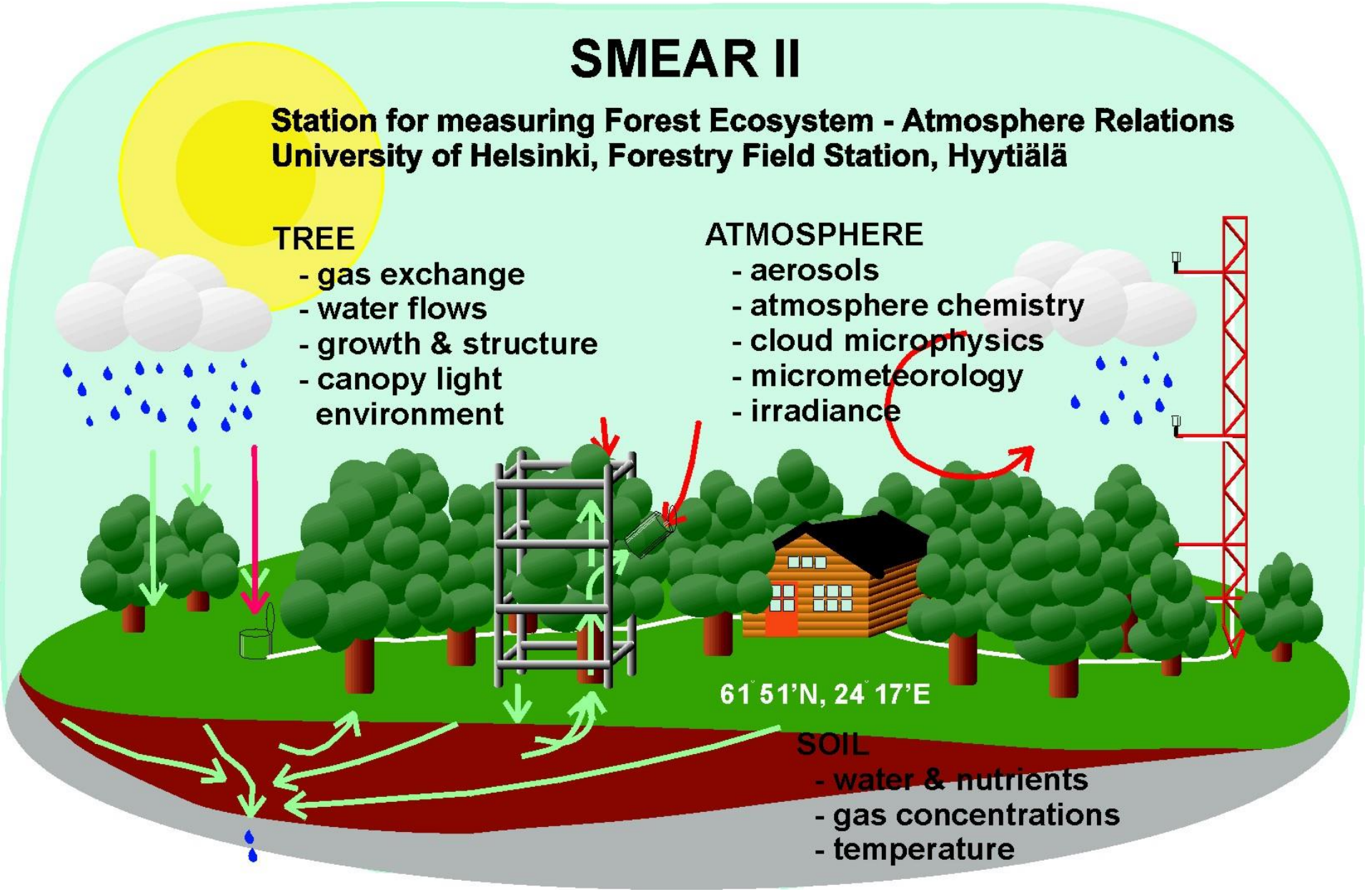
SMEAR

SMEAR stations

- SMEAR = Station for Measuring Ecosystem-Atmosphere Relations
- Univ. Helsinki, Dept Physics and Forest Sciences, SMEAR IV Univ. Eastern Finland and FMI
- Continuous long-term field measurements at the stations
  - meteorology, soil, vegetation, fluxes, atmospheric chemistry, aerosols
  - mostly time series at fixed locations
- Short-term campaigns at fixed sites or moving platforms
  - set of measurements in specific field
  - also geospatial data

