Data Sharing

Speaker: Kjell Petersen & **Korbinian Bösl** (UiB)
Moderator: Nazeefa Fatima (UiO)

elixir NORWAY

15 June 2022

Federated EGA Norway - a controlled access archival service for FAIR sharing of Norwegian human genome and phenome data

Kjell Petersen
Technical Coordinator ELIXIR Norway

Learning Objectives

In this session, we will learn:

- the main principles of a controlled access archive
- how it is possible to also share sensitive data in life sciences
  - …even if data sharing is tightly regulated
- how ELIXIR Federated Human Data sharing services supports FAIR data sharing

Session Take-Away

After completing this session, you will be:

- familiar with some key challenges and solutions to sharing of sensitive data
- able to understand how to get started

Learning Activity

Optional 10 min pair session
if you find the time (tomorrow)

Group activity; study the data submission and retrieval guidelines on https://ega.elixir.no. Discuss the steps researcher A need to go through to submit data to FEGA Norway and researcher B to retrieve the same dataset.

RDMkit

Data management

Data life cycle

Your role

Your domain

Your tasks

Your tasks

Sensitive data

Is your data sensitive?

Description

In general, the term “sensitive data” is used for any data that could do harm (for example to people, organisations, countries, or ecosystems) if it would be openly available. This can for example be pers commercial information, but also information such as breeding grounds of endangered species. Any must be protected against unauthorized access. What is considered sensitive information is usually by national laws and may differ between countries. You should be cautious when you are dealing wi sensitive, or potentially sensitive, information.

Considerations

- If you deal with any information about individuals from the EU, you are bound by the General D Protection Regulation (GDPR). In GDPR, such data is called “personal data”.
- In the context of GDPR “special category data” is a subclass of “personal data” that is potentially harmful, and GDPR prescribes very strict rules for dealing with this data. Article 9 of GDPR defin special categories as data consisting of racial or ethnic origin, political opinions, religious or phil beliefs, or trade union membership, genetic data, biometric data, data concerning health or dat concerning a natural person’s sex life or sexual orientation. Confusingly, these special categories sometimes colloquially called “sensitive data”. Note that this page is concerned with the broader

The way science work - and a scientific community

- Building on the work of others
  - "standing on the shoulders of giants" - or many, many, many researchers before me
- Life Sciences is a data driven/centric science field
- Sharing of data is essential in several phases
  - Among collaborators - in the active project phase
  - To the community and public - when publishing results, findings and interpretations
- Sensitive data such as human genomic data and phenotype information
  - Can not be shared open publicly for anyone to see
  - Controlled Access is required

Controlled Access Archival Service

- Dataset existence and description made public available, but not the data itself
- Access to data requires approval by authority controlling the dataset
  - Data Access Committee
- Download and reuse of data may require signed agreements

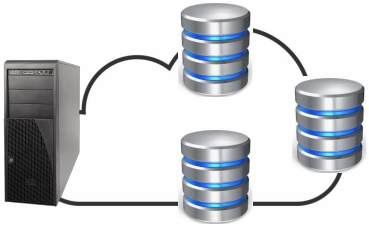
Overview of data management

FAIR data management from raw data provider to public repository

Data production

Data storage/sharing

Data compute



Data archiving

elixir Core Data Resource

Data management

TSD

SAFE

NTNU | HUNT Cloud

Federated European Genome-phenome Archive Norway

elixir NORWAY

Data for life

elixir NORWAY

Sensitive data

GDPR

§

Personvernforordningen

Personopplysningsloven

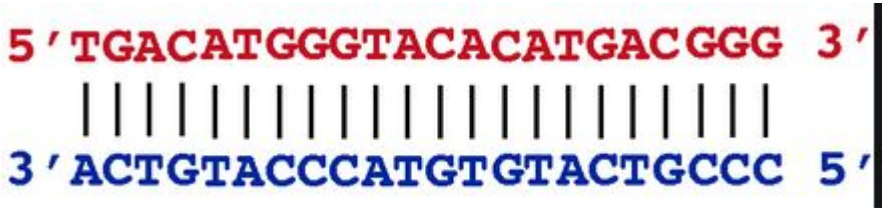
Helseforskningsloven

NORMEN

elixir

Examples of sensitive data

Genomic sequences



Disclaimer

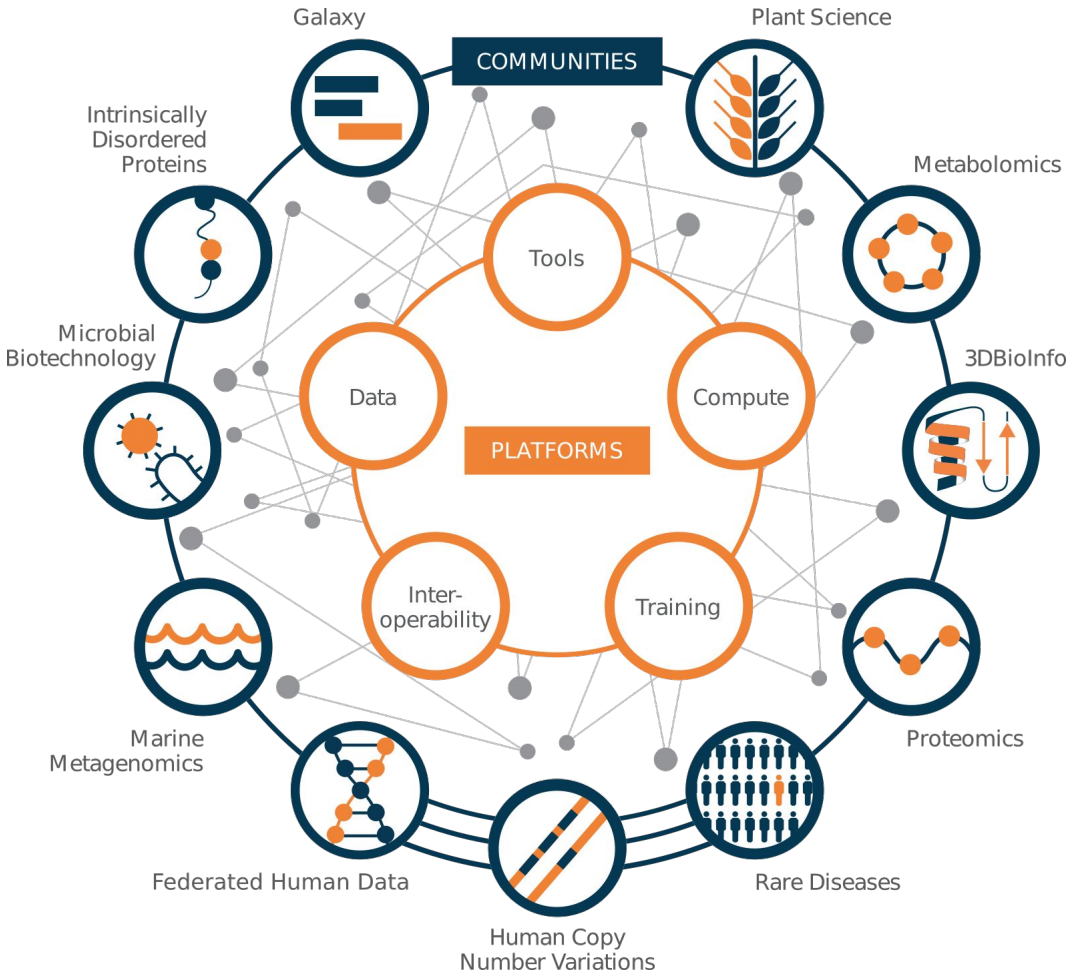
I do not have:

- any formal legal education
- authorization or official role mandated from UiB

Content overview

- Brief background on ELIXIR Europe and international context for sharing sensitive data
  - Standards and international initiatives

ELIXIR Structure



Human Data Communities

Cancer Data Focus Group

- tumour heterogeneity;
- investigation of new enzymes;
- the fight against contamination;
- tumour evolution;
- cancer progression;
- harmonization of data from electronic health records.

15 ELIXIR Nodes involved

Health Data Focus Group

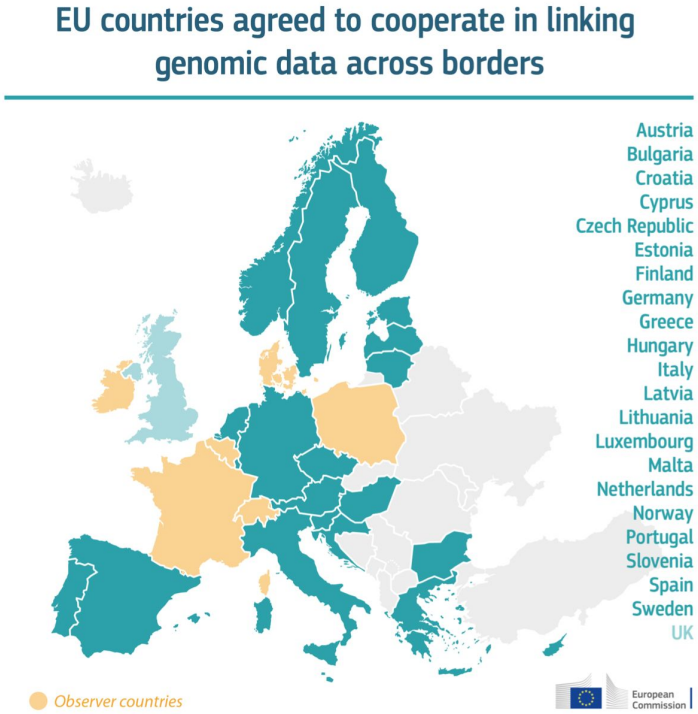
- omics
- clinical (structured and unstructured EHRs, and clinical imaging)
- lifestyle/patient questionnaire
- administrative
- real-world wearables/mobile devices data.

