User Identification and Authentication for Geophysical Data Centers: Exploring a Difficult Transition

Florian Haslinger, Jerry Carter, Helle Pedersen, Jonathan Schaeffer,
Robert Casey, Javier Quinteros, Angelo Strollo

and further contributions from

Lesley Wyborn, Elisabetta D‘Anastasio, Jonathan Hanson, Mark Chadwick,
Christos Evangelidis, Jens Klump …

The (seismological) world today – paradise, almost…

Open, unrestricted, unconstrained ***anonymous*** access to (waveform) data and associated metadata is a long-standing paradigm in seismology *(to large extents also in other disciplines, e.g. GNSS)* – founded in the realisation that

where global observations are needed to do science, open sharing of data is fundamental

- at the foundation of FDSN (Seismology) and IGS (GNSS)
- implemented in international data centers like IRIS and ORFEUS for decades, also in almost all national / institutional data centers globally
- regarded as a ‘role model‘ in other fields of (Earth) sciences / geophysics, that often adopted a similar approach

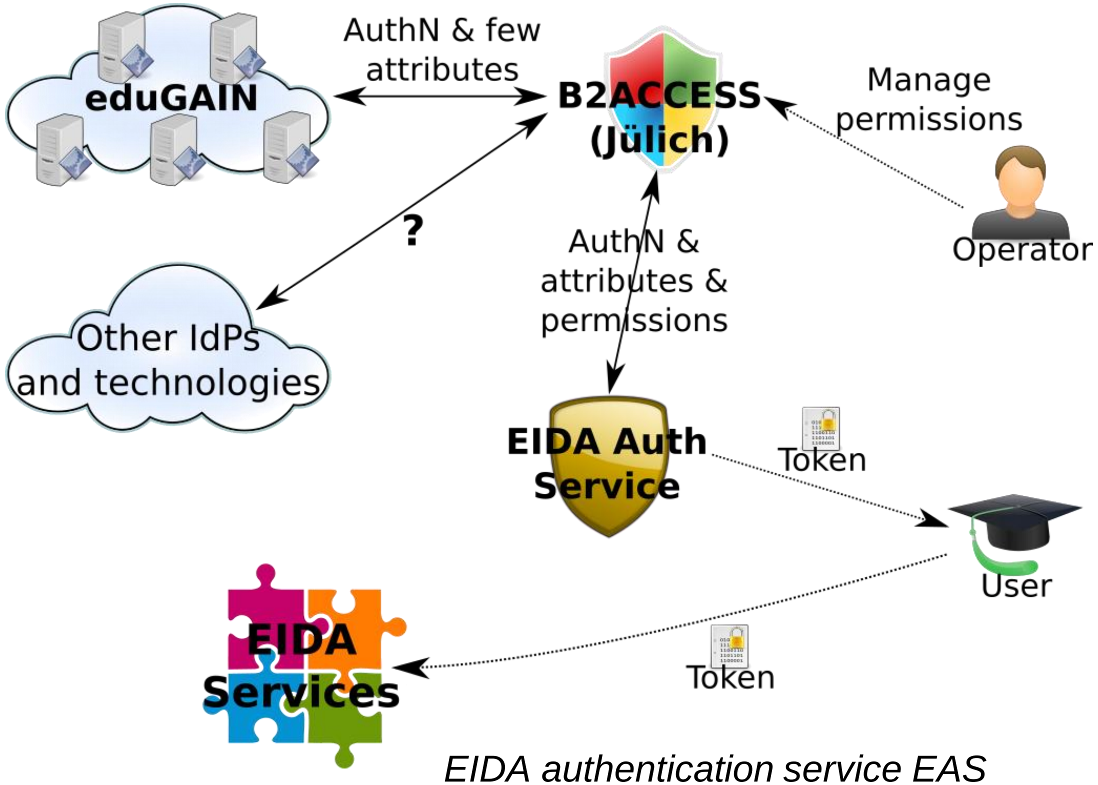
- today‘s tools and services to access and distribute data are built around that paradigm
  - while ***already also enabling the implementation of ‘access restriction***‘ e.g., for embargoed or otherwise restricted datasets or services, usually through specific user authentication and authorisation mechanisms
- in this way serving TB of data every day to the scientific community – and anybody else who would want it
- monitoring usage (if at all) by counting requests, volumes shipped, and (sometimes) their geographical origin

The challenge: *funders and other authorities want to know more… (I)*

Increasingly, data centers are asked by funders or other institutional authorities to report more details on ‘usage‘ of their data and services than they currently capture

To comply with that, ***user identification*** (*authentication*) will have to be implemented for (all) data access

- ✓ technically possible / feasible today (as part of established AAAI methods / infrastructures)
- ✓ partially already implemented as an option (e.g., EIDA authentication mechanism for fdsnws-dataselect, /queryauth request mechanisms)
  - ➢ making use of federated identity provision & management systems (GEANT / eduGAIN / B2access, …)
- but ‘generalisation‘ to any (data) access will be a clear paradigm shift for us (seismologists at least)