SMEAR II

Station for measuring Forest Ecosystem - Atmosphere Relations
University of Helsinki, Forestry Field Station, Hyytiälä



SMEAR II Hyytiälä

SMEAR III Hotel Torni

SMEAR II Siikaneva 1

SMEAR II Siikaneva 2

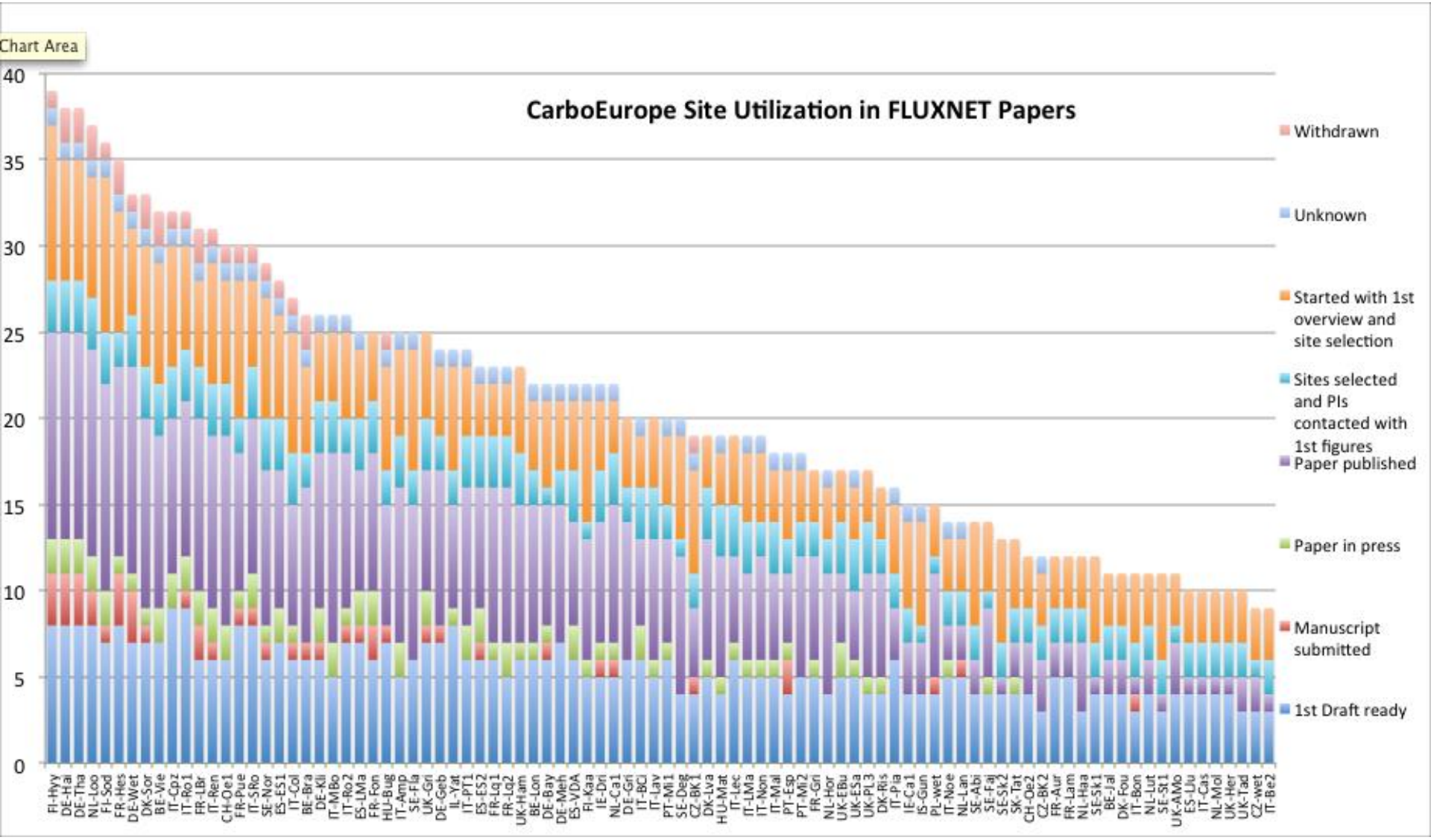


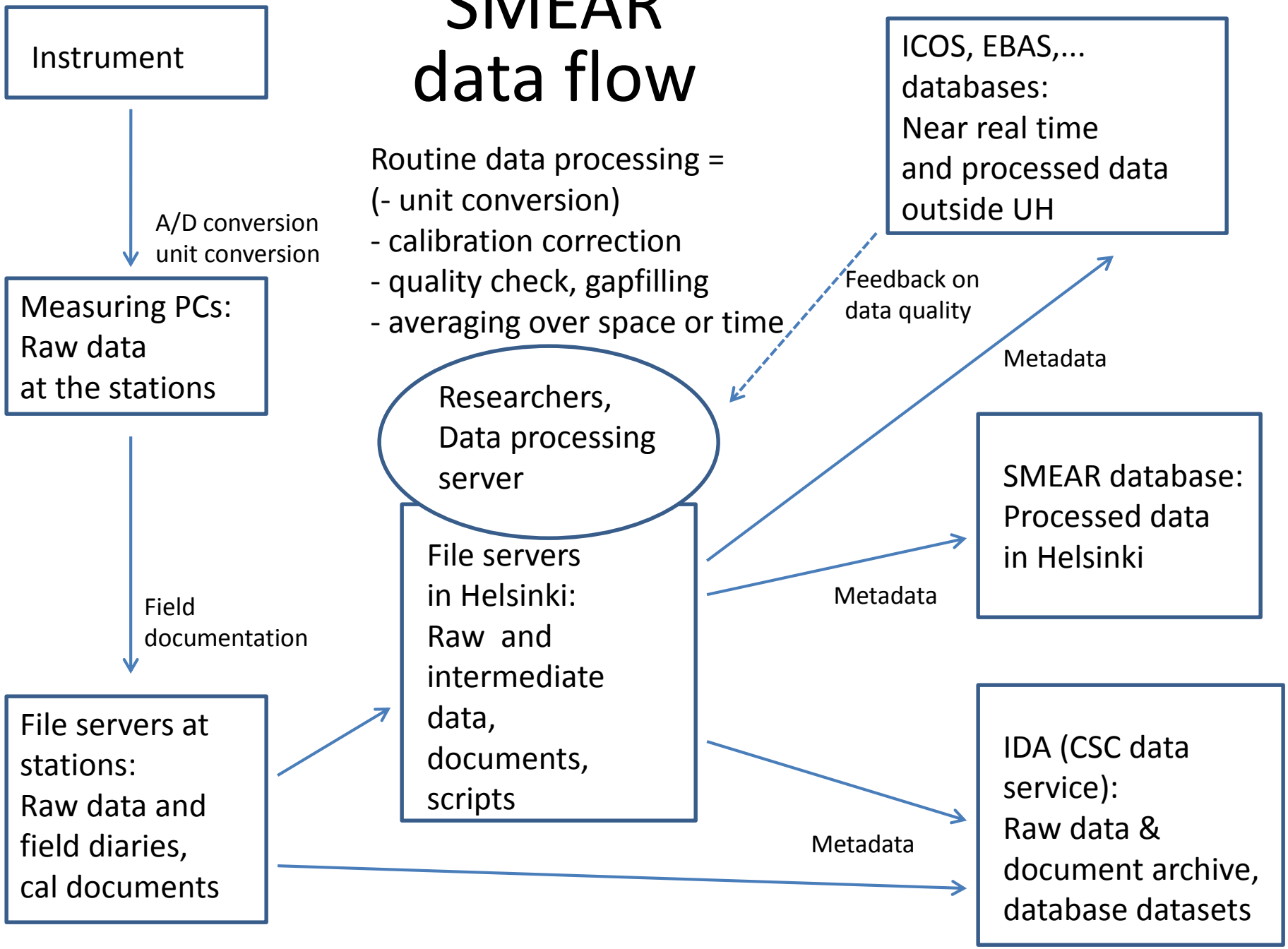
Chart Area
CarboEurope Site Utilization in FLUXNET Papers
Withdrawn
Unknown
Started with 1st overview and site selection
Sites selected and PIs contacted with 1st figures
Paper published
Paper in press
Manuscript submitted
1st Draft ready
0
5
10
15
20
25
30
35
40
FI-Hyy
DE-Hai
DE-Tha
NL-Loo
FI-Sod
FR-Hes