1+MG Declaration of cooperation - April 2018

DECLARATION OF COOPERATION

Towards access to at least 1 million sequenced genomes in the European Union by 2022

https://ec.europa.eu/digital-single-market/en/european-1-million-genomes-initiative



22 countries have now signed; 6 are observers

Aligning with international initiative

Working towards GA4GH standards, APIs and toolkits to be used throughout ELIXIR Nodes for human data discovery and access – **GA4GH into Europe**

ELIXIR coordinating data federation

Minimum recommendations for EU-wide infrastructure to access and analyse genomic data

Standards

Application Programming Interfaces

Secure, federated clouds

Tools, services & workflows

Legal and regulatory frameworks

Training

Capacity building programme

Necessary minimal infrastructure component	In development[a]	Implemented at scale[a]
Genomics data and clinical information standards geared towards specific disease communities	Yes	No
Common application programming interfaces to enable remote data discovery and access	Yes	Yes
Computational resources, including secure, federated cloud computing environments that offer secure access across national boundaries to raw data and interoperable results	Yes	Yes
Regulatory frameworks that enable access to and the processing of genomic data across borders, including the management of transnational user access and compliance	Yes	No
A repository of tools and services, including workflows to analyse deposited data while enabling these analysis workflows to operate on data across national borders. This will contribute towards data reproducibility and provenance, which are of high importance in both research and clinical practices	Yes	Yes
A training and capacity-building programme to develop the skills and workforce required for genomics and big data in health care as well as shift the culture towards openness and integration of research data across national boundaries	Yes	Yes

[a]'In development' and 'Implemented at scale' refer to locally defined status within ELIXIR and/or Biobanking and Biomolecular Resources Research Infrastructure.

Saunders G et al (2019)

nature REVIEWS GENETICS

Roadmap | Published: 27 August 2019

Leveraging European infrastructures to access 1 million human genomes by 2022

Gary Saunders, Michael Baudis, [...] Serena Scollen

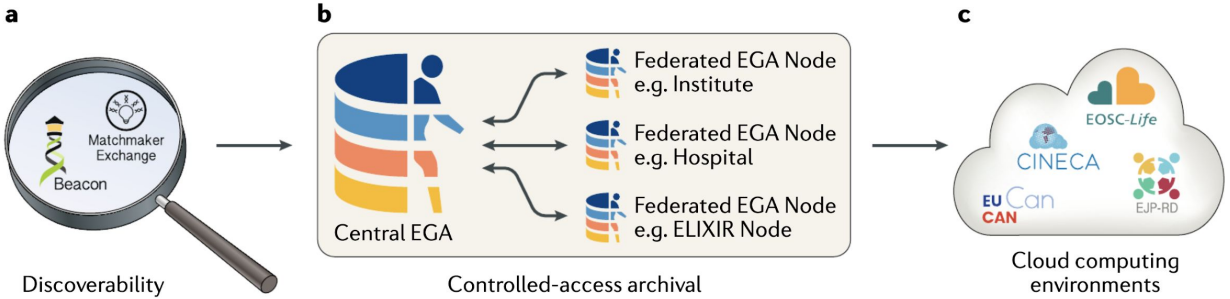
Nature Reviews Genetics (2019) | Download Citation ↓

Abstract

Human genomics is undergoing a step change from being a predominantly research-driven activity to one driven through health care as many countries in Europe now have nascent precision medicine programmes. To maximize the value of the genomic data generated, these data will need to be shared between institutions and across countries. In recognition of this challenge, 21 European countries recently signed a declaration to transnationally share data on at least 1 million human genomes by 2022. In this Roadmap, we identify the challenges of data sharing across borders and demonstrate that European research infrastructures are well-positioned to support the rapid implementation of widespread genomic data access.

Developing standards and driving use

Driving interoperability using GA4GH standards



Federated European Genome-phenome Archive Norway

TSD

SAFE

NTNU | HUNT Cloud

elixir

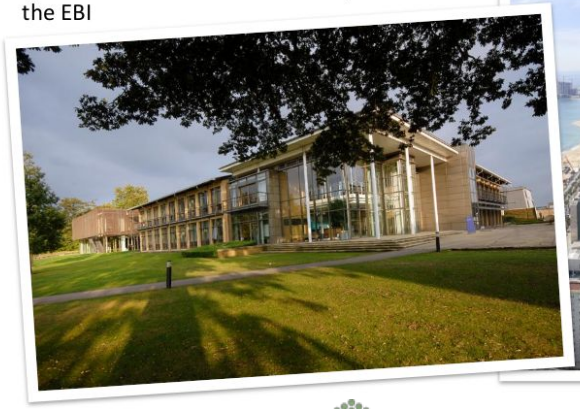
Crypt4GH: A secure method for sharing human genetic data

EGA brief history

- The EGA is a resource for permanent secure archiving and sharing of all types of potentially identifiable human genetic and phenotypic data
- Implementing the FAIR principles

How the EGA is managed

The EGA was launched in 2008 by the EBI



EMBL-EBI

In 2012, EBI and CRG started working together to establish EGA as a joint venture, which has grown in the context of ELIXIR

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EUROPEAN GENOME-PHENOME ARCHIVE

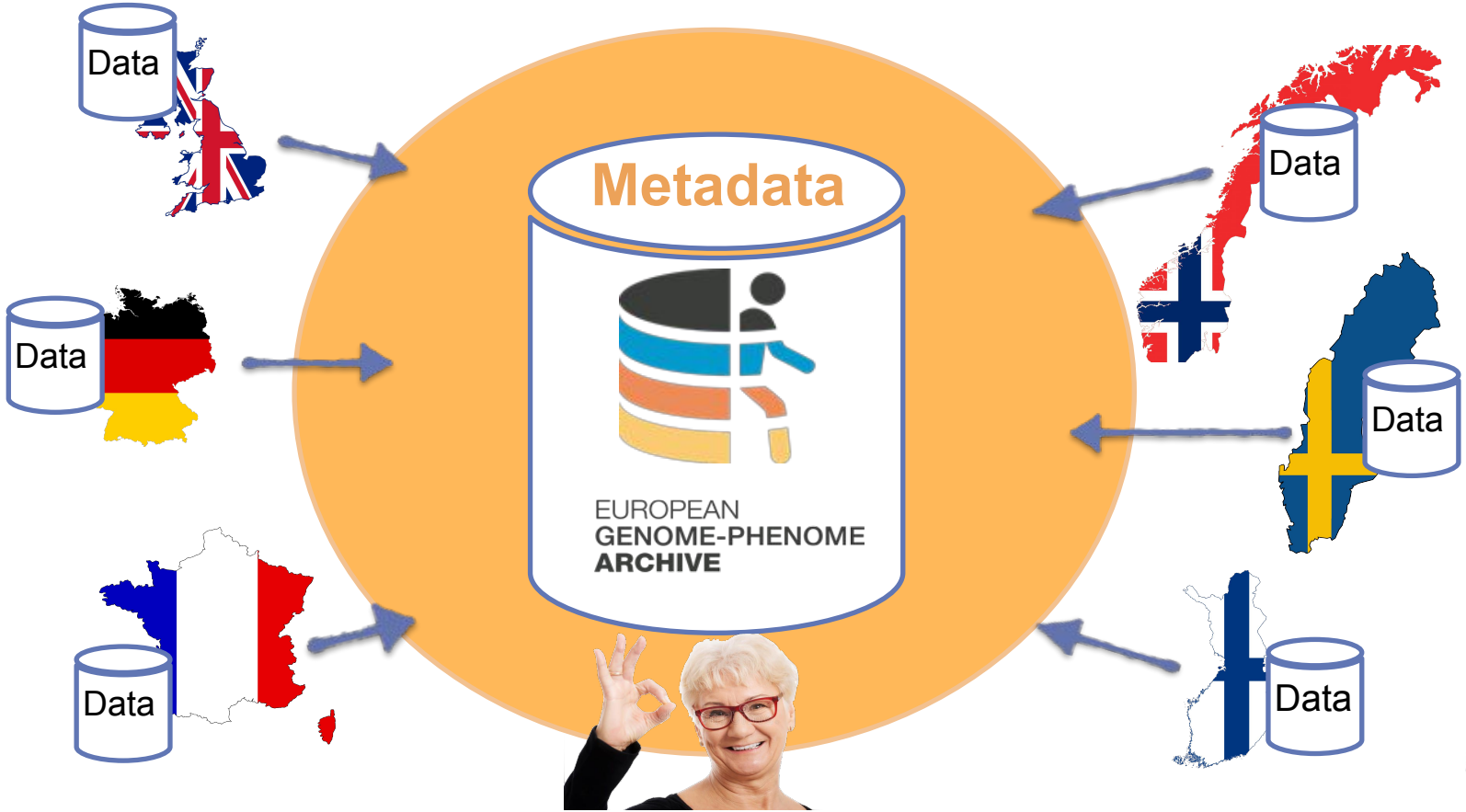
elixir NORWAY

Why - and how is this FAIR?

