

# Data Sharing 2032

Giorgia Bincoletto, Dorien Huijser, Enrico Glerean  
CIMeC Virtual Event  
2022-06-23

# Program

— — —

- 10:00-10:35: Scientific research using health data: Data Protection regime vs. Open Data - Giorgia Bincoletto
- 10:35-10:40: Break
- 10:40-11:15: Data sharing in practice: getting started - Dorien Huijser
- 11:15-11:20: Break
- 11:20-11:55: Open issues in data sharing - Enrico Glerean
- 11:55-12:00: Break
- 12:00-12:30: Final discussion

# Scientific research using health data: Data Protection regime vs. Open Data

**Giorgia Bincoletto**

Post-doc researcher

Faculty of Law, LawTech Group

University of Trento - Italy

— — —

Slides of the talk “Scientific research using health data: Data Protection regime vs. Open Data” can be found at:

<https://doi.org/10.5281/zenodo.6702892>

# Data sharing in practice: getting started

**Dorien Huijser**

Research data manager  
Utrecht University Library  
The Netherlands

# Step-by-step approach

— — —

1

## PLAN

- Ethics approval
- Data Management Plan
- Information letter & consent form

2

## EXECUTE

- Secure tooling
- Best practices
- Agreements

3

## ARCHIVE & SHARE

- FAIR
- Anonymous vs. personal data sharing

**Consult with your data steward / privacy officer for recommended tools & procedures**

# PLAN: Data management + privacy + ethics

— — —



## What to collect?

If personal data:

Minimise

Explicit purpose (e.g., research question or exploratory)



## What to share?

- If personal data: Anonymisation plan or informed consent for data sharing
- Type & size
- Underlying a paper vs. entire project



## Restrictions?

- Purposes - validation, new questions, etc.
- Audience - team, all researchers, public?
- Time (embargo)
- Agreements?



## Responsibility + costs

- Taking care of data requests
- Incidental findings
- Costs

**Include in ethics application, data management plan & participant information letter**

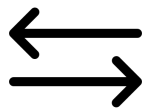
# PLAN: Brain data: personal or anonymous?



Header info in  
DICOM files



Facial details



Link through  
accompanying  
data



“Fingerprint” (both  
functional patterns and  
anatomical images)

Large number of data points → consider personal data, thus requiring:

- Legal basis, e.g., informed consent
- Protection:
  - Pseudonymisation
  - Minimisation: as little as needed, remove when not necessary anymore
  - Separate research data from contact information
  - Secure tooling, access control, etc.
- Enable data subjects' rights



# PLAN: Informed consent

Do: be transparent about which data you collect and which data you want to share, why, and who can access the data now and in the future.

Do: be explicit on what will happen to participants' data + risks involved

Do: provide contact information for questions and complaints

“To the best of our knowledge, data will not contain information that can directly identify you using reasonable means. [...] But: with additional data linked to your name (e.g., medical scans), your information could be associated with you”

Don't: promise anonymity if you are not certain you can achieve that.

Don't: promise that you will publish data publicly, if you cannot guarantee the participants' privacy

**Consult with your data steward / privacy officer for correct formulation in your institute**

Example from: [https://open-brain-consent.readthedocs.io/en/latest/gdpr/ultimate\\_gdpr.html](https://open-brain-consent.readthedocs.io/en/latest/gdpr/ultimate_gdpr.html)  
Bannier et al., 2021. The Open Brain Consent: Informing research participants and obtaining consent to share brain imaging data. <https://doi.org/10.1002/hbm.25351>

# EXECUTE

— — —

## Secure tooling

For storage, collection, analysis, sharing, etc.

## Best practices

Brain Imaging Data Standard: <https://bids.neuroimaging.io/>

Documentation (version control, README, etc.)

Minimise, separate, abstract, hide, inform, control, enforce, demonstrate

## Agreements

*Ask your institute's legal team!*

E.g., Consortium, data processing (e.g., external tools), data transfer (sharing)

# ARCHIVE & SHARE: Basic steps

— — —

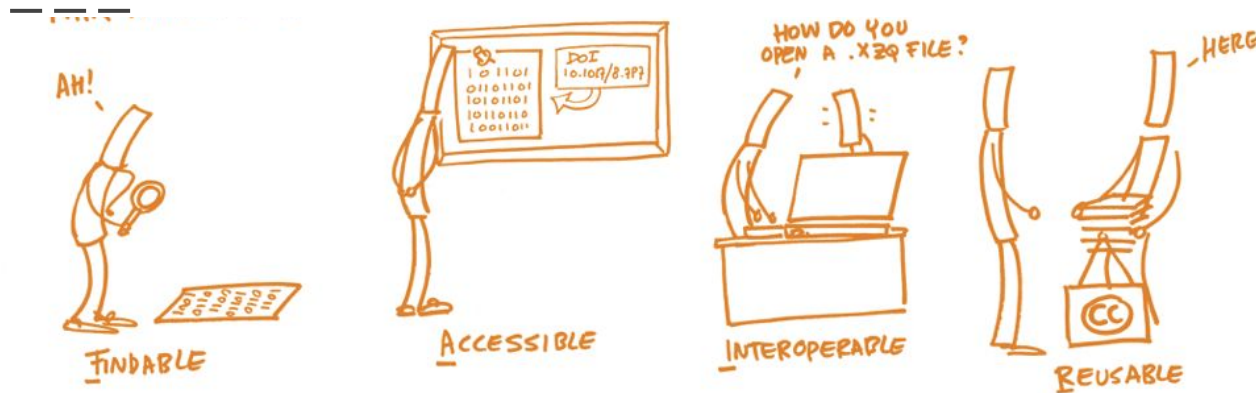
1. Pseudonymise (possibly anonymise) the data as early as you can
2. Publish at least the metadata & documentation in a data repository (FAIR)
3. Depending on your scenario, share data (characteristics):
  - a. Data are anonymous → Share!
  - b. Legal basis to share personal data → Share with restrictions
  - c. No legal basis to share personal data → Use alternatives