DE-Wet
DK-Sor
BE-Vie
IT-Cpz
IT-Ro1
FR-LBr
IT-Ren
CH-Oe1
FR-Pue
IT-SRo
SE-Nor
ES-ES1
IT-Col
BE-Bra
DE-Kli
IT-MBo
IT-Ro2
ES-LMa
FR-Fon
HU-Bug
IT-Amp
SE-Fla
UK-Gri
DE-Geb
IL-Yat
IT-PT1
ES-ES2
FR-Lq1
FR-Lq2
UK-Ham
BE-Lon
DE-Bay
DE-Meh
ES-VDA
FI-Kaa
IE-Dri
NL-Ca1
DE-Gri
IT-BCi
IT-Lav
PT-Mi1
SE-Deg
CZ-BK1
DK-Lva
HU-Mat
IT-Lec
IT-LMa
IT-Non
IT-Mal
PT-Esp
PT-Mi2
FR-Gri
NL-Hor
UK-EBu
UK-ESa
UK-PL3
DK-Ris
IT-Pia
IE-Ca1
IS-Gun
PL-wet
IT-Noe
NL-Lan
SE-Abi
SE-Fi
SE-Sk2
SK-Tat
CH-Oe2
CZ-BK2
FR-Aur
FR-Lam
NL-Haa
SE-Sk1
BE-Jal
DK-Fou
IT-Bon
NL-Lut
SE-St1
UK-AMo
ES-Llu
IT-Cas
NL-Mol
UK-Her
UK-Tad
CZ-wet
IT-Be2