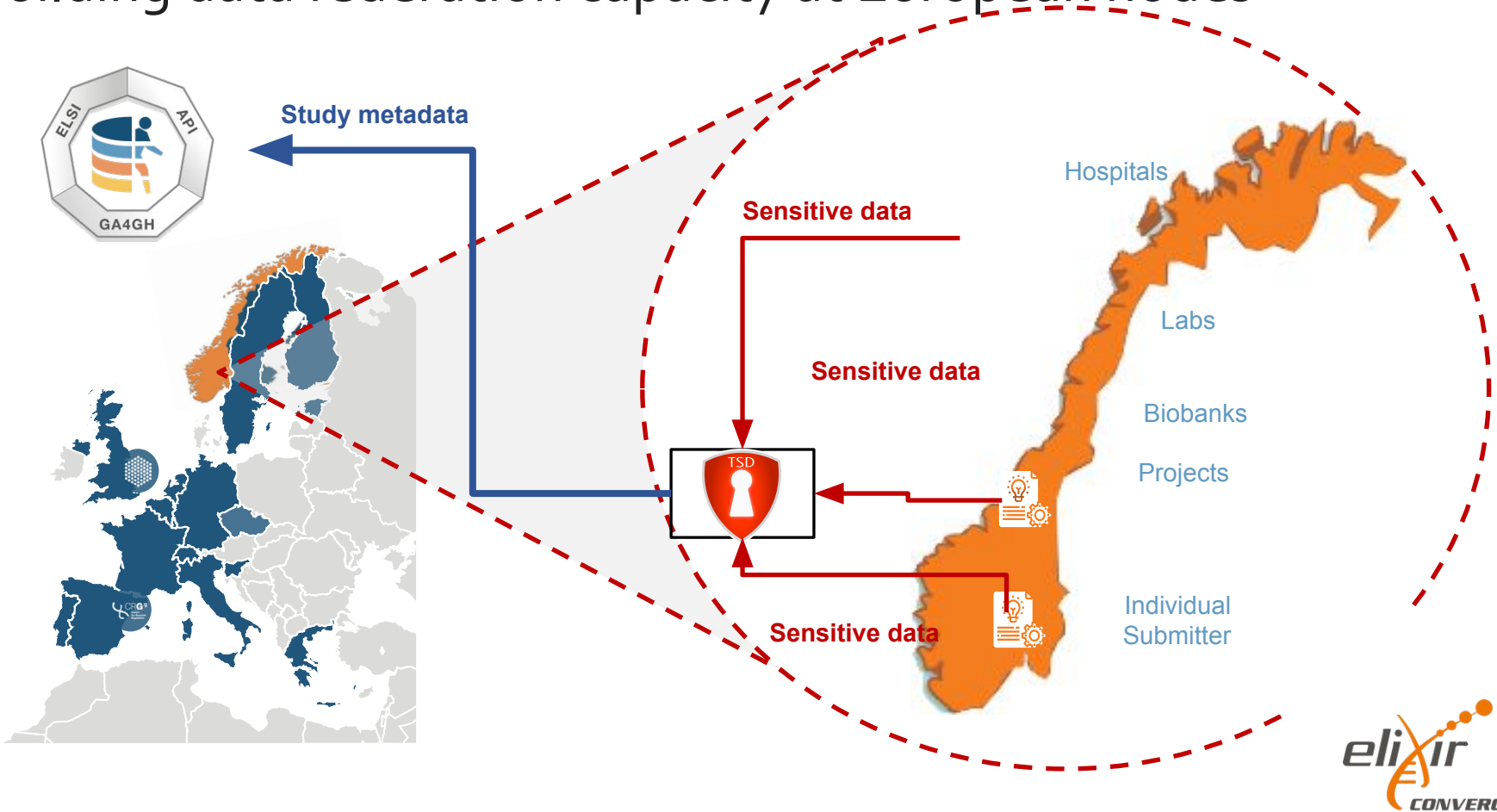
F : Findable

- Accession ID in publications to EGA / Federated EGA
- Your datasets existence are discoverable from the web portal

Federated EGA makes EGA scalable and GDPR-compliant



Building data federation capacity at European nodes



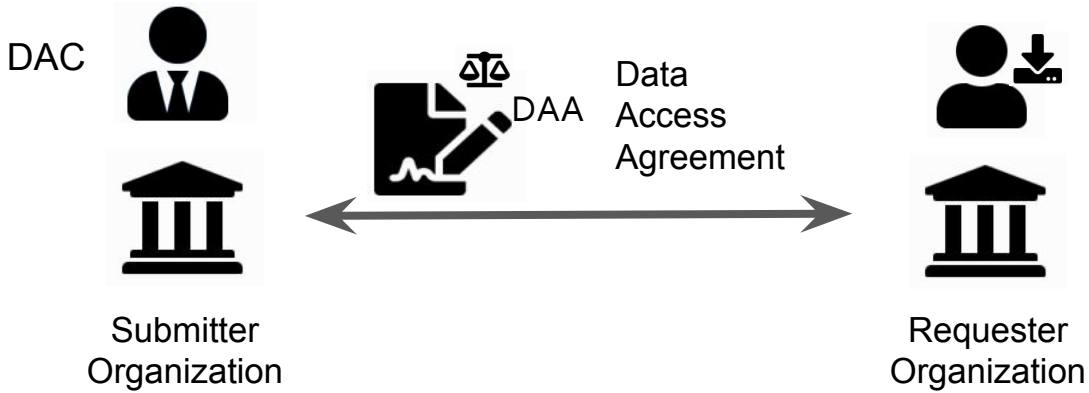
Some key GDPR concepts

Data
Service

Data Controller

Data Processor

Data Access Agreement when sharing data



https://ega.elixir.no

Federated European Genome-phenome Archive Norway

Docs and Info | Privacy Policy | Terms of Service | Contact Us | About

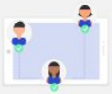
Federated EGA Norway node

LS LOGIN

The Federated EGA (European Genome-phenome Archive) is a distributed solution for sharing and exchange of human -omics data across national borders. Federated instance collects metadata of -omics data collections stored in national or regional archives and make them available for search through the main EGA portal.

Searchable

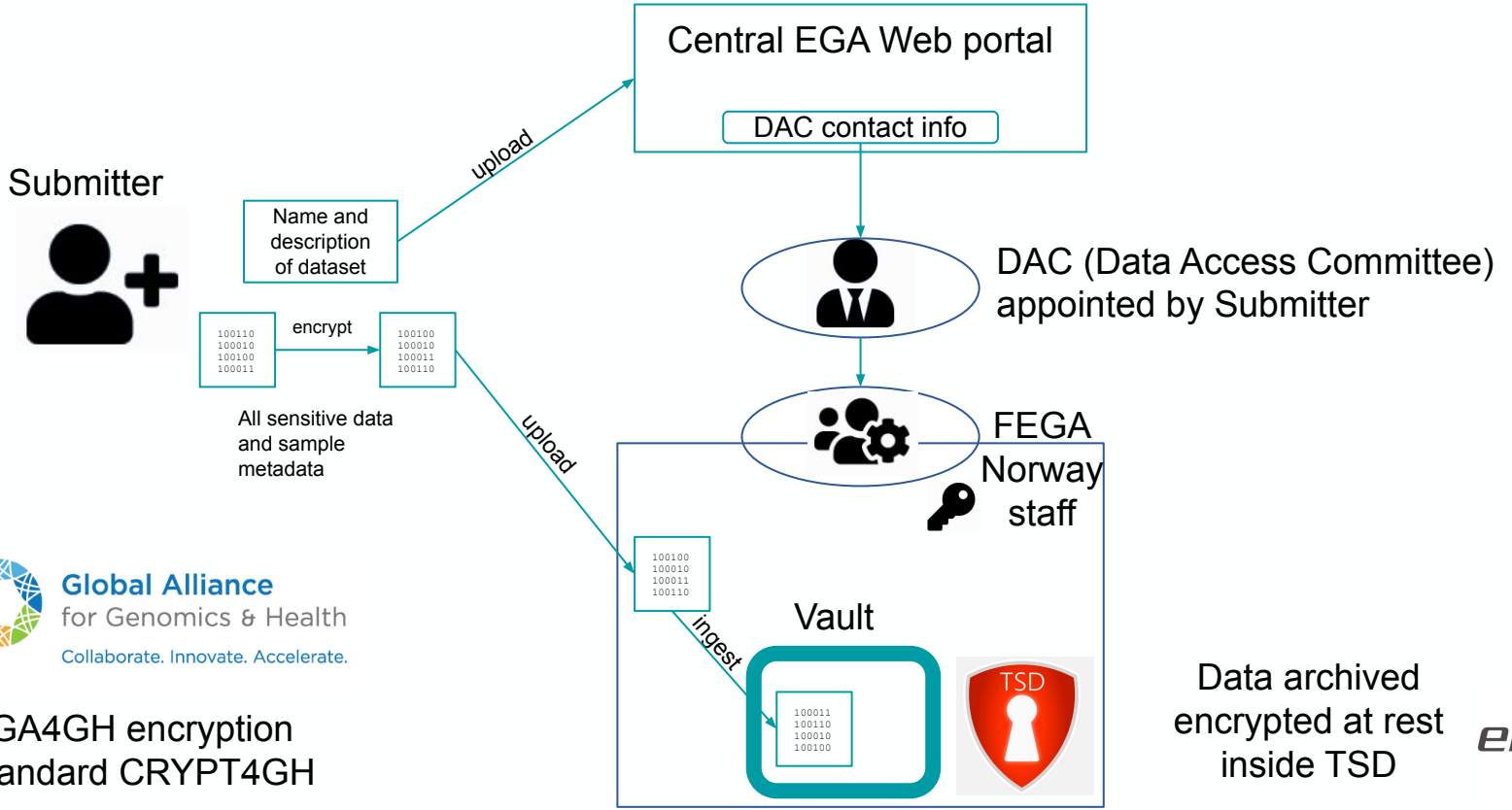
Secure



Shareable

elixir NORWAY

Data flow and data control - part I



Global Alliance for Genomics & Health
Collaborate. Innovate. Accelerate.

GA4GH encryption standard CRYPT4GH

Data archived encrypted at rest inside TSD

elixir NORWAY

https://ega-archive.org

EUROPEAN GENOME-PHENOME ARCHIVE

Search...

Tips on how to search

ABOUT | SUBMISSION | BROWSE | ACCESS | DOWNLOAD | METADATA | Helpdesk | Log in

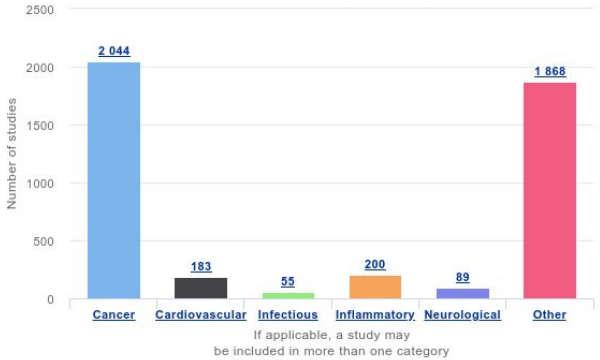
The European Genome-phenome Archive (EGA) is a service for permanent archiving and sharing of all types of personally identifiable genetic and phenotypic data resulting from biomedical research projects.

DISEASE TYPE | TECHNOLOGY | SAMPLE TYPES

What is in the EGA?