# ARCHIVE & SHARE: 1. Pseudonymise data

— — —

Data type	Pseudonymisation	Tools
Contact info (name, address, phone, email, etc.)	Remove If still needed: separate + control access	-
Other directly identifying info (e.g., photos, id numbers)	Remove, deidentify, or control access	
Demographic info (age, gender, education, etc.)	Generalise (e.g., birth date → age → age group), remove, separate	Amnesia: <a href="https://amnesia.openaire.eu/">https://amnesia.openaire.eu/</a> ARX: <a href="https://arx.deidentifier.org/">https://arx.deidentifier.org/</a>
DICOM images	<a href="#">Remove header info</a> Deface images Separate raw images	<a href="https://www.researchgate.net/post/Best_free_tool_for_DICOM_data_anonymization">https://www.researchgate.net/post/Best_free_tool_for_DICOM_data_anonymization</a> <a href="https://github.com/PeerHerholz/BIDSonym">https://github.com/PeerHerholz/BIDSonym</a>
Subject IDs	Replace, randomize or remove	
Free text (e.g., interviews, open survey questions)	Check, replace identifying info or remove	<a href="https://nlp.stanford.edu/software/CRF-NER.html">https://nlp.stanford.edu/software/CRF-NER.html</a>

# ARCHIVE & SHARE: 2. FAIR dataset: Publish metadata



Always publish **metadata + documentation** in a data repository, even if you cannot share the data itself.

Choose repository (gives DOI + basic metadata)

Specify access conditions

Use [BIDS](#) + open [formats](#)

**Documentation**

E.g., SOPs, analysis pipelines / code, codebooks, README, preregistrations, information letters, etc.

Interlink “data” + publication + code

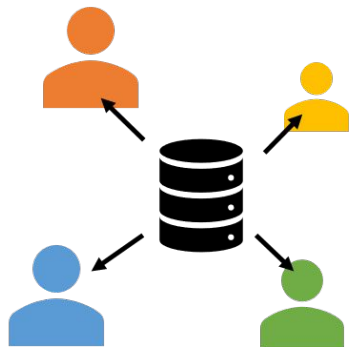
FAIR principles: <https://www.go-fair.org/fair-principles/>

How to FAIR: <https://howtofair.dk/>

Image: Foster ([https://www.openaire.eu/images/Guides/FAIRdataprinciples\\_foster.png](https://www.openaire.eu/images/Guides/FAIRdataprinciples_foster.png))

# ARCHIVE & SHARE: 3. Determine your scenario

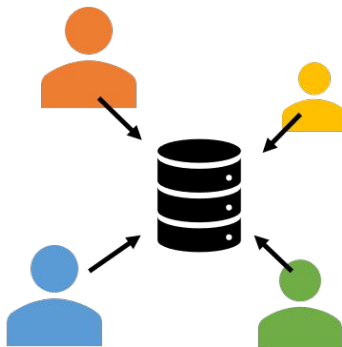
Data are fully anonymous



Publish data alongside metadata in a data repository

+ all other options →

Legal basis to share data

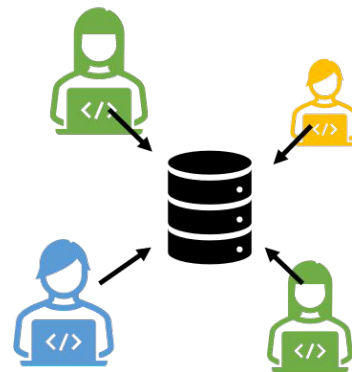


Share with restrictions, e.g.:

- Data repository with restricted access option
- Add requestor to infrastructure (no data transfer)

+ all other options →

No legal basis to share data



- Share derived data in a data repository
- (Federated) code-to-data
- Synthetic data
- Confidential computing

# Repositories to deposit (neuro)data



GIN



# Sharing with access restrictions

— — —

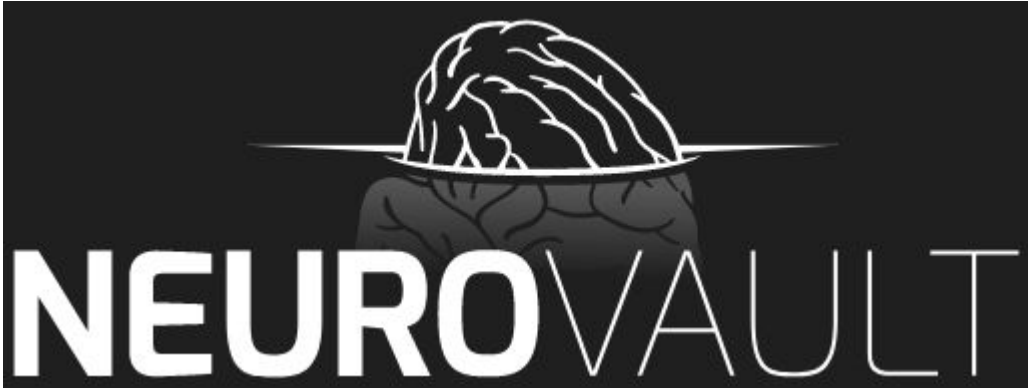
Personal data = Protect!

