

Compliant Sharing of Sensitive Data with Dataverse and Lohpi

Aakash Sharma, Thomas Bye Nilsen, Håvard D. Johansen
UiT The Arctic University of Norway

LOHPI^Φ Team



Thomas
Bye Nilsen



Aakash
Sharma



Dag
Johansen



Håvard D.
Johansen



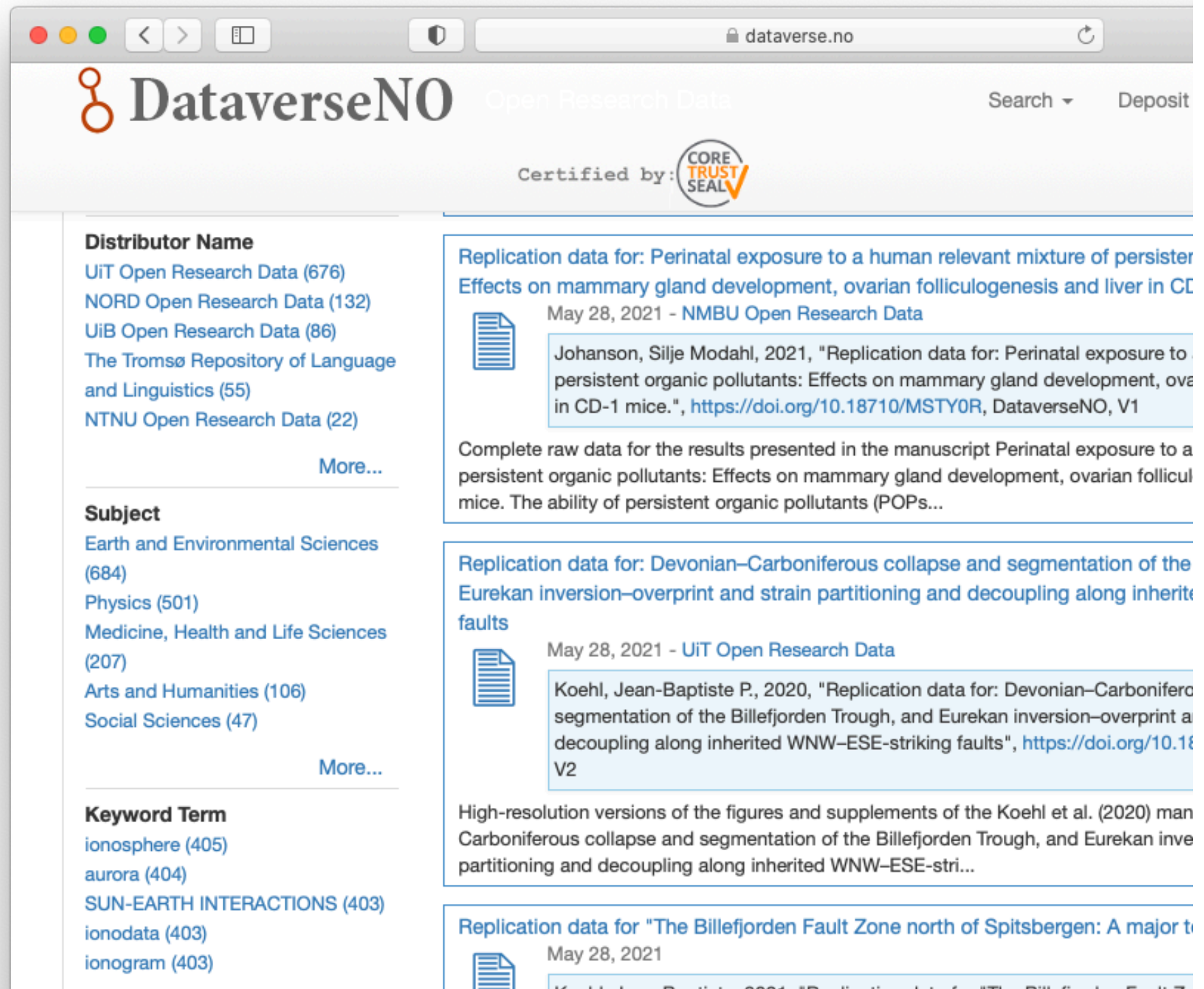
UiT The Arctic
University of Norway

Lab → NN

SpareBank 1
NORD-NORGE

Sharing data is the norm

- 147,000 Datasets
- 38.2 Million downloads
- Social Sciences 30,300 (21%)
- Medicine, Health and Life Sciences 7,420 (5.1%)



The screenshot shows the DataverseNO website interface. The header includes the logo, the text "Open Research Data", a search bar, and a "Deposit" button. A "Certified by: CORE TRUST SEAL" badge is visible. The main content area is divided into three columns: "Distributor Name", "Subject", and "Keyword Term".