Studies in the EGA by disease

Click on a column to view category subgroups



If applicable, a study may be included in more than one category

Latest studies

The growing need for controlled data access models in clinical proteomics and metabolomics. – *2021-10-01*

Telomere maintenance by telomerase activation or alternative lengthening of telomeres (ALT) is a major determinant of poor outcome in neuroblastoma. Here, we screen for ALT in primary and Read more →

Study 1 / 5

Next Study

Published in:

BROWSE

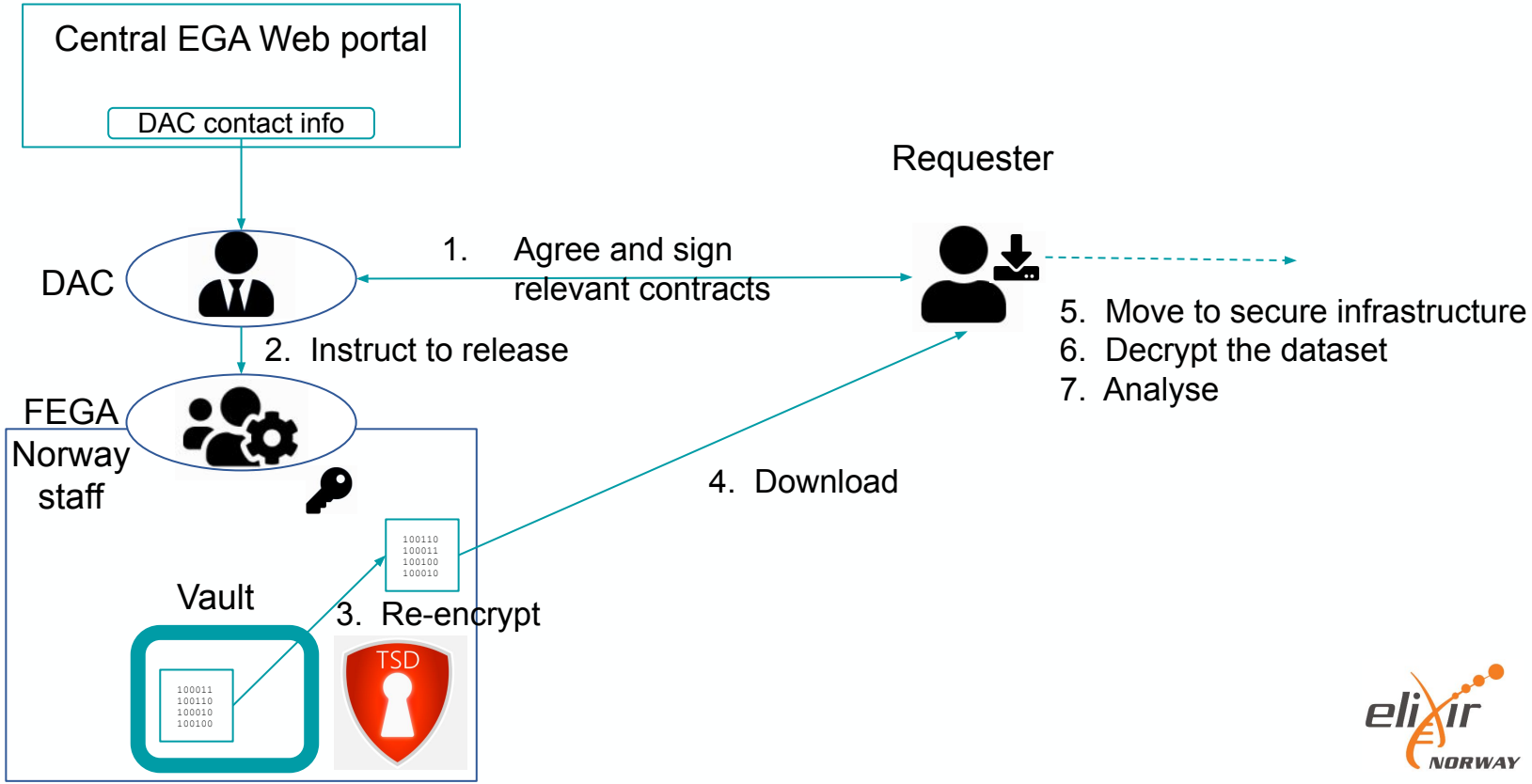
- Studies
- Datasets
- DACs

HELP

- FTP & Aspera
- Tools
- EGA Blog

elixir NORWAY

Data flow and data control - part II



GDPR required agreements and assessments

Data Processing Agreement

FEGA Norway provides a template agreement text

FEGA Norway

Submitter Organization

Data Controller

Service Provider Organization

Data Processor

Allows:

FEGA Norway to store data encrypted in TSD on behalf of the controller

To only release data to a requestor upon DAC´s instruction to do so

FEGA Norway

Submitter organization

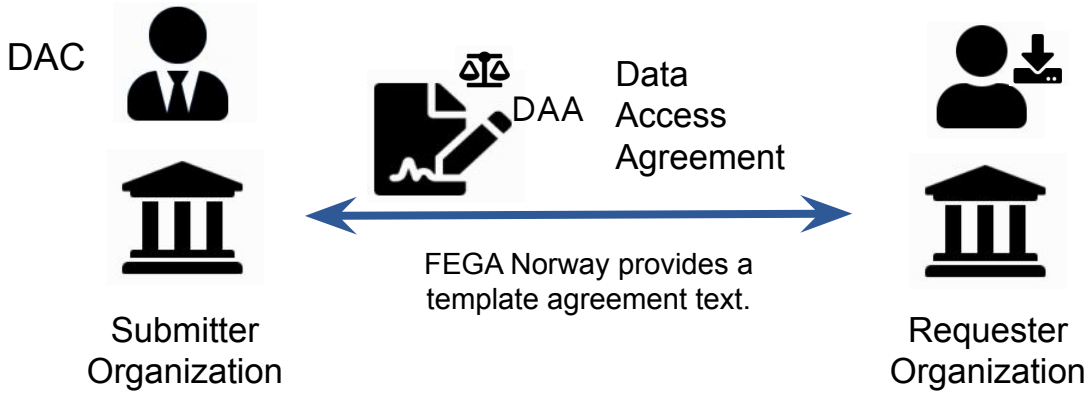
Must do ROS[1)] and DPIA[1)] according to UiO procedures required to operate the service

Must do ROS[2)] and DPIA[2)] according to their own procedures required for storing data in NFEGA

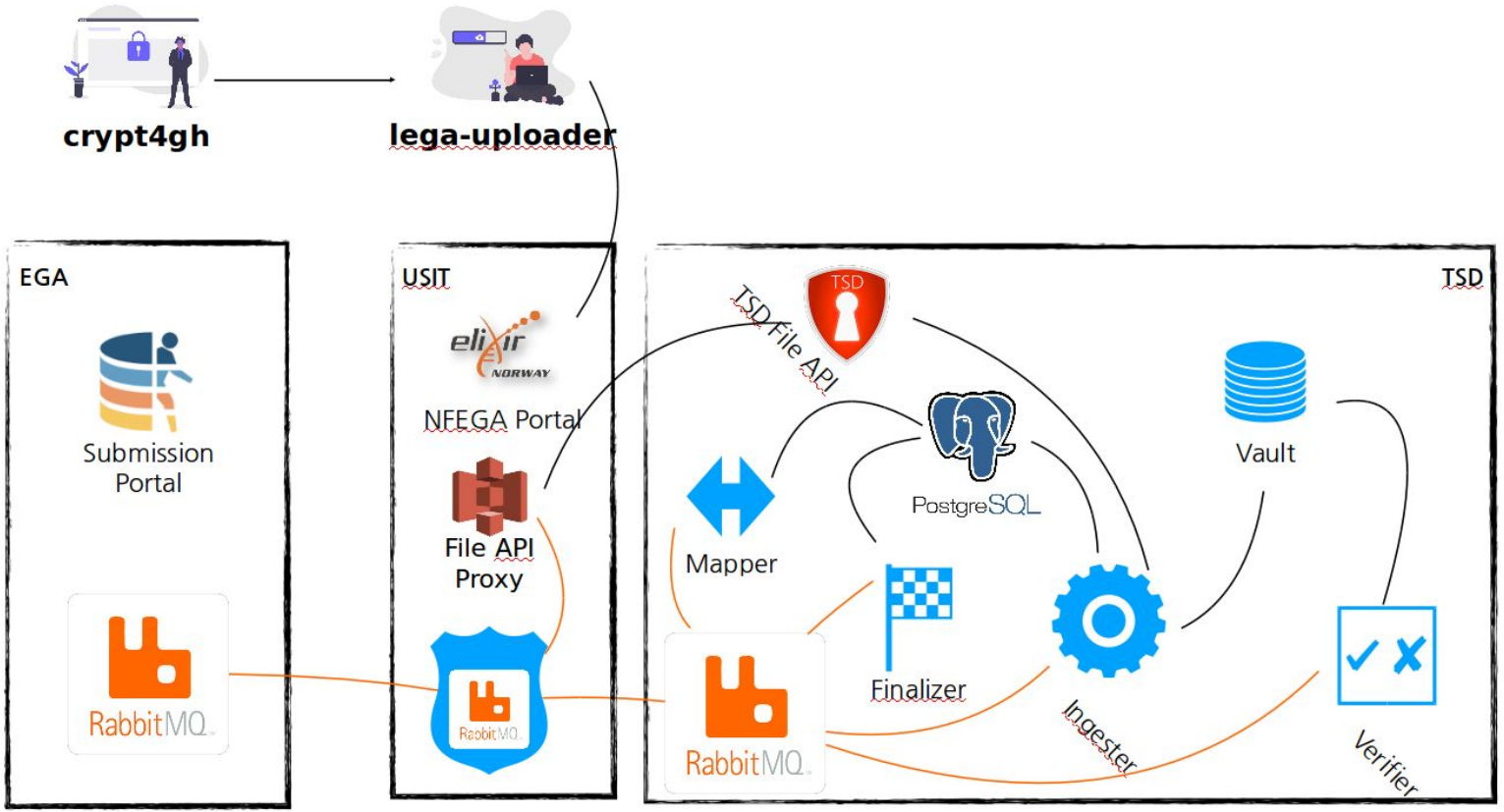
FEGA Norway provides ROS[1)] and DPIA[1)] as background for submitter organization to develop their ROS[2)] and DPIA[2)]

ROS: Risiko Og Sårbarhetsanalyse DPIA: Data Protection Impact Assessment Analysis