Access restrictions:

- Data sharing/transfer agreement
  - Restrict use of the data (e.g., don't reidentify, no third-party sharing, etc.)
  - Mandatory when sharing outside EEA, possibly also within EEA
  - Always discuss with legal department + authorized person has to sign!
- Possible other restrictions:
  - Research proposal
  - Affiliation with research institute
  - Agree to data use agreement (custom license)
  - Embargo
  - Etc.



# Derived data sharing

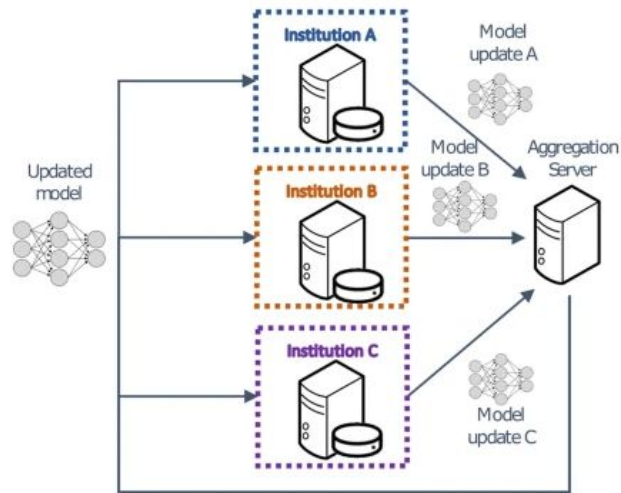


TemplateFlow

brainlife

# (Federated) Code to data

— — —  
Send code to data → run code locally → return results only



Federated/distributed analysis:  
Model trained partly locally, each run updating the aggregated model.

Examples:

- [DataShield](#)
- [ENIGMA](#) consortium
- [COINSTAC](#)
- [PySyft](#)

# Confidential computing, synthetic data



## Synthetic data

Same data structure, possibly same statistical properties

- Many R & Python packages
- MRI data: see <https://brainpower.readthedocs.io/en/latest/simulations.html>



## Confidential computing

Secure *analysis*

Encrypted containers on secure clusters

# TAKE-HOME MESSAGE

— — —

1. Plan data sharing well in advance
2. Communicate data sharing intentions to participants and in your planning documents
3. Always publish metadata and documentation
4. If you have personal data: pseudonymise and protect the data
5. You can always share derived and anonymous data
6. Use alternatives such as synthetic data & code-to-data solutions if needed

# Open issues in data sharing

**Enrico Glerean**

Staff scientist and data agent  
Aalto University, School of Science  
Finland

# 15 seconds about me

## Staff scientist + data agents at Aalto University

- daily helping researchers with issues related to **computational workflows** with sensitive data, data **minimisation**, preparing legal+ethical forms
- [trainings for the whole Aalto](#) on handling personal data in research and hands-on data **pseudo|anonymization**
- also research (**neuroimaging, medical imaging**)

# Why do we need to share our data?

# Transparency is at the core of Research integrity



Aalto University  
School of Science



# Transparency is at the core of Research integrity

The four main lines for Research Integrity according to  
**The European Code of Conduct for Research Integrity**

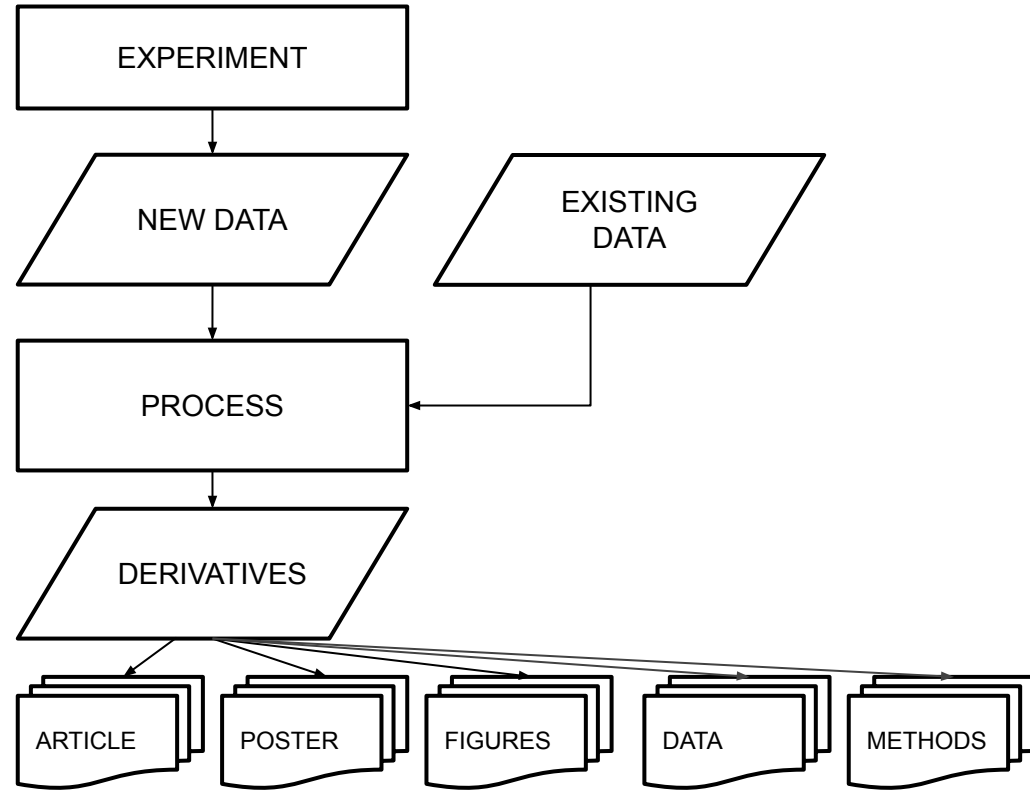
1. **Reliability** - concerns the quality and **reproducibility** of research
2. **Honesty** - concerns the **transparency** and objectivity of research
3. **Respect** - for the human, cultural and ecological environment of research
4. **Accountability** - concerns the implications of publishing the research