Distributor Name

- UiT Open Research Data (676)
- NORD Open Research Data (132)
- UiB Open Research Data (86)
- The Tromsø Repository of Language and Linguistics (55)
- NTNU Open Research Data (22)

Subject

- Earth and Environmental Sciences (684)
- Physics (501)
- Medicine, Health and Life Sciences (207)
- Arts and Humanities (106)
- Social Sciences (47)

Keyword Term

- ionosphere (405)
- aurora (404)
- SUN-EARTH INTERACTIONS (403)
- ionodata (403)
- ionogram (403)

Search results for replication data are displayed in a list:

- Replication data for: Perinatal exposure to a human relevant mixture of persistent organic pollutants: Effects on mammary gland development, ovarian folliculogenesis and liver in CD-1 mice.**
May 28, 2021 - NMBU Open Research Data
Johanson, Silje Modahl, 2021, "Replication data for: Perinatal exposure to persistent organic pollutants: Effects on mammary gland development, ovarian folliculogenesis and liver in CD-1 mice.", <https://doi.org/10.18710/MSTY0R>, DataverseNO, V1
Complete raw data for the results presented in the manuscript Perinatal exposure to a persistent organic pollutants: Effects on mammary gland development, ovarian folliculogenesis and liver in CD-1 mice. The ability of persistent organic pollutants (POPs...
- Replication data for: Devonian–Carboniferous collapse and segmentation of the Billefjorden Trough, and Eurekan inversion–overprint and strain partitioning and decoupling along inherited faults**
May 28, 2021 - UiT Open Research Data
Koehl, Jean-Baptiste P., 2020, "Replication data for: Devonian–Carboniferous collapse and segmentation of the Billefjorden Trough, and Eurekan inversion–overprint and strain partitioning and decoupling along inherited WNW–ESE-striking faults", <https://doi.org/10.18710/2020.05.18>, V2
High-resolution versions of the figures and supplements of the Koehl et al. (2020) manuscript: Devonian–Carboniferous collapse and segmentation of the Billefjorden Trough, and Eurekan inversion–overprint and strain partitioning and decoupling along inherited WNW–ESE-striking faults...
- Replication data for "The Billefjorden Fault Zone north of Spitsbergen: A major tectonic feature in the Caledonides of the Arctic region"**
May 28, 2021
Koehl, Jean-Baptiste P., 2021, "Replication data for "The Billefjorden Fault Zone north of Spitsbergen: A major tectonic feature in the Caledonides of the Arctic region"

<https://dataverse.org/metrics>

Sensitive datasets cannot be public

- Trust issues [Bongartz et al. 2017]
- Easy to identify individuals [Salerno et al. 2017, Goodman and Meslin 2014]
- Consent revocations (GDPR, GPDPR)

T

SPØRRESKJEMA

T

Dato for utfylling

Dag mnd år

Kjente sykdommer:

1. Har du hatt angina eller hjerteinfarkt..... Ja Nei Usikker

2. Har du hatt hjerneslag eller drypp..... Ja Nei Usikker

3. Har du diabetes mellitus (sukkersyke)..... Ja Nei Usikker

Hvis ja: nåværende behandling..... Kun diett Tablett Insulin

Diagnoseår: Antall år på insulin

Røyking:

Har du noen gang vært daglig røyker..... Ja Nei

Har du noen gang vært daglig sigarillorøyker..... Ja Nei

Har du noen gang vært daglig piperøyker..... Ja Nei

Jeg har aldri vært daglig røyker

Hvis du er eller har vært daglig røyker:

Alder ved røykestart:..... Alder i år

Alder ved siste røykeopphør:..... Alder i år

Antall år med røyking:..... År

Antall sigaretter (eller tilsvarende) per dag:..... Antall

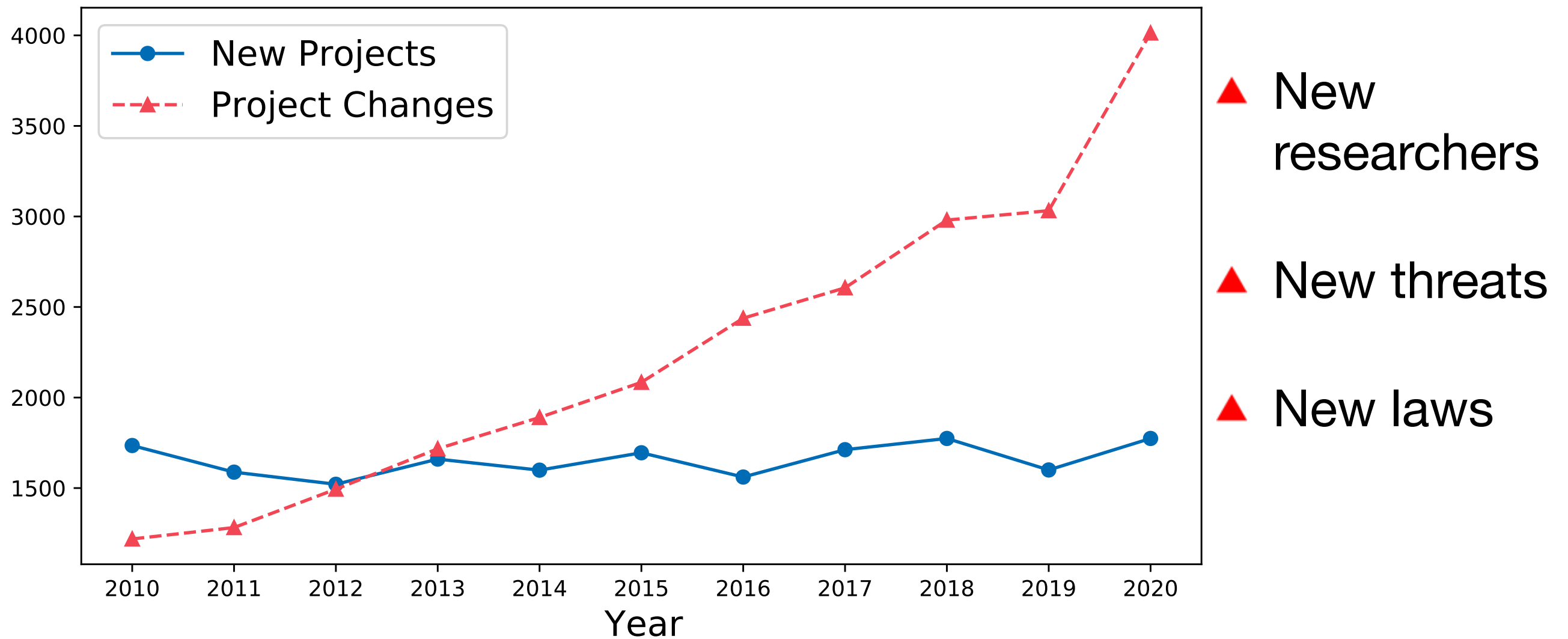
1 sigar = 4 sigaretter, 1 sigarillo = 2 sigaretter, 1 pipe = 2,5 sigaretter

Familieanamnese:

Antall søsken (inkludert deg selv):.....