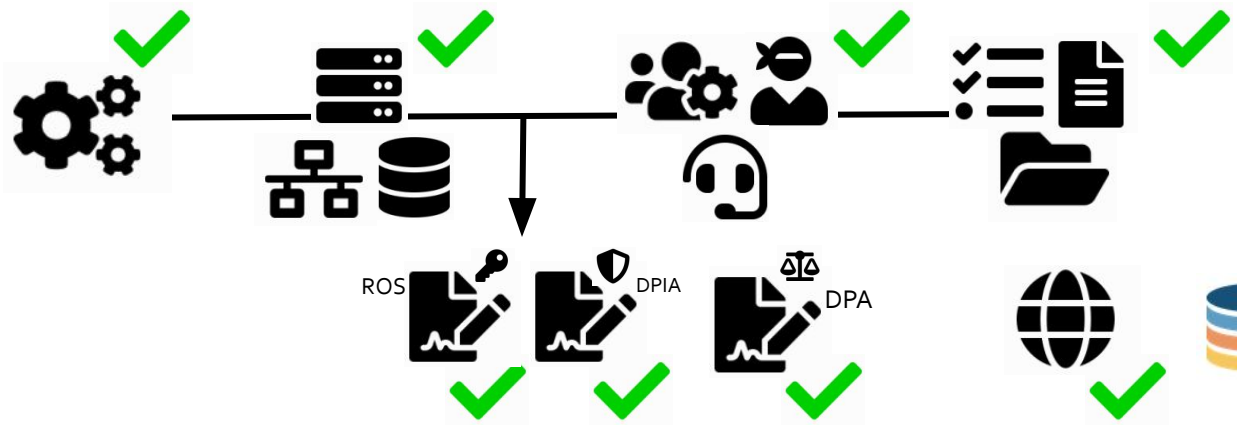
Data Access Agreement before releasing the data to Requester



FEGA Norway- Data-In



What is already in place for FEGA Norway ?



23.05.2022

elixir NORWAY

First submitter with dataset ready to be uploaded

Cancer Registry of Norway

Janus Serum Bank

The Janus Serum Bank has blood samples from 318 628 Norwegians. The biobank is reserved for cancer research, and is internationally unique regarding size and number of cancer cases.

- "Small non-coding RNA as early detector cancer biomarkers"
- Close to 2000 Samples
  - 10 different cancer types
  - 1 control group
- miRNA-seq data
- DAC appointed

elixir NORWAY

How can XUH or UiX support their researchers with FEGA data submission?

- DPA (+DPIA and ROS) of "Submitting my research dataset to FEGA Norway" is daunting for individual researchers
  - Involves competence from and communication with many OUS units/roles:
    - Data Protection Officer (Personvernombud)
    - Data Security Officer
    - Legal advisors

- DPA and related work is specific to the dataset in question, but there are large parts that are common for all datasets submitted to FEGA Norway.
  - Can be handled efficiently at right organizational level through templates and standardised agreements with appendix for dataset specific details.

https://ega.elixir.no

Protecting

The sensitive data stored at Federated EGA instance is encrypted with Crypt4GH - a new standard file container format from the Global Alliance for Genomics and Health (GA4GH), allowing genomic data to remain secure throughout their lifetime, from initial sequencing to sharing with professionals at external organizations.

Strong encryption both in transit and at rest are critical requirements for Genomics England and other genomics in healthcare initiatives.

Augusto Rendon, Crypt4GH implementation leader at Genomics England

Hosted by

UiO : Universitetet i Oslo

elixir NORWAY

Deposit data

Please follow the instructions on how to perform a submission.

Search data

Please follow the instructions on how to perform a search.

Retrieve data

Please follow the instructions on how to perform a retrieval.

Approve access

DAC, please follow the instructions of data approval.

fega-norway-support@elixir.no

Design by Papaya. Illustrations from Undraw.

elixir NORWAY

Where to work on your data before publication?

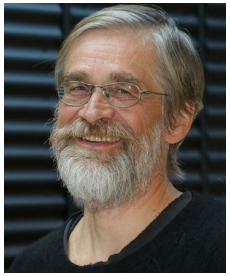
NTNU | HUNT Cloud

elixir NORWAY

Acknowledgements

Dmytro Titov
(former team member)

Eivind Hovig
Fabian L. M. Bernal
Milen Kouylekov
Ahmed Ghanem
Arash Azamifard





Gard Thomassen
Leon Charl du Toit
Abdulrahman Azab

Useful Resources

- Federated EGA Norway node, https://ega.elixir.no
- TSD (UiO) - https://www.uio.no/tsd
- SAFE (UiB) - https://www.uib.no/safe (English link top right)
- Hunt Cloud (NTNU) - https://www.ntnu.edu/mh/huntcloud
- RDMkit Sharing step: https://rdmkit.elixir-europe.org/sharing

Licensing

In this part, we will learn about data sharing with respect to:

- code versioning
- code deposition
- licensing
- data availability

Instructor: Korbinian Bösl

Learning Activity

Data Sharing Quiz Link

- Data life cycle +
- Your role +
- Your domain +
- Your problem -
  - Compliance monitoring
  - Data analysis
  - Data management plan
  - Data organisation
  - Data protection
  - Data publication
  - Data quality
  - Data storage
  - Data transfer
  - Identifiers
  - Licensing
  - Documentation and metadata
  - Sensitive data
- All tools and resources
- Tool assembly +

RDMkit

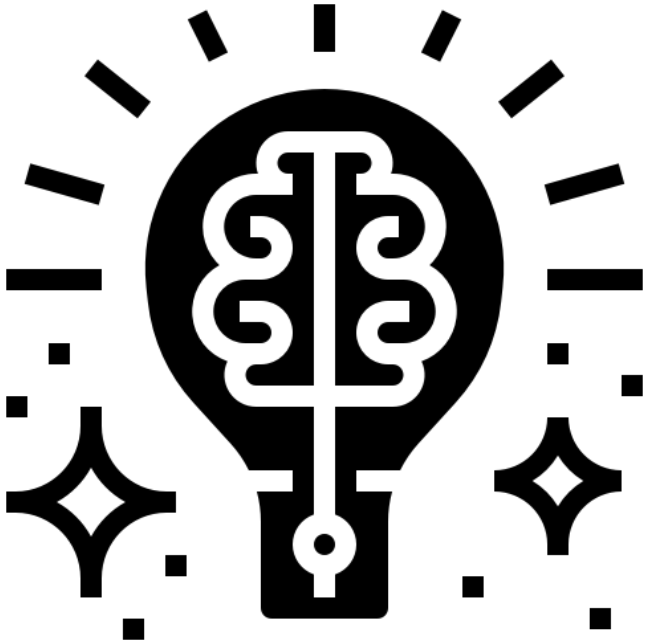
Link to RDMkit:
https://rdmkit.elixir-europe.org/

Intellectual property rights (IPRs)

Patent: Protects novel, non-obvious, inventions

Copyright: creative products: software, writing, figures, photos, some datasets, this presentation

Trademark: Protects a name/brand



IPR often regulated in work contract

Why should I licence my research outputs?

Why should I licence my research outputs?

Legal security for users (Accessibility)

Why should I licence my research outputs?

Legal security for users (Accessibility)

Increase of willingness to reuse outputs (Reusability)

Why should I licence my research outputs?

Legal security for users (Accessibility)

Increase of willingness to reuse outputs (Reusability)

Allows deposition/mirroring in 2nd databases (Findability)

Concepts in open licenses

Waive all your interests that may exist in your work

Copy left:

Concepts in open licenses

Waive all your interests that may exist in your work

Copy left:

Credit for the original creation

Concepts in open licenses

Waive all your interests that may exist in your work

Copy left:

Credit for the original creation

License new creations under identical terms

Concepts in open licenses

Waive all your interests that may exist in your work

Copy left:

Credit for the original creation

License new creations under identical terms

Non-commercial

Concepts in open licenses

Waive all your interests that may exist in your work

Copy left:

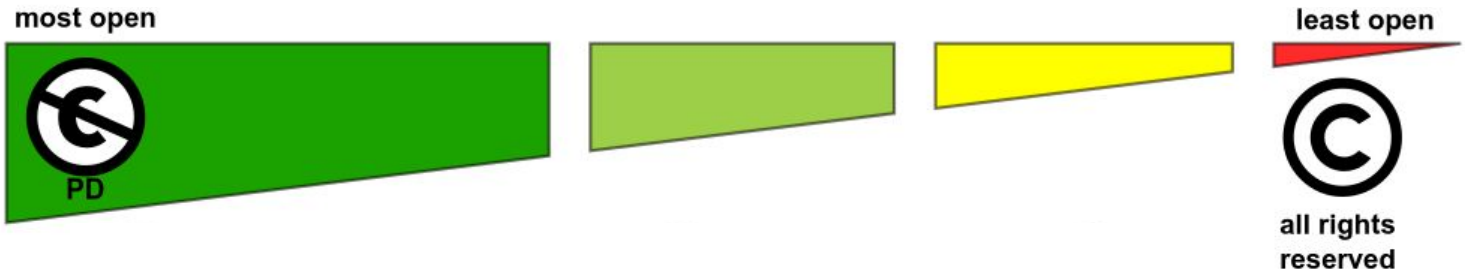
Credit for the original creation

License new creations under identical terms

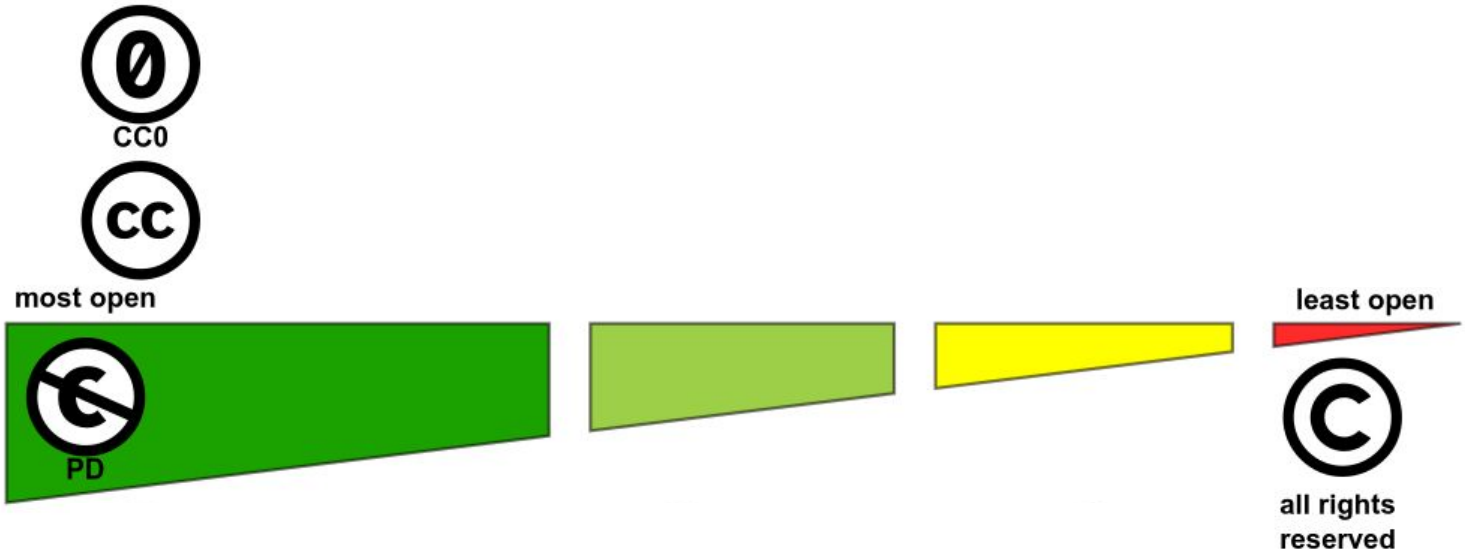
Non-commercial

Cannot be shared with others in adapted form