# Transparency in the research process

- Simple pipeline of research work **with personal data that cannot be fully anonymised**
- From very rich data formats to documents containing 2D colourful pictures, tables and text
- E.g. a dataset of fingerprints to train AI to decode gender

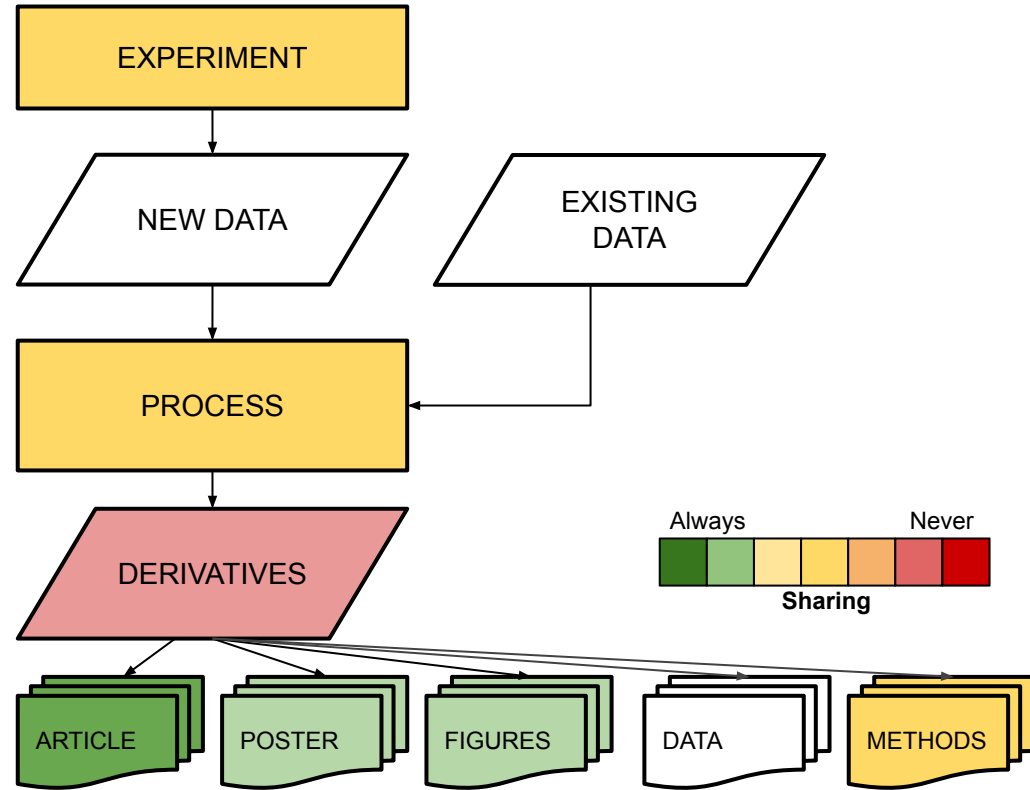


Image from <https://link.springer.com/article/10.1007/s40747-019-0099-y>



# Transparency in the research process

- Reviewers should be able to **scrutinize each “box”**
- Data should be able to be **re-used** (secondary use)
- The missing colours depend on
  - A) the country where this research is happening
  - B) the country where the data is going