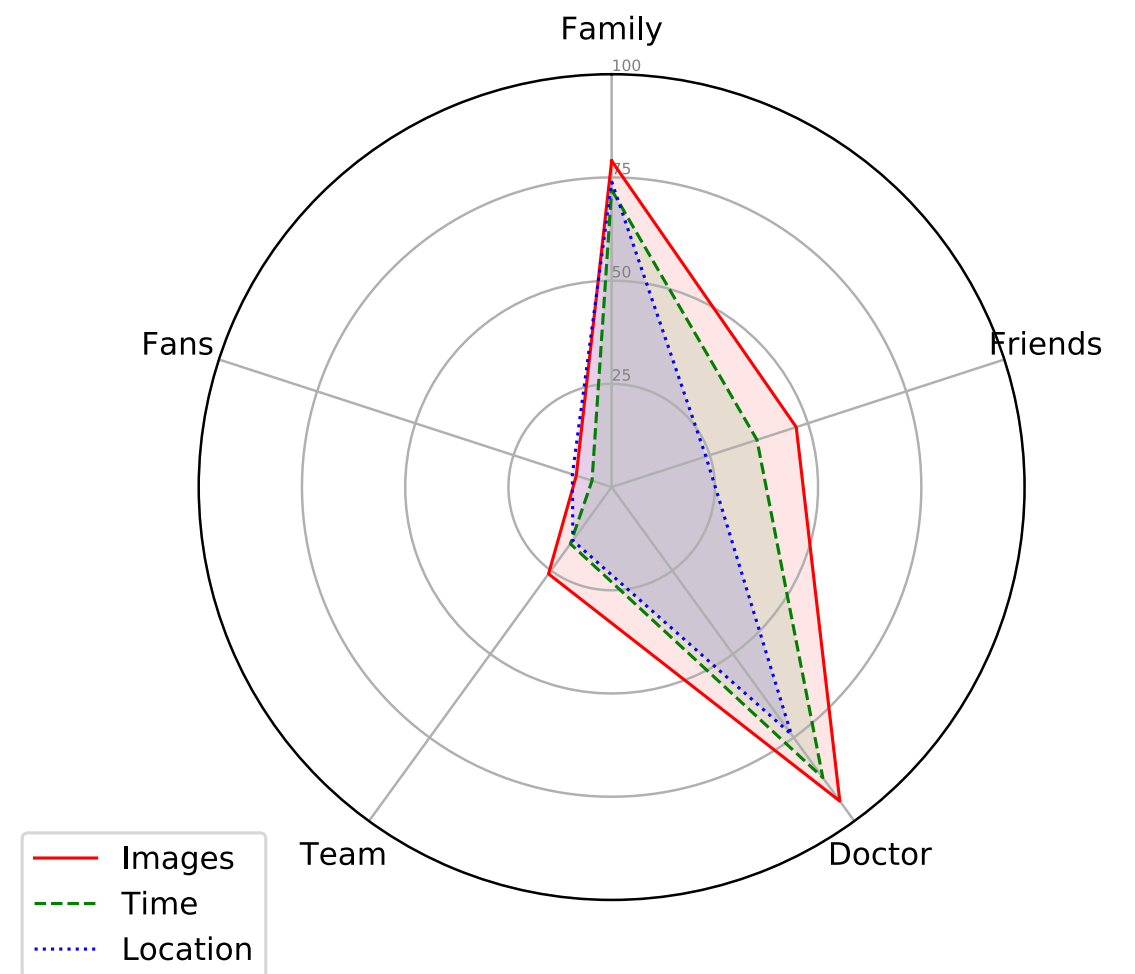
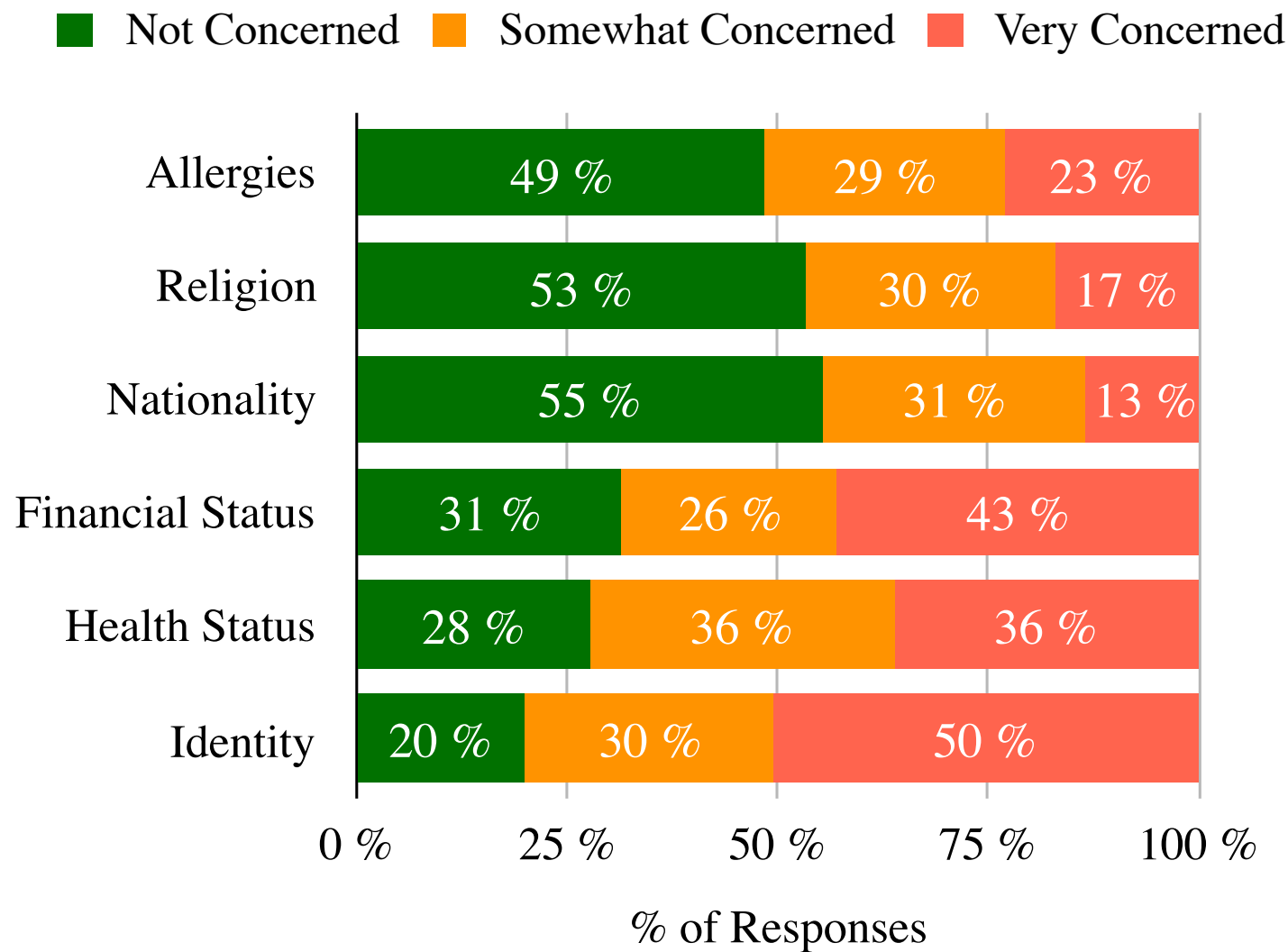