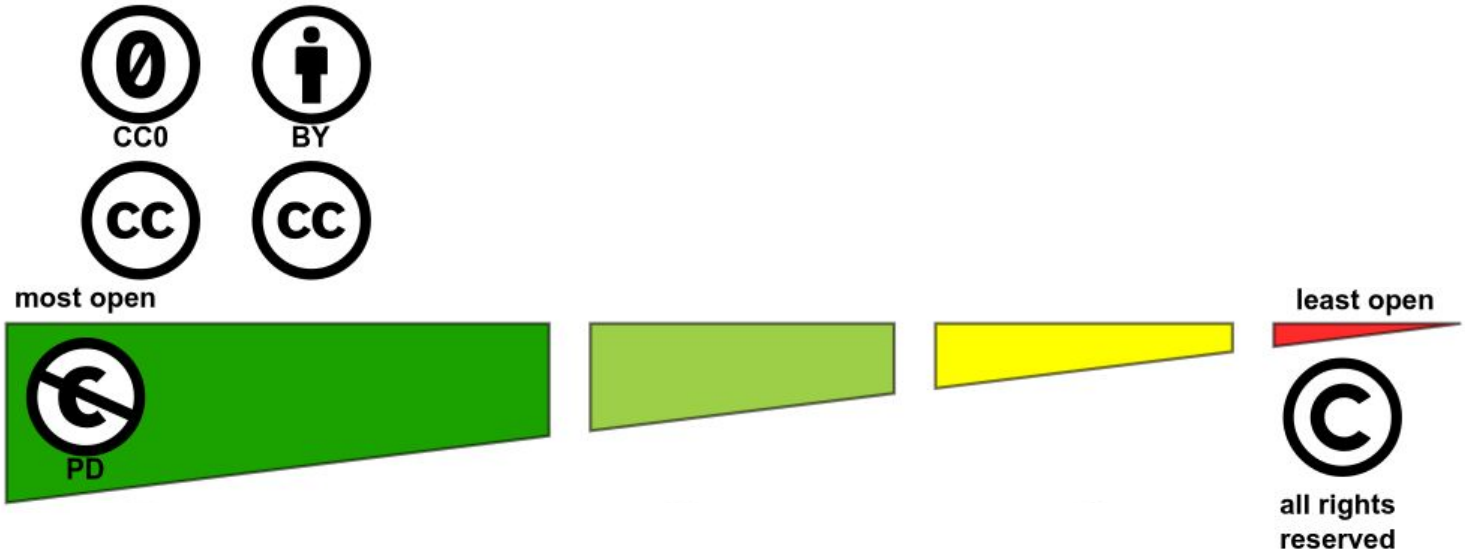
creative commons



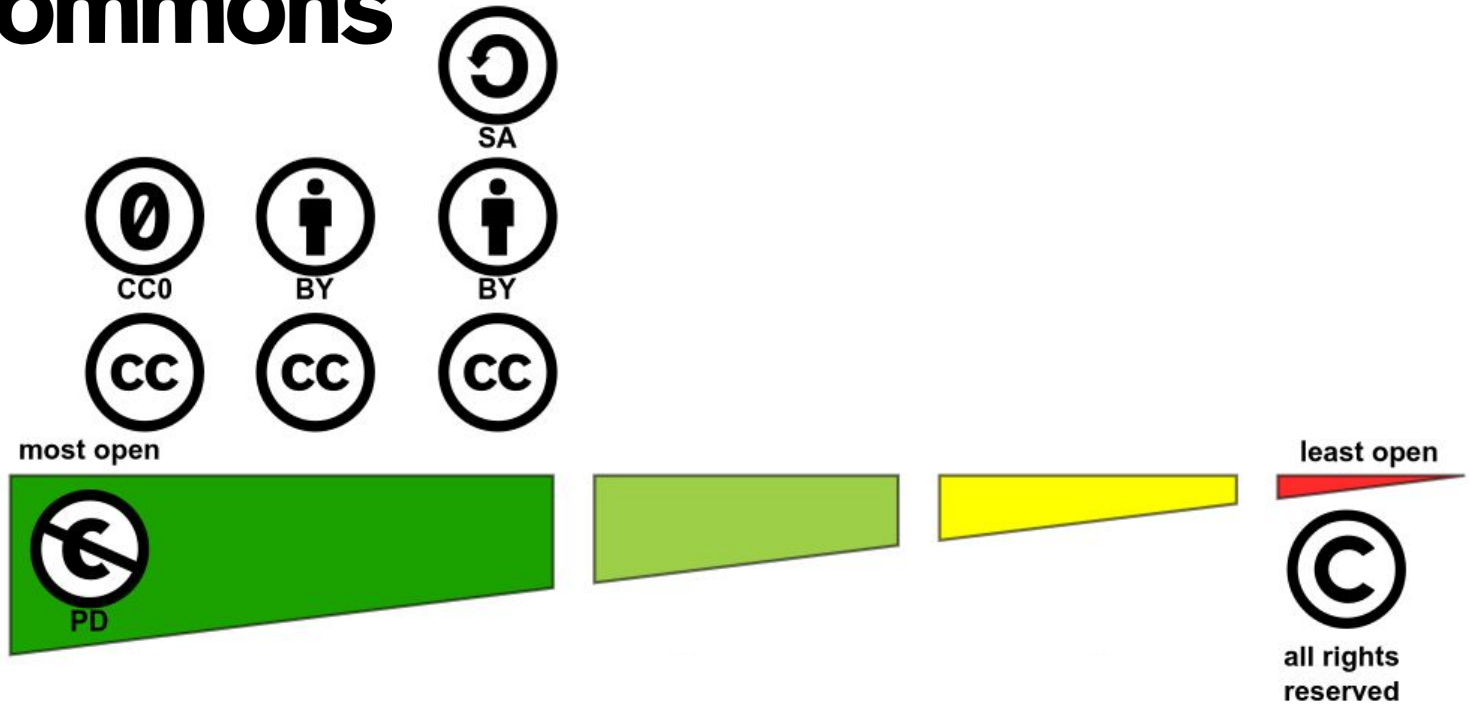
creative commons



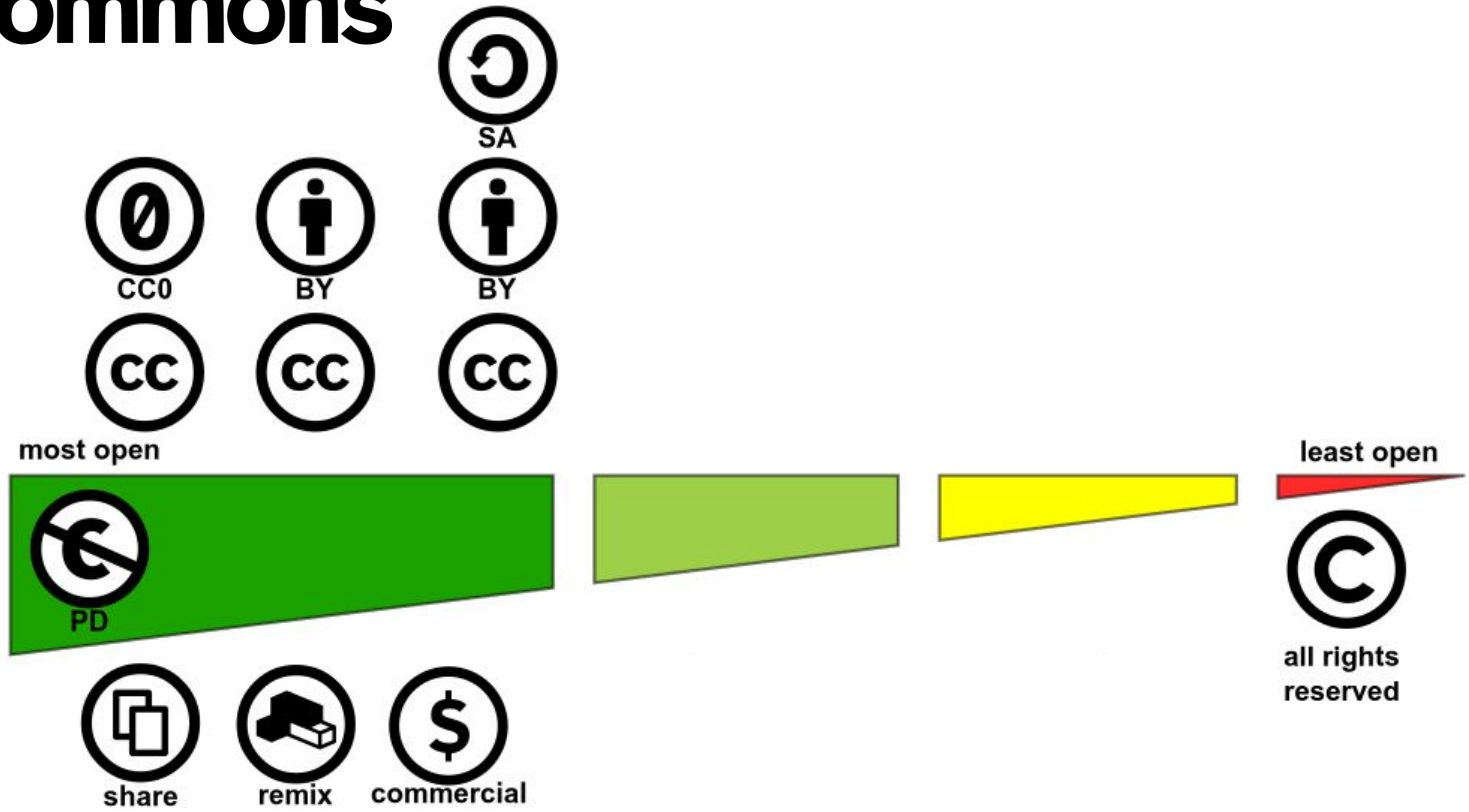
creative commons



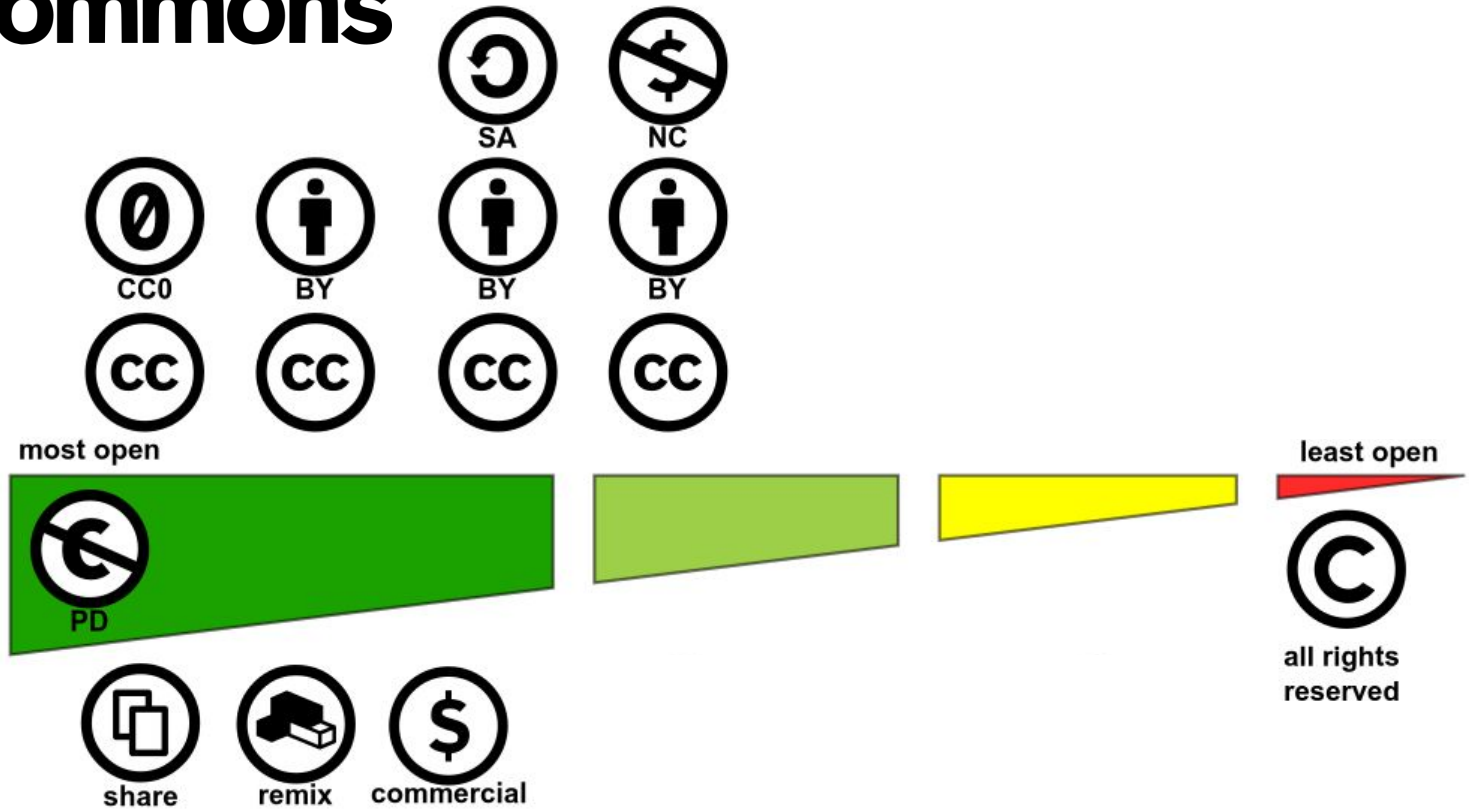
creative commons