Example of data flow

1. Measurement
> 2. raw observation
> 3. corrections, conversions
> 4. 'final observation'
> 5. processing
> 6. processed data
> 7. analysis
> 8. analysis results
> 9. visualizations
>10. publication
>11. published data

1. Thermometer
> **2. voltage record**
> 3. mV→ K, calibration correction
> **4. temperature record in K**
> 5. averaging and filtering
> **6. mean temperature during 1 min**
> 7. frequency analysis
> **8. analysis results**
> 9. figure of periodicity in T
>10. 'Hyytiälä T has 1-day periodicity'
>11. published data of T

SMEAR data flow



- 1500 variables measured
- 40 Gb day

SMEAR data project

Started in 2011

Aims

- Reliable long-term data storage
- Easy access to all data
- Formal documentation (metadata)

Planning and implementation with CSC – IT Centre for Science (Tieteen tietotekniikan keskus Oy)

SmartSMEAR

- Browser interface of SMEAR database
  http://avaa.tdata.fi/web/smart/smear
  - visualization and download of observational data, derived variables and calculated air mass back-trajectories
  - opened in Dec 2013, at the moment provides limited set of data but will be expanded to cover all continuous measurements at SMEAR stations
- Metadata as headers in exported .csv files, .hdf5 files with embedded metadata
- Temporal averaging, filtering by quality level

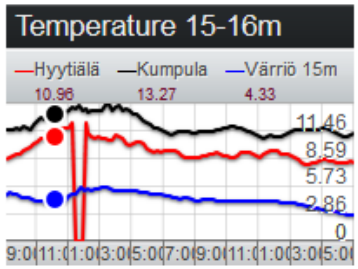
Variables:

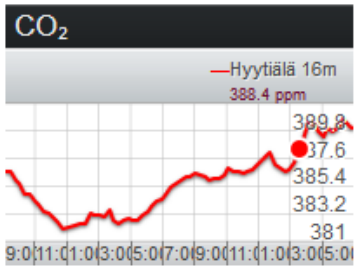
- Hyytiälä SMEAR II
  - Meteorology
  - Gas
  - Radiation
  - Soil
    - Soil surface temperature
    - Soil temperature A
    - Soil temperature B1
    - Soil temperature B2
    - Soil temperature C
    - Soil surface water content
    - Soil water content A
    - Soil water content B1
    - Soil water content B2
    - Soil water content C
    - Soil water potential A
    - Soil water potential B
    - Surface runoff
    - Surface runoff (2)
    - Subsurface runoff
    - Subsurface runoff (2)
    - Soil heat flux
  - Flux
- Kumpula SMEAR III
- Värriö SMEAR I
- Helsinki Hotel Torni

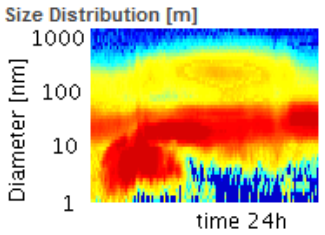
From: 2014-09-27 To: 2014-09-28 Shift: << Day >> Make Query

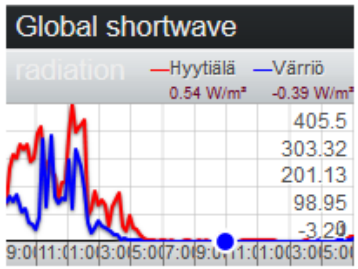
Quality Level: Any Averaging: None Averaging Type: None Arrival Height: 100m

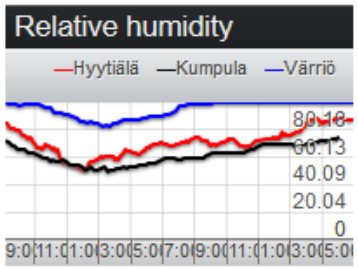
Reload main view

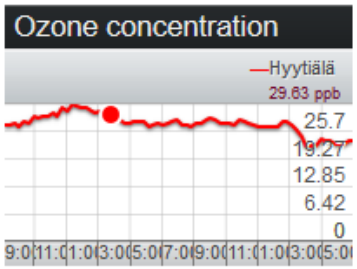


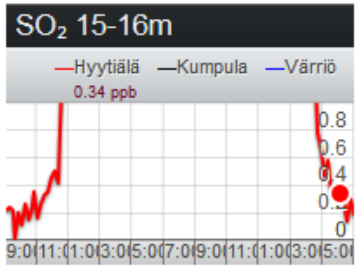


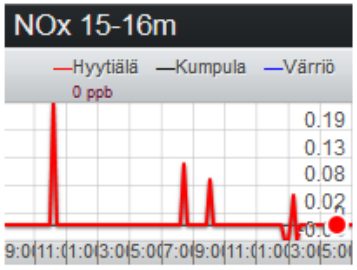


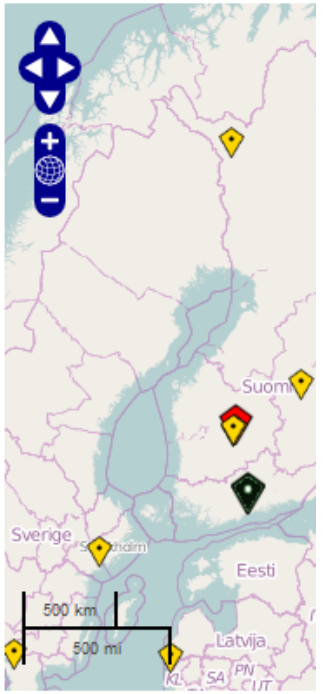












SmartSMEAR is data visua
atmospheric, flux, soil, tree
research stations of the Univ
for studying the connection b
the stationary measureme
measurements is provided a

The page consists of selec
variables for the desired time

The stations and variables
measured variables divided
variable name shows its dat
unit and source instrument c