Projects are not static



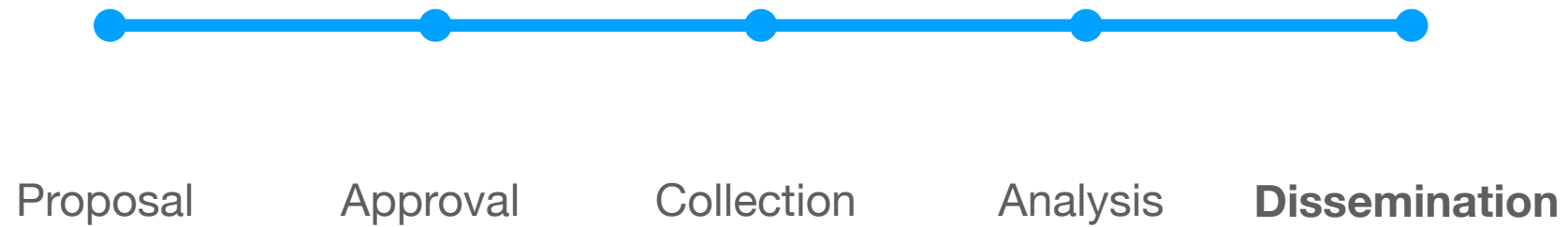
<https://rekportalen.no>

Opinions change

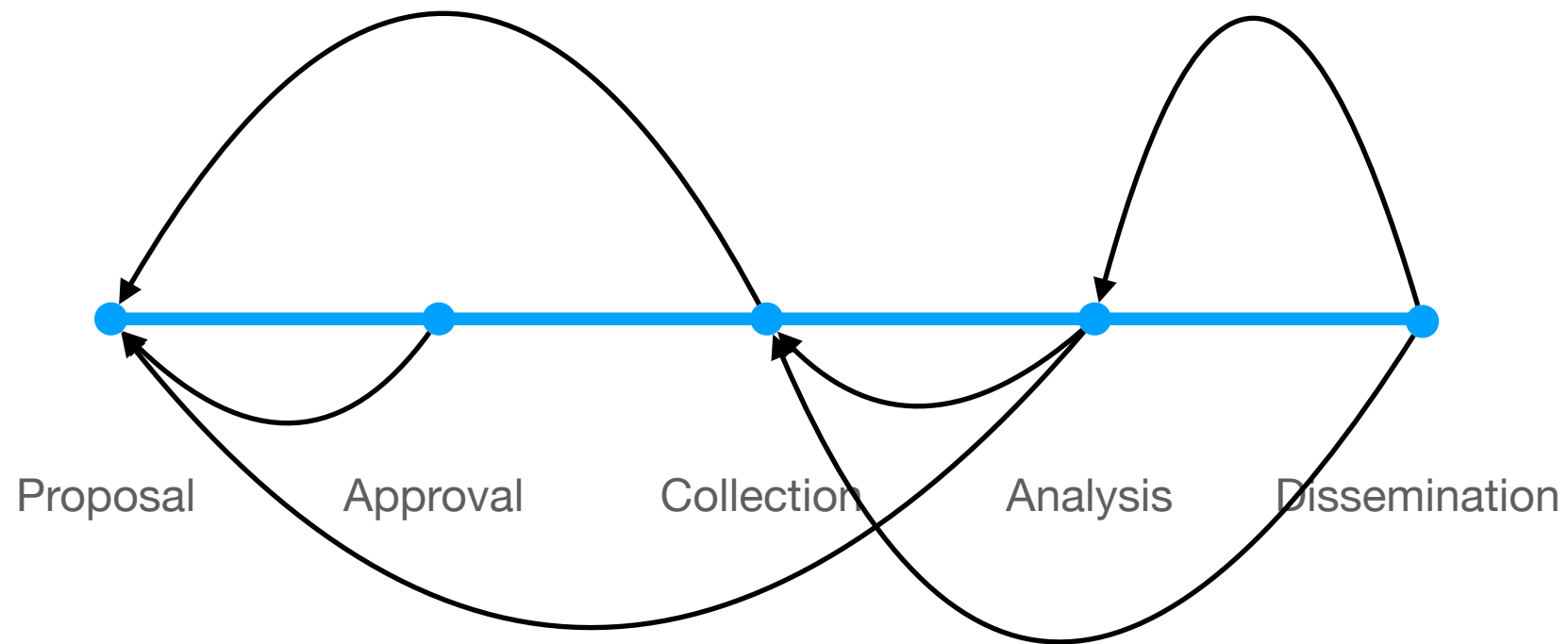


Sharma, Aakash, et al. "Privacy Perceptions and Concerns in Image-Based Dietary Assessment Systems: Questionnaire-Based Study." JMIR Human Factors 7.4 (2020): e19085.

A project's lifecycle



A project's lifecycle is non-trivial