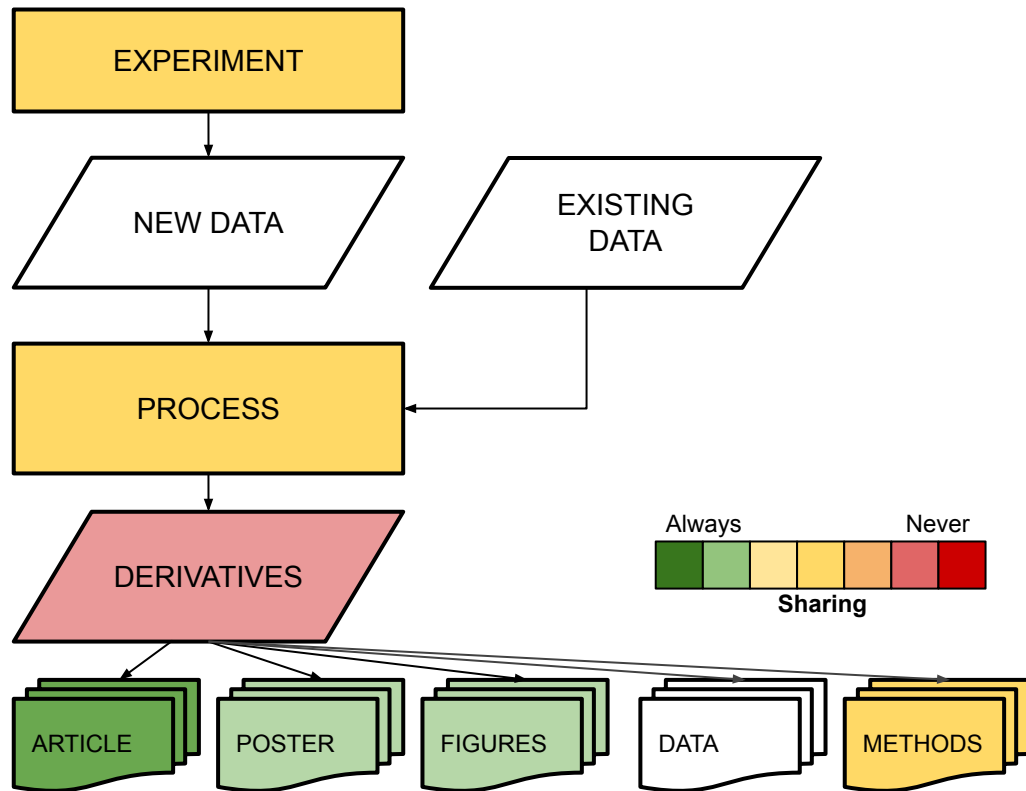


*Disclaimer: colors based on personal experience as a reviewer of neuroimaging and experimental psychology articles*

# Transparency in the research process

## The missing colours:

- Under GDPR we are able to share the fingerprints **with certain countries after a contract between parties**
- Under GDPR we are **not able to give** data access to an anonymous peer reviewer
- Under GDPR we are **not able to re-use the data for a new purpose**



*Disclaimer: colors based on personal experience as a reviewer of neuroimaging and experimental psychology articles*

# Possible consequences of a strict control of research personal data

- Researchers from GDPR countries can be **transparent only with certain selected audiences**
- Researchers from GDPR countries might be in **disadvantage** (e.g. less citations if data cannot be shared)
- **Sensitive datasets are collected in countries where these laws do not apply** (e.g. <https://arxiv.org/abs/1807.10609>)
- **Sustainability of research is affected** by difficulty in reuse (<https://www.nature.com/articles/s41431-020-0596-x>)

# We need “broad consent”

The core ethical and legal issue is

**How can we ask for consent to an individual taking part to a study, so that other researchers - who we do not yet know - can reuse the subject's personal data for answering research questions that are not yet known?**

# Ongoing work on “broad consent”

- “Broad consent” concept emerging from genetics and biobank literature, where individuals are asked for a wide consent e.g. “we reuse your data for future research”
- European Data Protection Board issued a clarification  
[https://edpb.europa.eu/our-work-tools/our-documents/other-guidance/edpb-document-response-request-european-commission\\_en](https://edpb.europa.eu/our-work-tools/our-documents/other-guidance/edpb-document-response-request-european-commission_en)

*The GDPR cannot be interpreted to allow for a controller to navigate around the key principle of specifying purposes for which consent of the data subject is asked*

**What can we do in the meantime while we wait that EDPB issues new guidelines on processing personal data for scientific research purposes?**