Example of metadata supply file

title|Tree gas exchange at Värriö research station in 2011|
title.lang|en|
description|Chamber measurements of pine shoot gas exchange at SMEAR I at Värriö research station. Data columns: datetime ISO8601, quality flag (1=online processed, 2=quality checked), pine shoot CO2 exchange in µg CO2 per m2 all-sided needle area in second|
subject|photosynthesis|
subject|transpiration|
discipline|http://www.yso.fi/onto/okm-tieteenala/ta1183|
discipline|http://www.yso.fi/onto/okm-tieteenala/ta1171|
availability|direct_download|
rights|Licensed|
rightsDeclaration|http://creativecommons.org/licenses/by/4.0/|
publisher|University of Helsinki, Department of Physics, Division of Atmospheric Sciences|
owner|University of Helsinki, Department of Physics, Division of Atmospheric Sciences|
contact.email|atm-data@helsinki.fi|
contact.accessURL|http://www.atm.helsinki.fi/SMEAR|
reference|Kolari P, Lappalainen HK, Hänninen H, Hari P. 2007. Relationship between temperature and the seasonal course of photosynthesis in Scots pine at northern timberline and in southern boreal zone. Tellus B 59, 542-552.|
reference|https://wiki.helsinki.fi/display/~pkolari@helsinki.fi/Gas+exchange+of+pine+shoots|
author|Pasi Kolari|
author|http://orcid.org/0000-0001-7271-633X|
contributor.systemDesigner|Erkki Siivola|
temporalCoverage|start=2011-01-01T00:00:00;end=2011-12-31T23:59:59;|
...

Terms of data use

- Original data producer always has the intellectual rights
  - in some cases difficult to point out single person or whose contribution is important enough →ENVRI reference model
- Some projects may set their own terms
- Free public access and use of the data
  - principle also expressed by Finnish government
  - data mostly produced by public funding
- Data unofficially delivered under Creative Commons 4.0 Attribution licence
- In practise fair scientific use: acknowledge, cite, offer co-authorship

Interaction with other databases

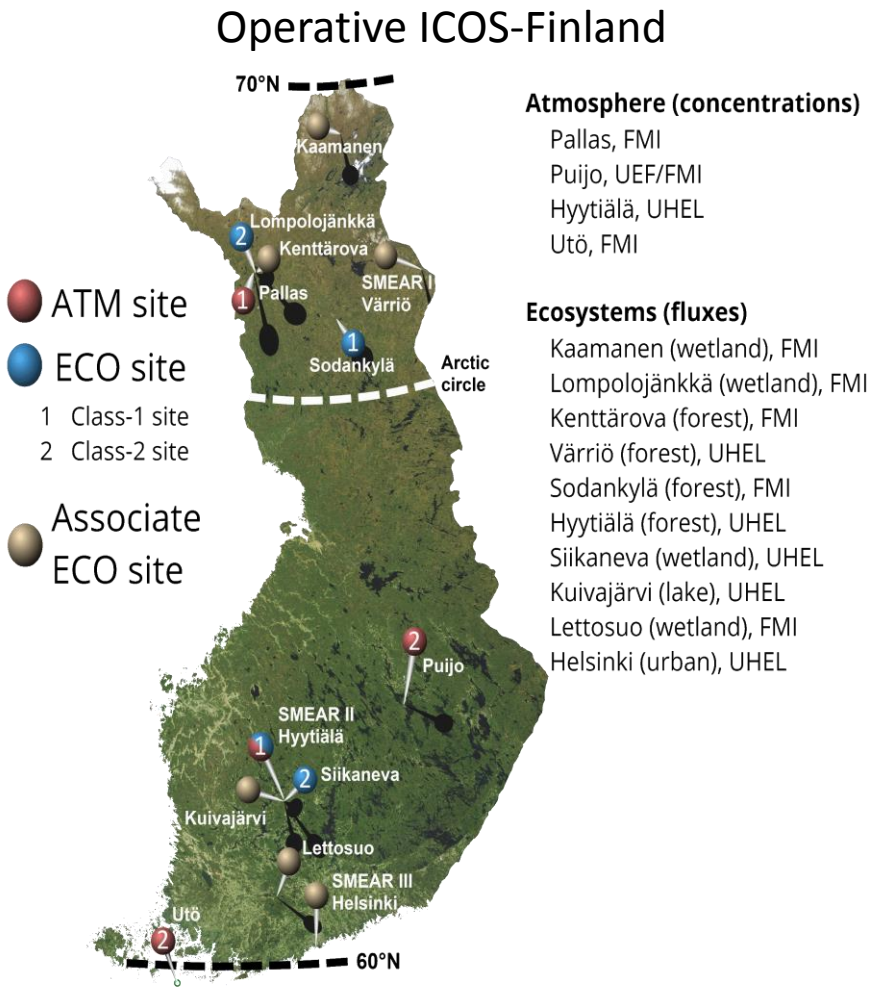
- Continuous
  - automated daily submission of raw instrument output files to ICOS ATC, metadata embedded
  - automated daily submission of online-processed data (1 or 30 min time resolution) to ICOS ETC database, metadata via browser User Interface
- Campaign/project databases
  - normally submission of processed data after the campaign has ended
  - diverse formats of metadata