EIDA authentication service EAS
© Javier Quinteros, GFZ Potsdam

Usage data collection today: *ORFEUS-EIDA Data Centres*

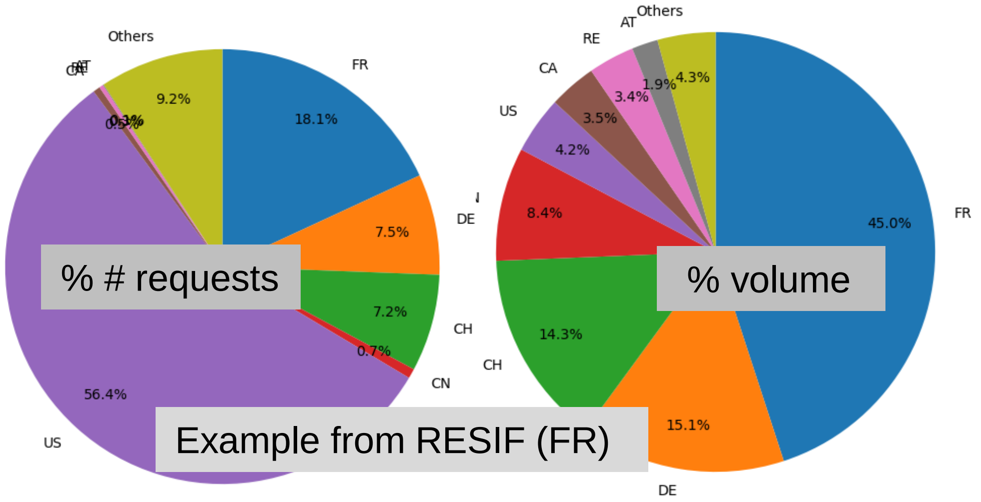
For **access to restricted data**, ORFEUS has an Authentication/ Authorization System (EAS) in production supporting eduGAIN (via B2ACCESS).

- If users log in at their home institutions, only attributes provided by eduGAIN are part of their profile.
- These attributes always respect the normatives of each origin region (e.g country, institution).
- Users receive a token to access the data.
- After some short time logs of the requests received are anonymized and stored in cumulative form

For open access, only IP-address of requester is logged, but also anonymized and deleted after some time.

Statistics are collected and hosted on a single database allowing analysis about

- use of each datacenter
- data distribution from shared networks (like AlpArray)



Statistical information logged (anonymized)

Datacentre,
Date,
Seismic Network,
Station Code,
Location,
Channel,
Country,
Cumulative amount of:
    Bytes,
    Requests,
    Successful requests,
    Failed Requests.

The challenge: *funders and other authorities want to know more… (II)*

Increasingly, data centers are asked by funders or other institutional authorities to report more details on 'usage' of their data and services than they currently capture
To comply with that, ***user identification*** (*authentication*) will have to be implemented for (all) data access

- what information exactly is expected by those asking is often not clearly defined (yet?)
  - levels of usage characterisation / user individualisation; counting requests and/or volumes; access 'by dataset'; … => ***potential issues with PII / GDPR*** => (data) management overhead
  - authentication alone (***confirming an identity***) ***may not be enough*** – profiling (purpose of use) needs even more information (and may change for same user from access to access)
- (anecdotal) experience of others indicates that ***usage may drop with enforced authentication***
- requiring authentication is an access restriction that may not be in line with open science 'best practice' (debatable)

  and likely creates at least some issues e.g. for 'ad-hoc' group activities involving access (teaching, training, outreach)

The consequences

Implementing user identification at data centers meets with some technical and managerial issues:

- ! Information management, privacy & security:
  - ? how to **avoid** / minimize the **collection of ‘sensitive personal data‘** (different interpretation in different legislations…)
  - ? what is needed to ‘manage‘ the **unavoidable personal data collections** (legal compliance of technical and managerial setups) => ref. personal identifyable information PII in particular in Europe / GDPR context
- ! user experience:
  - ? how to **ensure** that ‘**everybody**‘ (anybody on Earth with access to a computer) – even ‘non-individuals‘ (independent machines / software agents…) **can authenticate** & access – at **any time**
  - ? what does it take to adapt our existing (standard) data access services and the tools built upon & around them
    - In particular as authentication technologies are still evolving fast – adaptation will **not be one-time**
- ! resource needs:
  - even if **implementing user authentication** can be largely streamlined, it **will require** (some) **resources** at the data center - from a usually already strained budget (the more exhaustive the profiling, the more resources…).
    In turn, these resources are **not** available to **improve** user experience and **services**

Hey, but wow … *tracking usage may offer benefits for data centers and users*

There are some (apparent) benefits arising from personalized user tracking
– aside from fulfilling funder requirements

- ✓ informing service & tool development
- ✓ assisting users (with failed or ‘sub-optimal‘ requests)
- ✓ informing data owners / contributors about usage & users
- ✓ identifying the audience and patterns of use by that audience
- ✓ monitoring / managing outbound data volumes per user

Could these benefits also be realized (more effectively) through other means and activities?