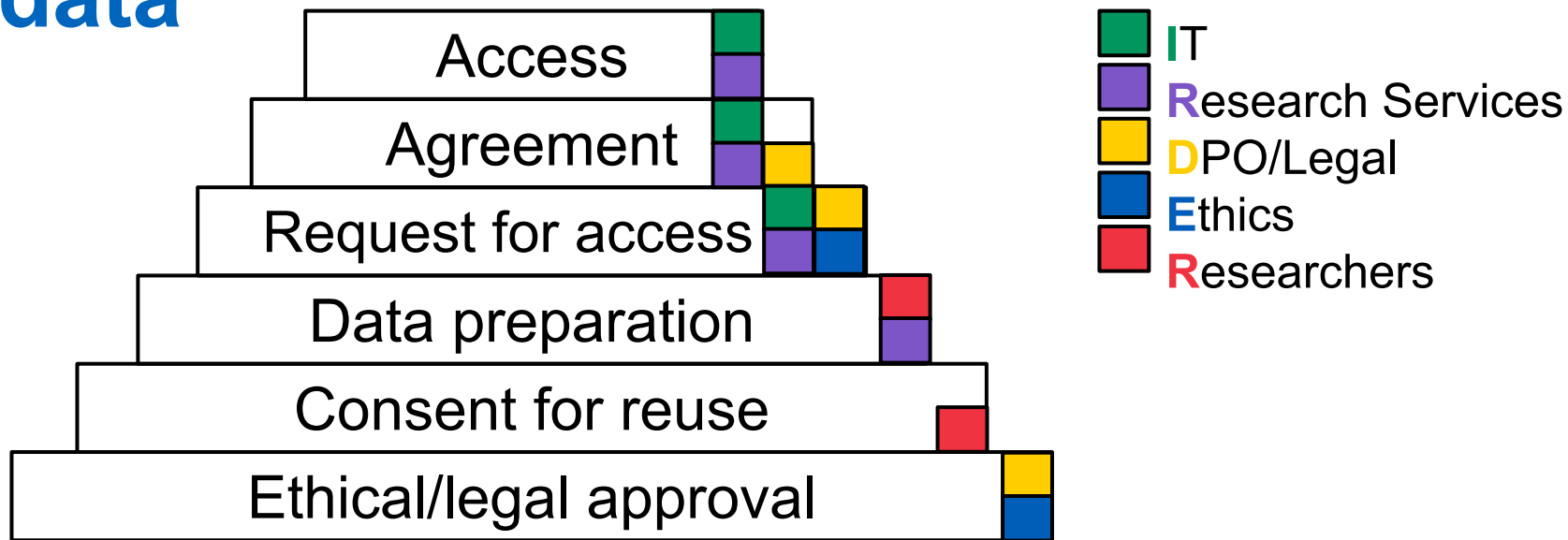


# What can we do to start sharing personal data for research purposes?

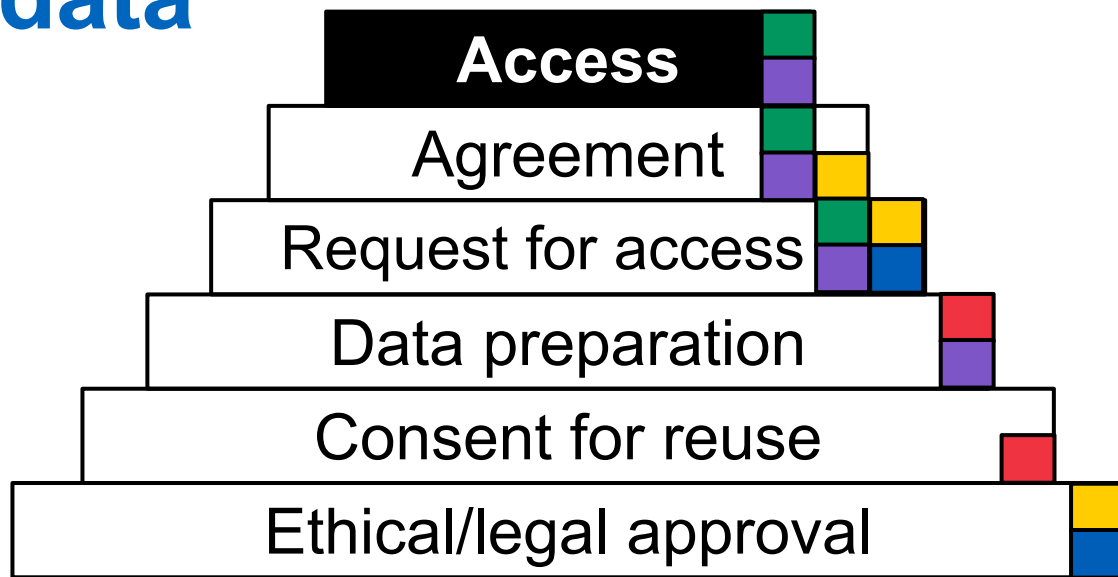
1. We make sure that the data is **findable** (as in FAIR) and **available on request**
2. We make sure that **we have a process** so that other researchers can request access to the data
3. We make sure that data are **prepared using best practices**
4. We stipulate a **contract** (DTA/DMA/DUA/DPA) with the requestor
5. We make sure that we have an **IT system for secure sharing** of such data

...where's ethics?

# The pyramid of ethically and legally sharing research data with personal data

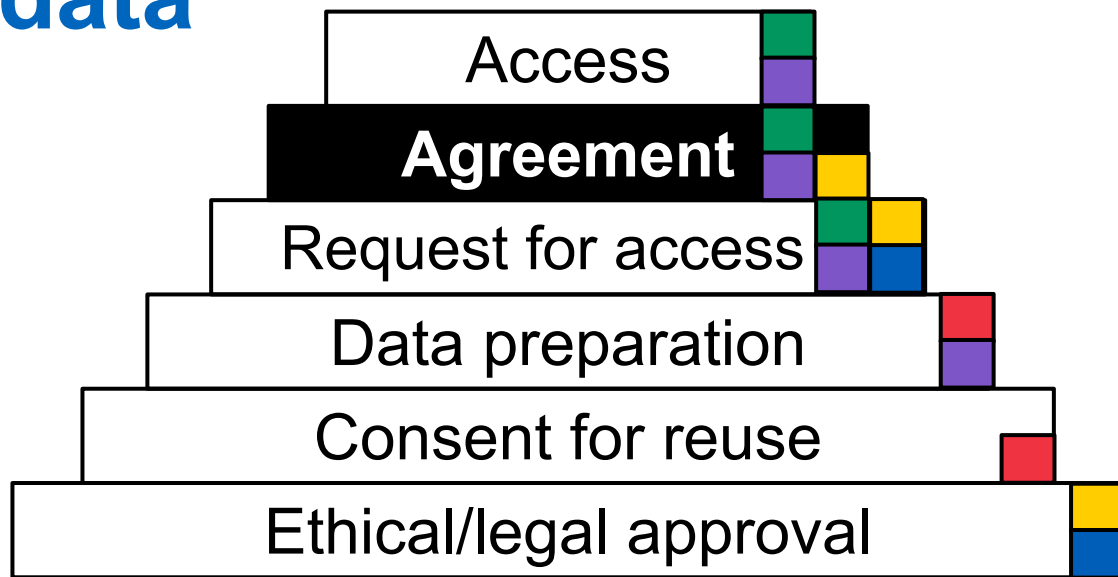


# The pyramid of ethically and legally sharing research data with personal data



**Solutions:** Access can be given so that requestor can copy data (e.g. UK biobank) or requestor uses data on local system (e.g. SD by CSC in Finland, Findata environment, TSD in Norway). Local storage can also be having a mirror of an existing dataset. Local computing can allow other models to be tested on restricted data (e.g. federated analysis).