ICOS – Integrated Carbon Observation System

- Presently 12 countries
- Atmosphere, ecosystems, oceans

ICOS-Finland

- ICOS observes greenhouse gases, water vapour and carbon, water and nitrogen cycles
  - Long time series
  - Quick data transmission and processing; harmonization and standardization of measurements and data processing
- >30 persons



- ICOS is a network with standardized stations and a common open data licence that only requires attribution via PID
- PID: data version, location, documentation
- ICOS data can be downloaded from the Carbon Portal.
- Each download request will be stored and connected to the PID.
- Users should cite the PID in their products (e.g. scientific publications)

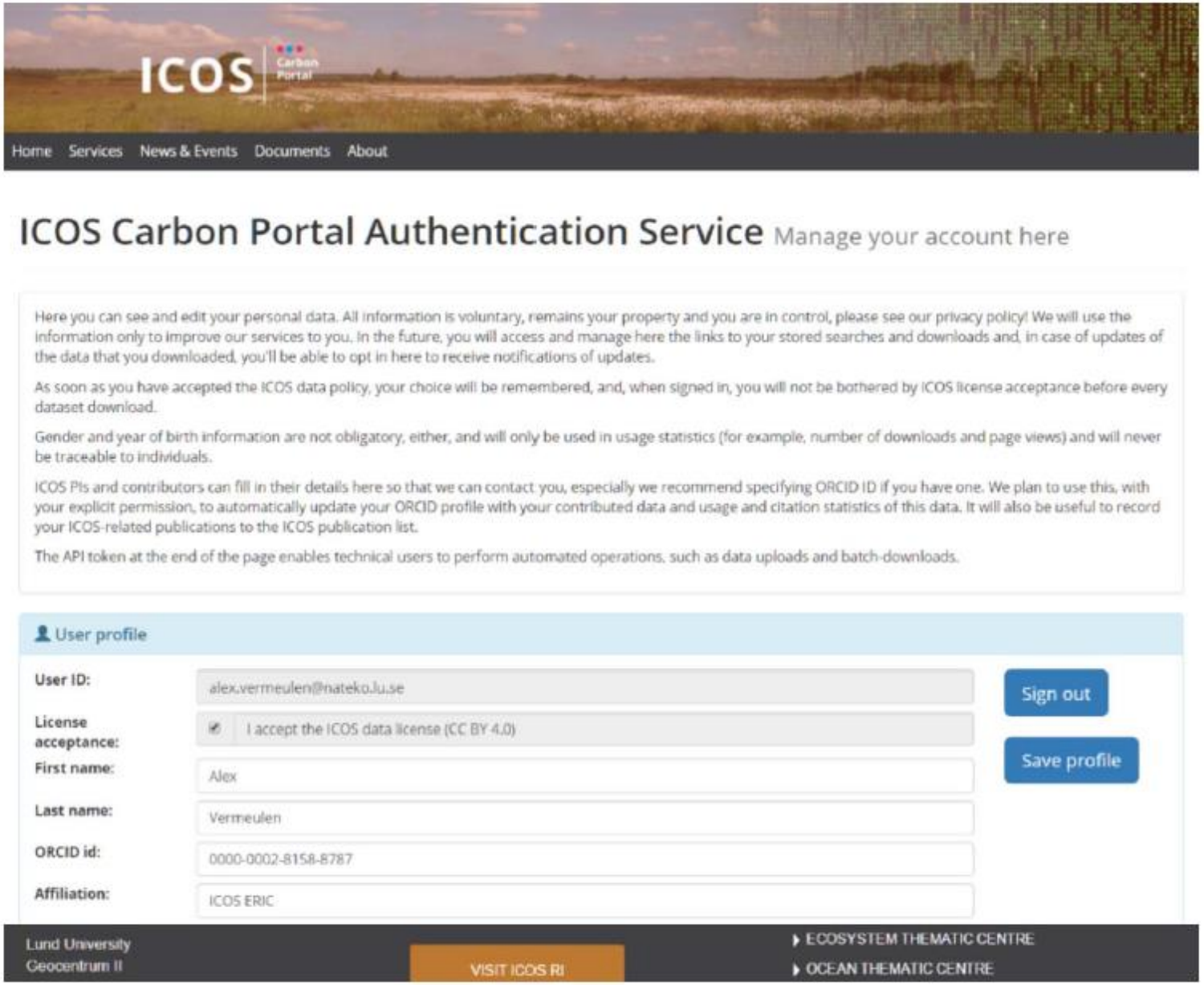
ICOS Data

- Level 0
  - raw sensor output (either mV or physical units)
- Level 1/NRT
  - calibrated and automatically assured data
- Level 2
  - final observation data products
- Level 3
  - elaborated data products, ICOS data

ICOS Data Policy

- Open and free of charge access to ICOS data products
- License: Creative Commons 4.0 BY
- Attribution required, ICOS requests:
  - Good scientific practice of use
  - Citation needed, ICOS provides suggestion
  - Acknowledgement or co-authorship when reasonable
- Redistribution is allowed but not necessary:
  - ICOS provides free, easy and most up to date access to all data!
  - No need for own storage, ICOS provides the best and easiest!
- Users are encouraged to contribute their products and distribute through ICOS CP through CC 4 BY

Authentication



What have we learned?

Benefits of data portals

- Easy discovery and retrieval of data
- Adds the visibility of the data
- Added value from data visualization
- Coordinated data processing and documentation
- Possible links to other data sources

These can also make data provider’s life easier

Distributed data storage

- Every project wants to set up own database - not good