- ❖ improving user feedback and communication mechanisms (fora, blogs, surveys …)
- ❖ promoting application, use and uptake of relevant (persistent) identifiers and solve existing issues (granularity, aggregation, (deep) resolution)
- ❖ …

OK, so let‘s move on …

Authentication & authorisation mechanisms are **required anyway** at our data centers at least **for some services**

- access to restricted data sets, connecting to cloud-based or HPC services, offering personalized work spaces, …

so **let‘s** keep coordinated and **develop common solutions** – in seismology but also beyond

- making use of evolving AAAI standards and technologies, improving ease-of-use as well as ease-of-maintenance

A general / generic user authentication requirement for everything should be (re)viewed very critically

- In dialog with the ‘requesting entity‘ and **mindful of the role of (community) data centers in the research life-cycle**
    - data centers serve as guaranteed long-term repositories for research output (and public data collection), ensuring its FAIRness
    - they are key players in the development and promotion of (community) standards for data and services
- with respect to user equity, implementation and maintenance effort, any other fallout

Usage data collection tomorrow – IRIS data services

IRIS Data Services will soon be implementing an identity management system to:

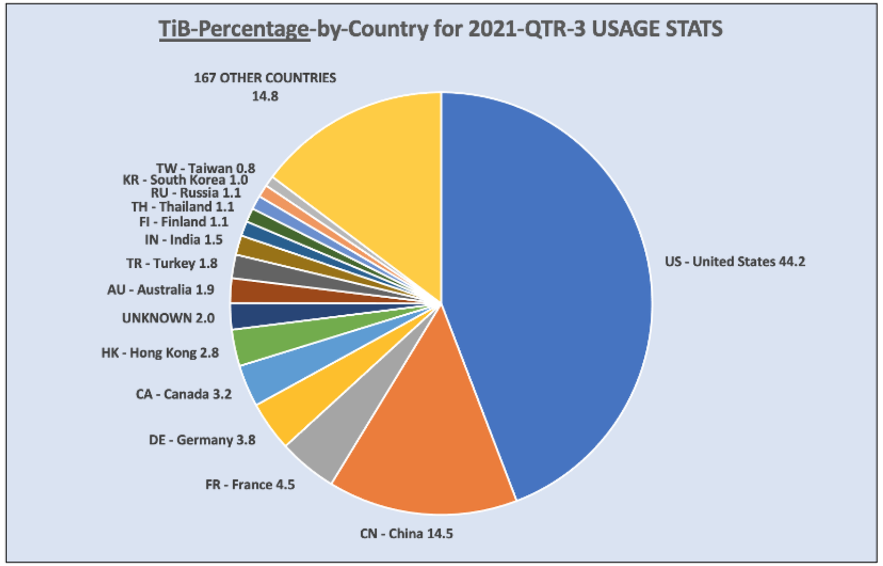
- Provide accountability to stakeholders for the funds we receive
- Gather information on data access to better serve our user community

Users who download data:

- Will register and build a profile
- Will receive a token to access data

Instead of tracking by IP Address:

- We will track by unique identifier codes
- We can understand their purpose of access
- We can tabulate institutional activity



Identity Profile (Example):

Name

Institution

Location

- Country
- State/province
- City

User Class

- Education (grp)
- Academic Res.
- Government
- Commercial
- Public

… and maintain the paradise (sort of)

The authors of this presentation came together in an ad-hoc manner triggered by IRIS‘ announcement that they would implement user identification for their data services by summer 2022.
We are currently discussing both the technical and the governance & strategic issues.

Technical issues will be further discussed and promoted through FDSN mechanisms (for seismology)
– expect some communication there soon

Governance & strategic issues will be further discussed in other upcoming venues (IUGG 2023, …)
and brought to relevant other bodies & initiatives (RDA, CODATA, ISC, …)

- Including all those connected issues in the FAIR data and open science context
  - identifiers, attribution, licenses and IPR, data protection and security, long term curation, long tail of science …

If you are interested to join the discussion, get in touch!

Further reading (suggestions)

The links below point to documents and other resources that we consider relevant and/or interesting in the context of the topic of this presentation

UNESCO recommendations on Open Science, 2021:
https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en

(federated) identity management, FAIR and open access from a different discipline (Biology):
https://www.fim4l.org/wp-content/uploads/2021/03/Open-Access-and-FIM-v4.pdf

two complementary reports by OECD/GSF and ICSU/WDS on international research data networks and sustainable research data repositories:
https://doi.org/10.1787/e92fa89e-en
https://doi.org/10.1787/302b12bb-en

A study from Germany / DFG on issues related to data tracking and use of usage data by academic publishers:
https://www.dfg.de/download/pdf/foerderung/programme/lis/datentracking_papier_en.pdf