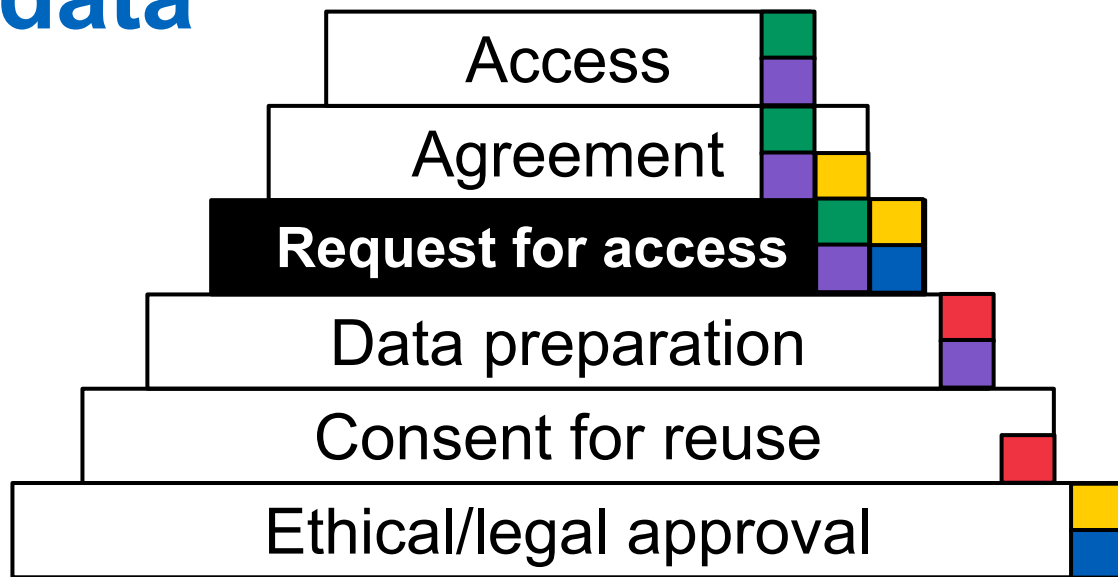
# The pyramid of ethically and legally sharing research data with personal data



**Solutions:** Data Transfer Agreement / Data Processing Agreement / Data Use Agreement are stipulated between parties.

**Open issues:** who checks that agreements are respected? What happens when something goes wrong?

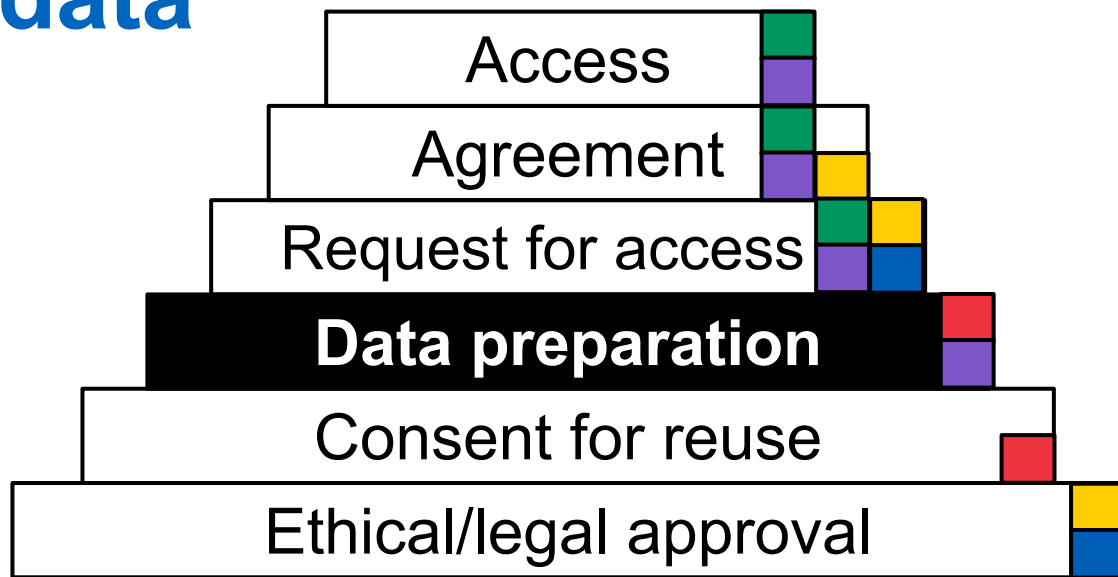
# The pyramid of ethically and legally sharing research data with personal data



**Solutions:** Example at Donders

**Open issues:** when a request for access comes, who checks it? Who approves it? Should ethics team have a saying? Who checks the requestor? Should some countries be automatically excluded? We should aim at **similar standardized processes across all organizations in EU**

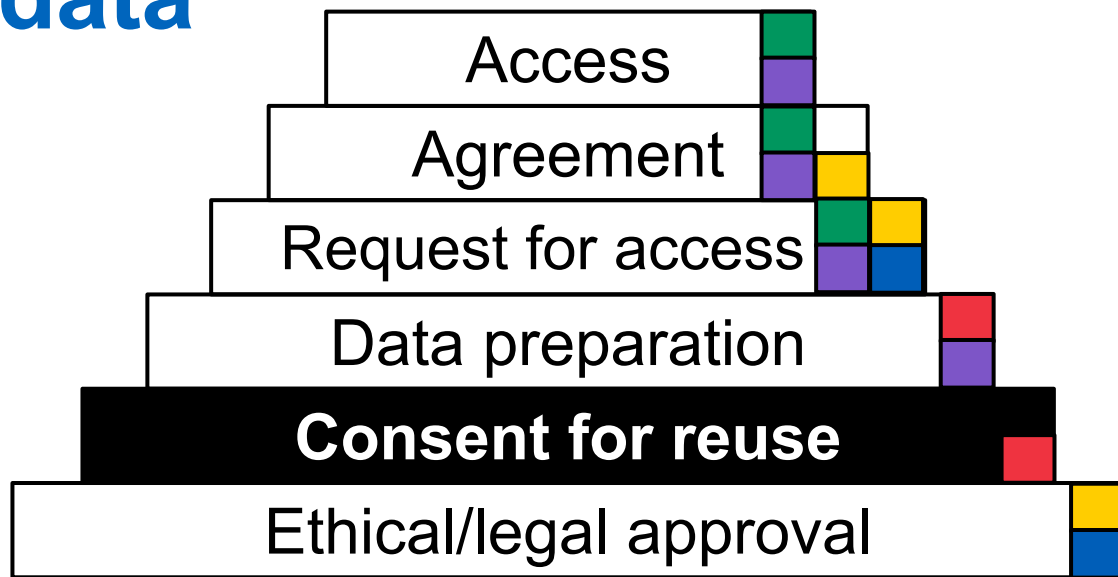
# The pyramid of ethically and legally sharing research data with personal data