We need system support for dynamic security policies

Sensitive data leads to silos

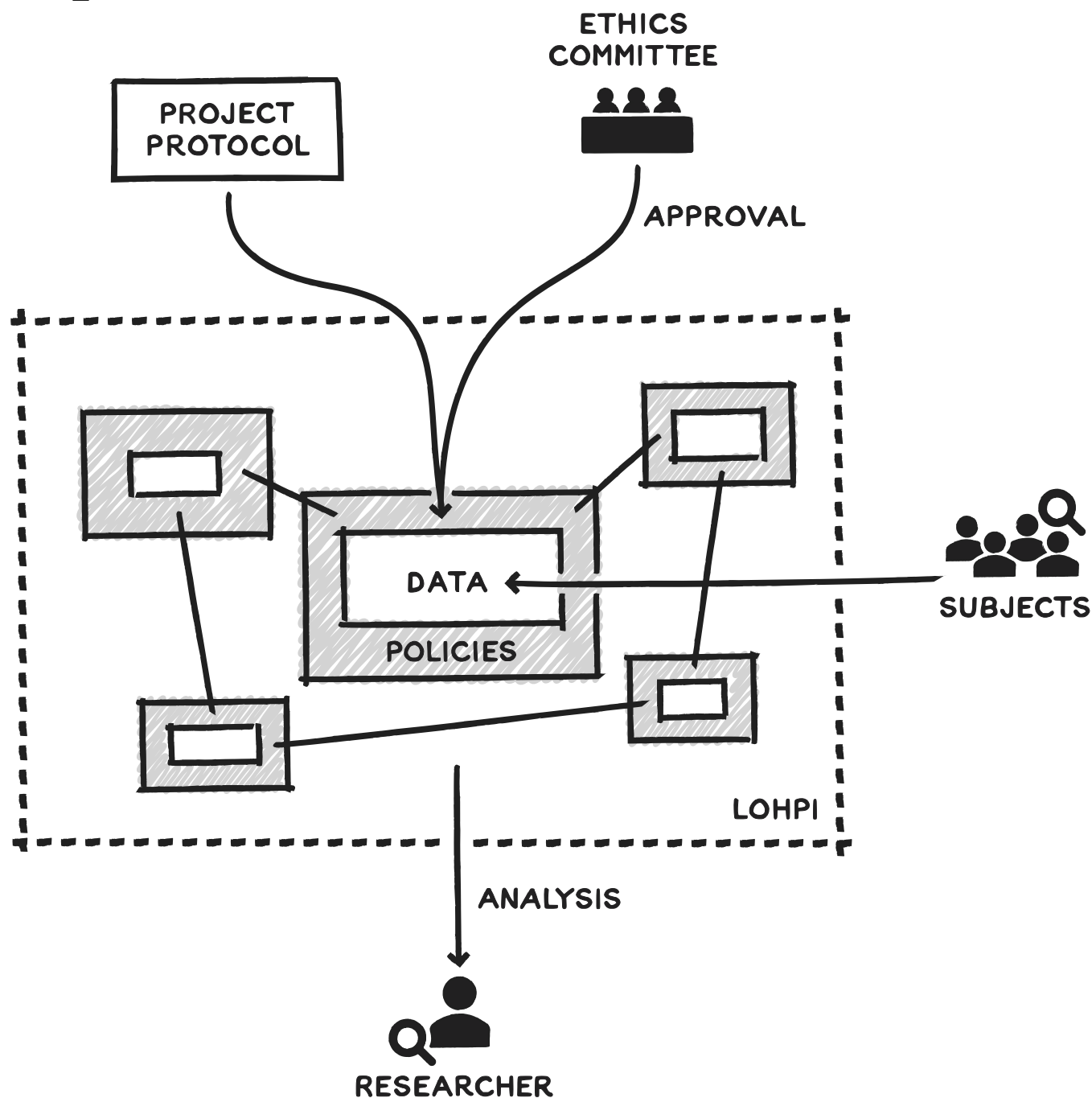
Tension between

- Open research (Dataverse)
 - ▶ Decentralized research model
 - ▶ Operate on institutional infrastructure
- Privacy risks
 - ▶ Projects operate in silos
 - ▶ Shared computational infrastructure (trusted by all parties, TSD)

