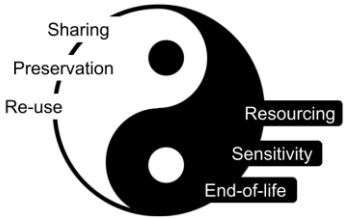
What Do We Really Need to Know? Balancing User Data Privacy vs Quality of Service vs Impact Reporting

Jens Klump, Lesley Wyborn, Rhys Francis

CSIRO

Australian National University



We acknowledge the Traditional Owners of the land, sea and waters, of the area that we live and work on across Australia. We acknowledge their continuing connection to their culture and we pay our respects to their Elders past and present.

What is personal information?

“[...] information or an opinion about an identified individual, or an individual who is reasonably identifiable (whether true or not or whether in a material form or not).”

- Privacy Act

The Situation

- Data repositories play a crucial role in supporting open data and open science by curating datasets and providing access services.
- Monitoring the usage of data repositories has traditionally involved simply logging access to datasets and collecting user feedback.
- Enhanced usage statistics can include finer geographical resolution and capture the purpose of data access, aiding funders and stakeholders in assessing the impact and outcomes of research infrastructures.
- Funders increasingly ask for more detailed usage statistics to gauge the impact of the services provided by a data centre.

The Complication

- The implementation of more granular user identification and usage tracking may introduce barriers to data access.
  - Sign-in is a deterrent and access barrier.
  - Sign-in causes problems in the anonymous review of data publications
- Concerns exist about the costs of implementing usage data collection, privacy legislation compliance, and the feasibility of providing detailed user information.

A real world example: from seismology

Lesley Wyborn

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 …

Warning:

This eclectic group came together in ad-hoc manner triggered by announcement of the US IRIS Seismology Data Center that they would implement user identification for their data services by summer 2022 to comply with requests from NSF.

Image Credit:
https://upload.wikimedia.org/wikipedia/commons/thumb/1/17/Warning.svg/2219px-Warning.svg.png

The (seismological) world was paradise, almost…

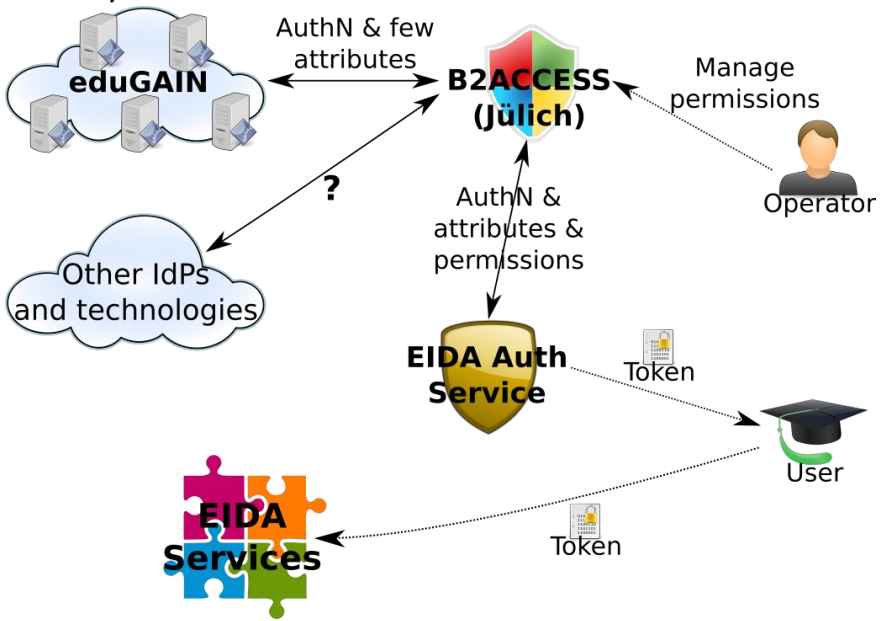
- Open, unrestricted, unconstrained ***anonymous*** access to (waveform) data and associated metadata was a long-standing paradigm in seismology founded in the realisation that

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

- Open science was implemented in international data centers like IRIS and ORFEUS for decades, also in almost all national / institutional data centers globally
- It was regarded as a 'role model' in other fields of (Earth) sciences that often adopted a similar approach
  - 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

But then the *funders and other authorities want to know more…*

- Increasingly, data centers are being 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*) had to be implemented for (all) data access
- Technically possible / feasible today (as part of established AAAI methods/infrastructures)



EIDA authentication service EAS

The challenge: *funders and other authorities want to know more: challenges*

- Levels of usage characterisation / user individualisation; counting requests and/or volumes; access ‘by dataset‘ => ***potential issues with legislation e.g. Personal Identifiable Information (PII)/ EU General Data Protection Regulation (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‘ (created issues with publishers)
- But is compatible with the OECD 2019 Scientific Paper No 65: *Reference framework for assessing the scientific and socio-economic impact of research infrastructures.*

Further reading (suggestions)

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

1. (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

1. 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

1. 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

The Question

- What are the benefits of implementing more detailed statistics and tracking mechanisms in data repositories?
- Can we do so without compromising user data privacy?
- What are the opportunities and challenges associated with enhancing usage monitoring?
- How do we balance usage monitoring against the needs of open science and data-driven research?

- What do we really need to know to measure impact and improve services?

Discussion