**Solutions:** Researchers and research services (e.g. research software engineers) prepare the data so that it is pseudo-anonymised without direct identifiers.

**Open issues:** who checks that it is de-identified? Should ethics/legal team have a saying on the anonymization strategy?

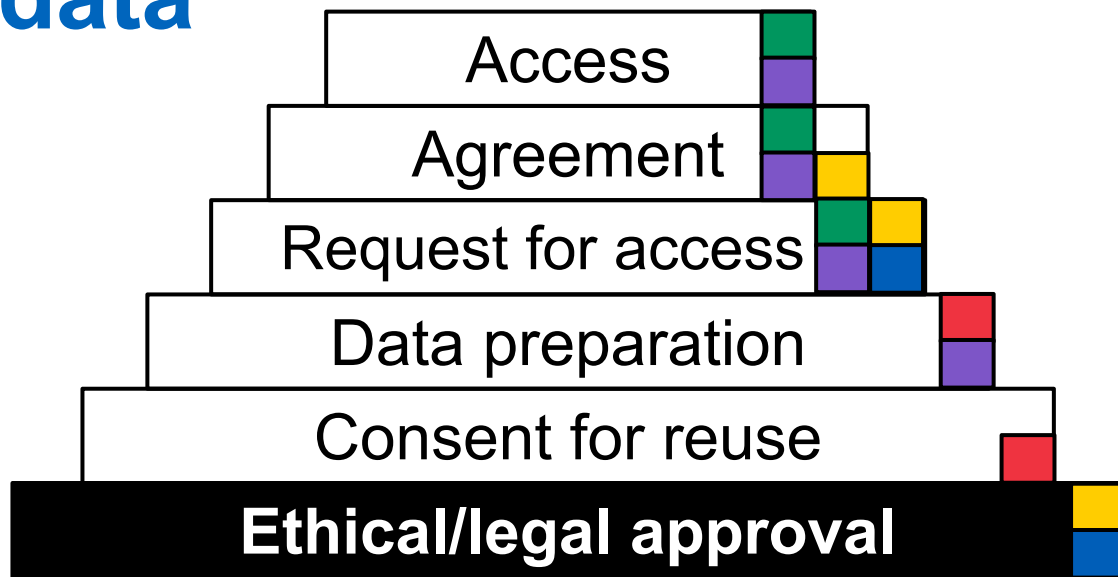
# The pyramid of ethically and legally sharing research data with personal data



**Solutions:** Potential solutions like the “open brain consent”

**Open issues:** EDPB still unclear on broad consent, how can we ethically and lawfully ask for consent for reuse/sharing to study participants? Limit the scope? Making sure that the participants understand the (small) risks and accepts them?

# The pyramid of ethically and legally sharing research data with personal data



**Solutions:** Ethics review process to take into account how re-use/sharing will happen

**Open issues:** similar as “consent for reuse”



**Yes! Let's data share!  
but...who pays the bill?**

***You can't have your cake and eat it***

# Sharing while protecting data subjects comes with lots of extra work... and work costs money

- **"Share your data, it's important!"**
  - EU policies/EOSC/ -> Ministries of education/national grant organization -> Universities and research institutions boards -> professors/P.I. -> doctoral researchers and postdocs
- **Personal data sharing just adds on top of all other tasks of our junior researchers**
  - It is a new task with no hours allocated for it (nor money)
  - It is not officially recognized as part of career advancement
  - There is no incentive for them beyond the internal "be good, do no evil, maybe you'll get citations"

# Sharing while protecting data subjects comes with lots of extra work... and work costs money

- Very often **universities own the copyright of the data, they should take the workload**, but with economic uncertainties this cost is not prioritized.
- Even when budgets might be allocated by governments/national grant agencies for the sole purpose of data opening/sharing, **data appraisal** must come into play: **which data is worth sharing?**

# Possible solutions

- Sharing one dataset (with personal data) is better than no sharing:  
**universities should pick one and start working on it**
- Universities should **invest in research support services** for both data appraisal and practical help for data sharing/opening to junior researchers
- (Multi-site) **data collection initiatives** should be prioritised with data sharing at the core. We can learn from UK biobank, and standard data collection protocols could be added to each custom small datasets people collect
- **We need more data(bio)banks** which can remove the burden from universities and researchers while ensuring highest standard of data protection (but well, who pays for these then? :) )

# Take home messages

1. **Not sharing research data** containing personal data is the easiest solution, but it is **against the transparency principle of research integrity**
2. **Protecting the subject should not come at the expenses of the transparency and reproducibility of research**, especially since we often protect study participants from close-to-impossible scenarios
3. Researchers, legal, ethics, and services teams must work on solving this together to **inform study participants of the risks of data reuse/sharing and understanding the importance of broad consent**
4. Sharing everything costs: **Data appraisal, data collection initiatives, and data(bio)banks should be a priority of EU and national grant agencies**