  - keeping the data up to date and consistent across multiple databases is tedious
  - We already have problems keeping data consistent on file server <> database <> archive
- If the data are already in some international database link to that instead of submitting with primary data
  - typically ancillary meteorological data
- Interoperability of metadata and data output formats are important for cataloguing and discovery of the data

Metadata

- Documentation needs strict formats and good guidance, otherwise people write there anything or nothing at all
- Metadata format should be such that it's possible to map its attributes to widely used standards (Dublin Core, DDI)

Acknowledgements and terms of data use

- Researchers sometimes don't want to give their data before they have published it themselves
  - What does "publishing the data" mean?
- Loss of information about contributors
  - e.g. in Fluxnet database just PI and possible “collaborators” are mentioned as data providers
  - in some cases difficult to point out primary author/owner or those whose contribution is important enough to warrant ownership
- Metadata should provide more diverse contributor options than just “PI” or “owner”
  - ENVRI workflow reference model helps in the future?

Final words

- There are lots of benefits in collecting data and metadata under one data portal
- Data submission is nuisance to many researchers
- Reward from data submission and especially documentation needed
  - Technical aid in data processing
  - Consistent and transparent data quality check, feedback
  - Multiple processing procedures → uncertainty estimates (ICOS flux data)
  - Formal metadata files and different data file formats produced by the portal and exported
  - Visibility of all contributors in the metadata
  - Digital identifiers for datasets → acknowledgement in case journal citation does not exist and co-authorship is too much

ATM/UHEL SMEAR stations
Cumulative numbers