Compliant sharing with Lohpi

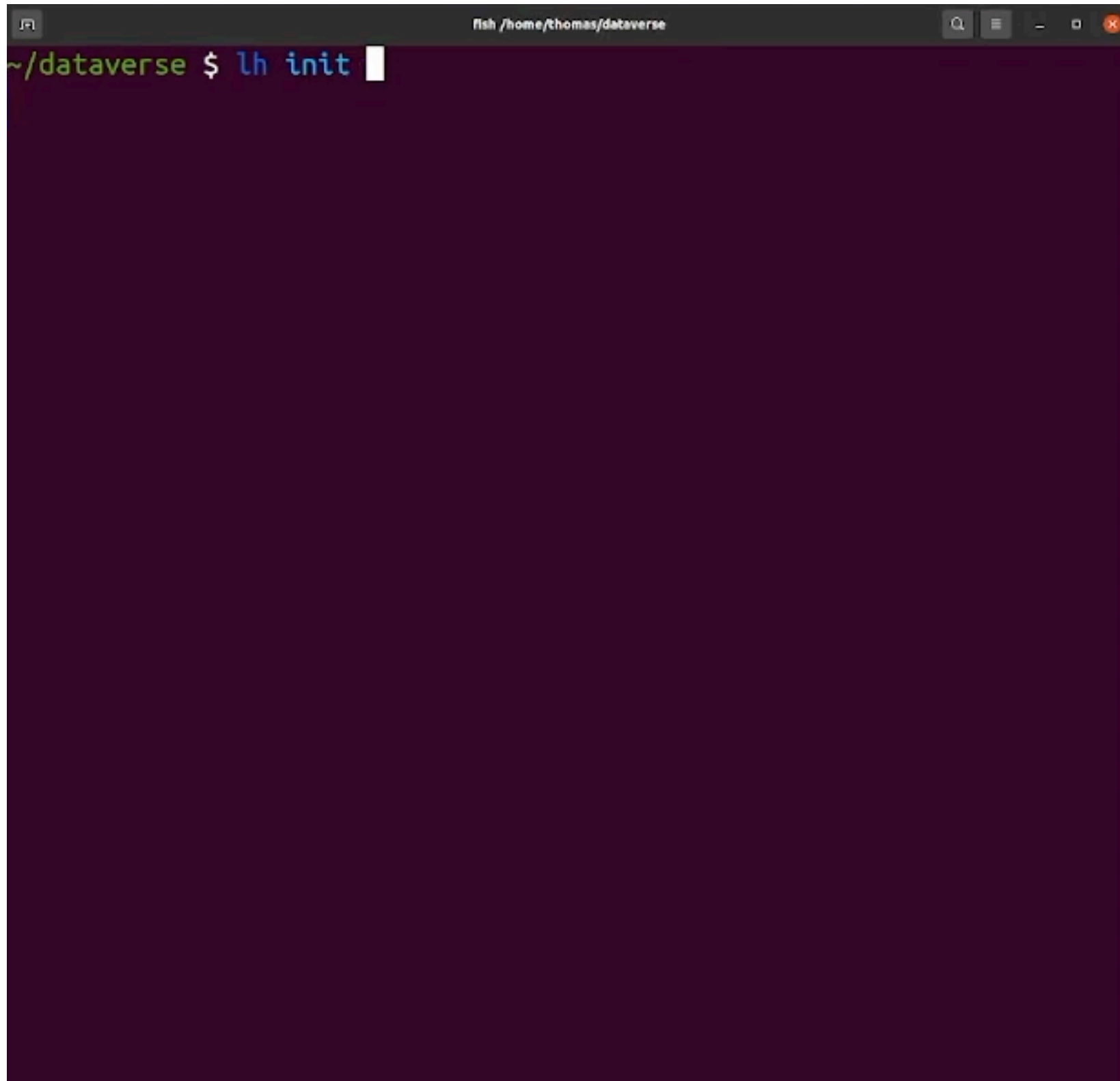
A distributed **metadata layer** that enables compliant data sharing.

A dataset's security policy can be **updated in near real-time**.



LOHPI®

Lohpi workflow

A terminal window with a dark purple background. The title bar at the top reads 'fish /home/thomas/dataverse'. The prompt is '~ /dataverse \$' and the command 'lh init' is entered, followed by a white cursor. The rest of the terminal is empty.

```
fish /home/thomas/dataverse  
~/dataverse $ lh init
```

- ▶ Integrates with existing authentication services

Lohpi workflow

```
fish /home/thomas/dataverse
~/dataverse $ lh init
Initialized Lohpi directory
~/dataverse $ lh login
Login link 🌐:
https://login.microsoftonline.com/c628abe3-0fe1-4381-bcc0-4721d4387973/oauth2/v2.0/authorize?&state=12345&client_id=14ea51f7-5500-4a45-bae1-3f308cfade90&response_mode=query&response_type=code&scope=https%3A%2F%2Fgraph.microsoft.com%2FUser.read%20&redirect_uri=http%3A%2F%2Flocalhost%3A5000%2FgetAToken&prompt=select_account&code_challenge=1qaz2wsx3edc4rfv5tgb6yhn1234567890qwertyuiop&code_challenge_method=plain
~/dataverse $ lh list
{
  "Identifiers": [
    "Data.bin",
    "Fish_Dataset.zip",
    "noMNIST.zip"
  ]
}
```

- ▶ Integrates with existing authentication services
- ▶ Public list of available datasets (discovery)

Lohpi workflow

```
fish /home/thomas/dataverse
~/dataverse $ lh checkout notMNIST.zip
Unauthorized: 401 Unauthorized
~/dataverse $
```

- ▶ Integrates with existing authentication services
- ▶ Public list of available datasets (discovery)
- ▶ Seamless data checkouts

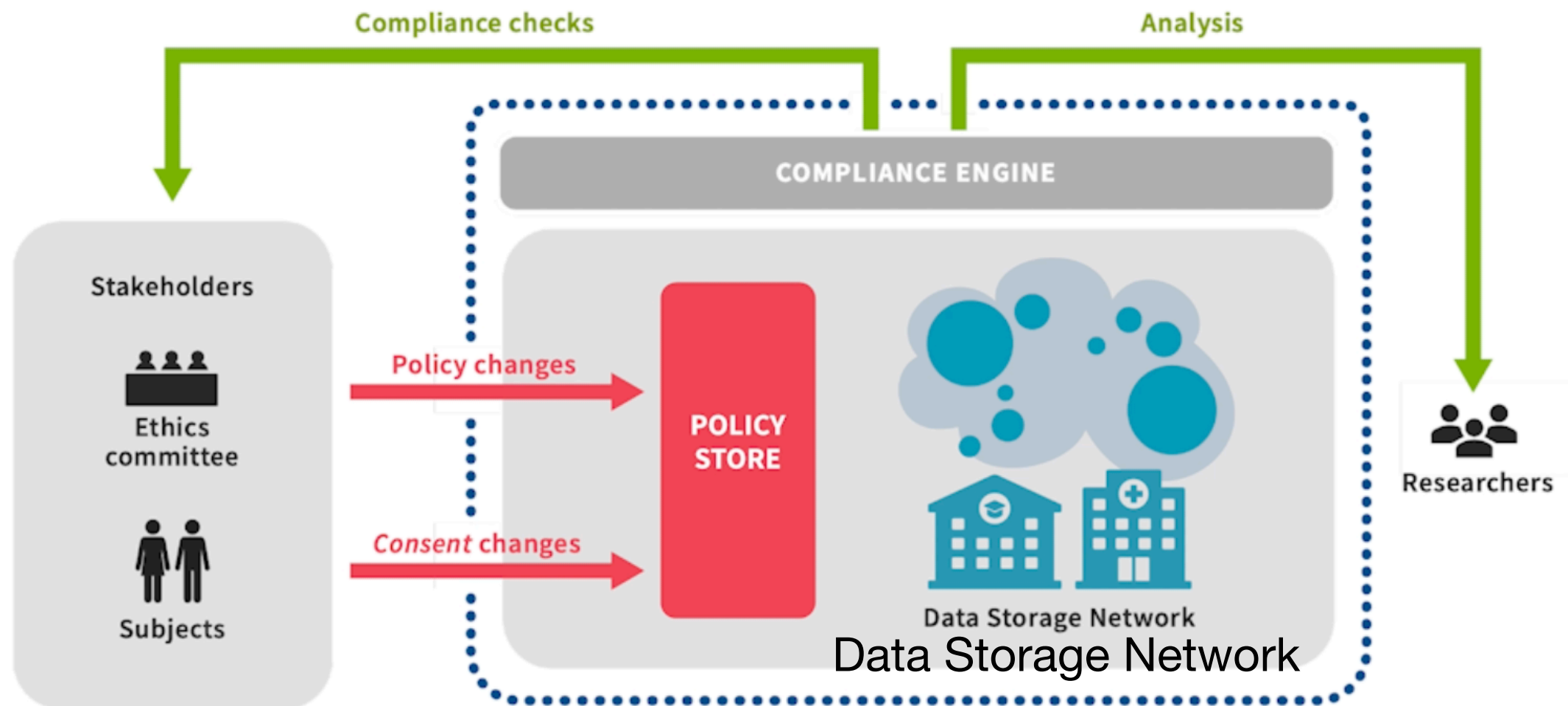
Lohpi workflow

```
fish /home/thomas/dataverse
~/dataverse $ lh checkout notMNIST.zip
Unauthorized: 401 Unauthorized

~/dataverse $ lh checkout notMNIST.zip
Downloading "notMNIST.zip"...
Success! Location: notMNIST.zip
~/dataverse $ lh comply notMNIST.zip
Dataset checkout is compliant
~/dataverse $
```

- ▶ Integrates with existing authentication services
- ▶ Public list of available datasets (discovery)
- ▶ Seamless data checkouts
- ▶ Up-to-date data security policies

Lohpi architecture



Sharma, Aakash, et al. "Up-to-the-minute Privacy Policies via gossips in Participatory Epidemiological Studies." *Frontiers in big Data* 4 (2021).

Secure dissemination with gossips



Jenkins, Kate, Ken Hopkinson, and Ken Birman. "A gossip protocol for subgroup multicast." Proceedings 21st International Conference on Distributed Computing Systems Workshops. IEEE, 2001.

Johansen, H. D., Renesse, R. V., Vigfusson, Y., & Johansen, D. (2015). Fireflies: A secure and scalable membership and gossip service. ACM Transactions on Computer Systems (TOCS), 33(2), 1-32.

Remove bottleneck

DIRECT MESSAGING

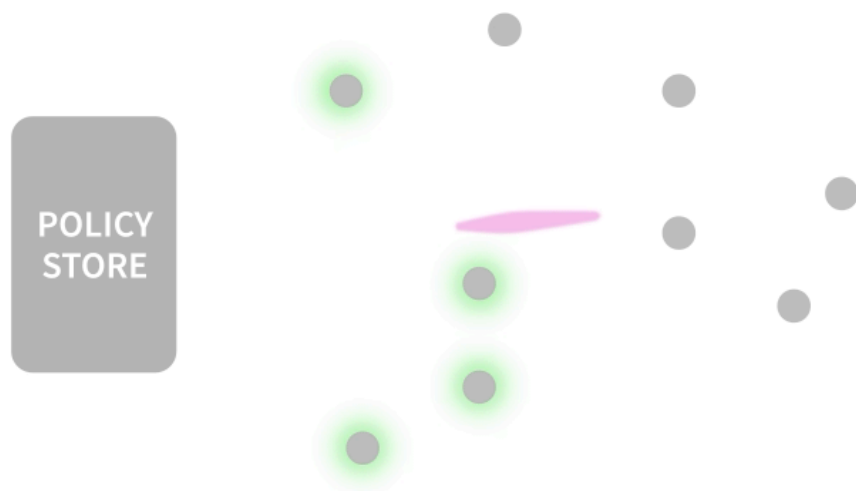


GOSSIPING

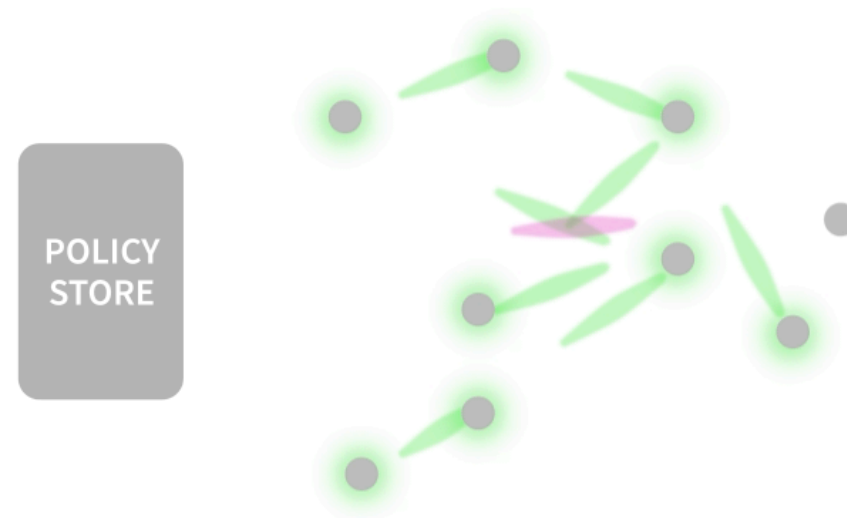


Scalability

DIRECT MESSAGING



GOSSIPING



What's in a gossip?

Info

MessageID ad528370

Signature Algorithm SHA-256 ECDSA

Signature 71 bytes: 30 45 02 20 7E 3C 48 DA B9 A5
B0 2F 5A 31 EC C4 25 6B 04 ...

Index

Index version 1.0

Policy ID#1 (0f5fbb8f, 2.23)

Policy ID#2 (c5bdaf2f, 2.5)

Policy ID#3 (bd9a29dd, 1.3)

Updates

Policy (0f5fbb8f,

{

Location code: EU-NO,

Authority: REK-NOR-8392,

Intents: research-only...

})...

Ongoing work

- Low-code policy language
- Compliance (formal proofs)
- Policy enforcement (Intel SGX, FUSE)
- Cloud-based service for clients
- Host sensitive datasets (Sports, Fisheries Crime)

Conclusion

Lohpi

- A distributed infrastructure to support compliant data sharing and analytics.
- Scalable across multiple ethics committees.
- Cloud-based or local infrastructure.

References

1. Bongartz, H., Rübsamen, N., Raupach-Rosin, H., Akmatov, M. K., & Mikolajczyk, R. T. (2017). Why do people participate in health-related studies?. *International journal of public health*, 62(9), 1059-1062.
2. Salerno, Jennifer, et al. "Ethics, big data and computing in epidemiology and public health." *Annals of Epidemiology* 27.5 (2017): 297-301.
3. Goodman, Kenneth W., and Eric M. Meslin. "Ethics, information technology, and public health: duties and challenges in computational epidemiology." *Public Health Informatics and Information Systems*. Springer, London, 2014. 191-209.
4. Sharma, Aakash, et al. "Privacy Perceptions and Concerns in Image-Based Dietary Assessment Systems: Questionnaire-Based Study." *JMIR Human Factors* 7.4 (2020): e19085.
5. Jenkins, Kate, Ken Hopkinson, and Ken Birman. "A gossip protocol for subgroup multicast." *Proceedings 21st International Conference on Distributed Computing Systems Workshops*. IEEE, 2001.
6. Johansen, Håvard D., et al. "Fireflies: A secure and scalable membership and gossip service." *ACM Transactions on Computer Systems (TOCS)* 33.2 (2015): 1-32.
7. Sharma, Aakash, et al. "Up-to-the-minute Privacy Policies via gossips in Participatory Epidemiological Studies." *Frontiers in big Data* 4 (2021).

LOHPI^Φ Team



*Thomas
Bye Nilsen*

*thomas.bye.nilsen
@uit.no*



*Aakash
Sharma*

*aakash.sharma
@uit.no*



*Dag
Johansen*

*dag.johansen
@uit.no*



*Håvard D.
Johansen*

*havard.johansen
@uit.no*



Uit The Arctic
University of Norway

Lab → NN

SpareBank 1
NORD-NORGE