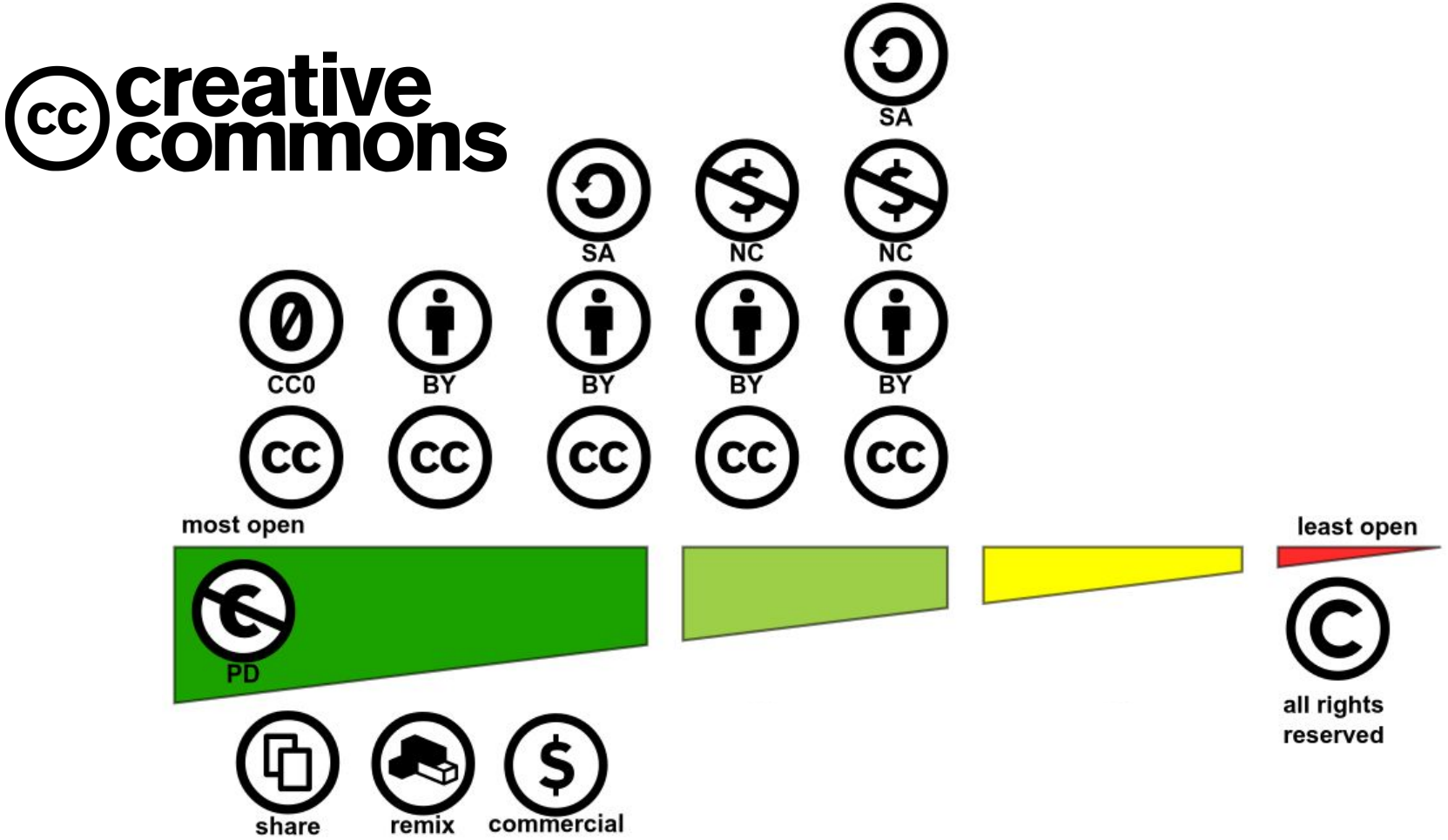


creative commons

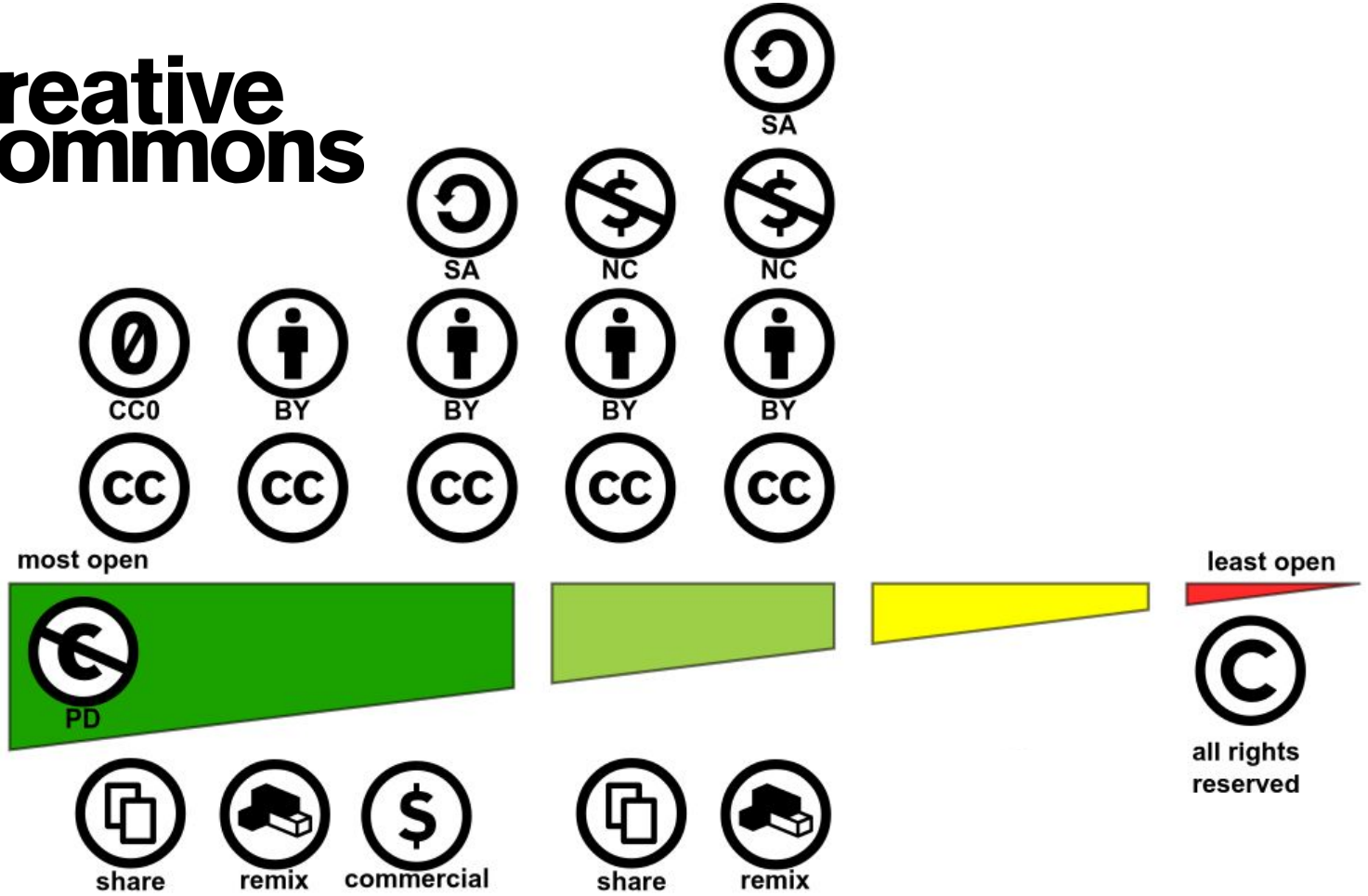


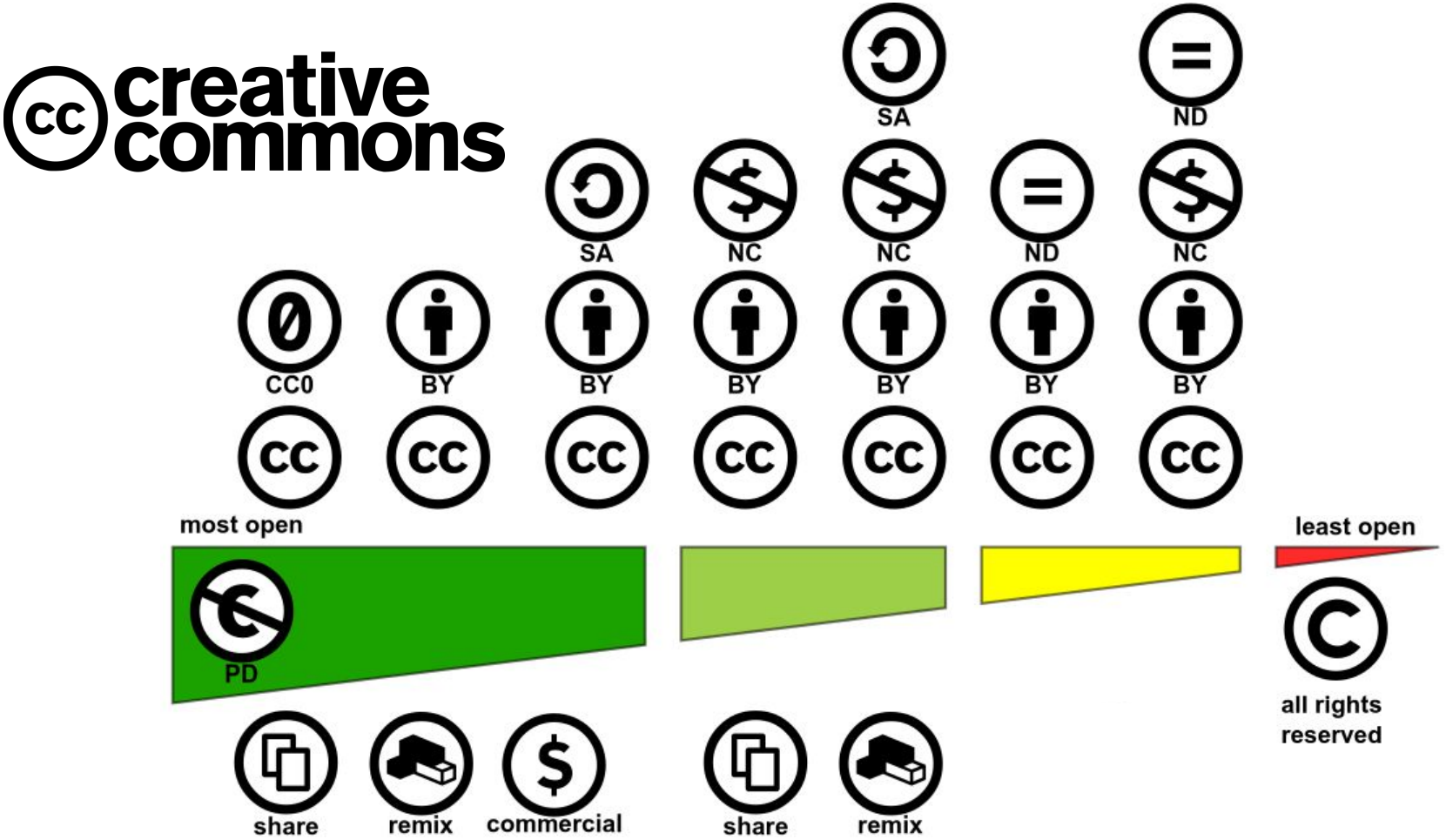
creative commons



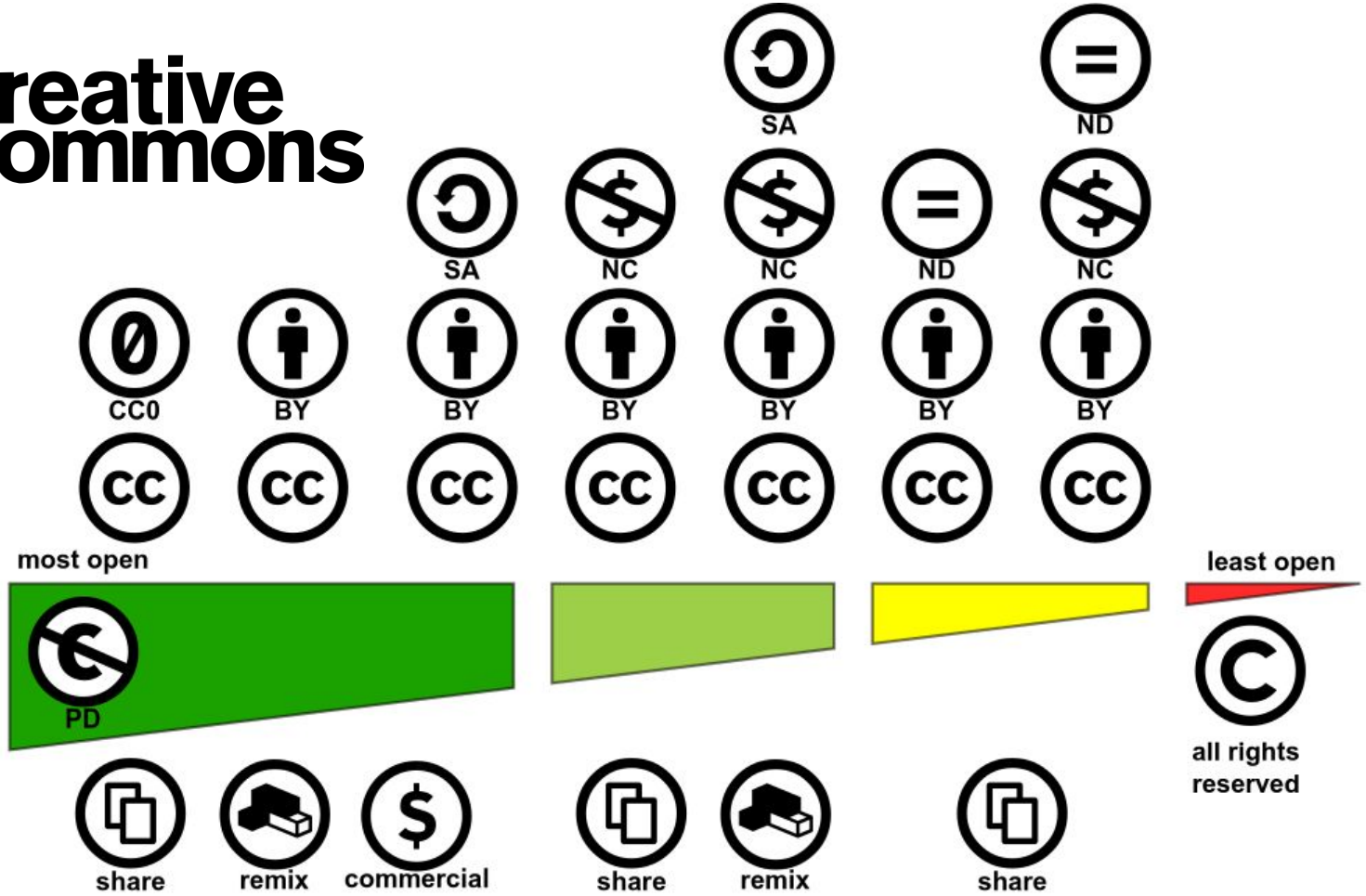


creative commons





creative commons

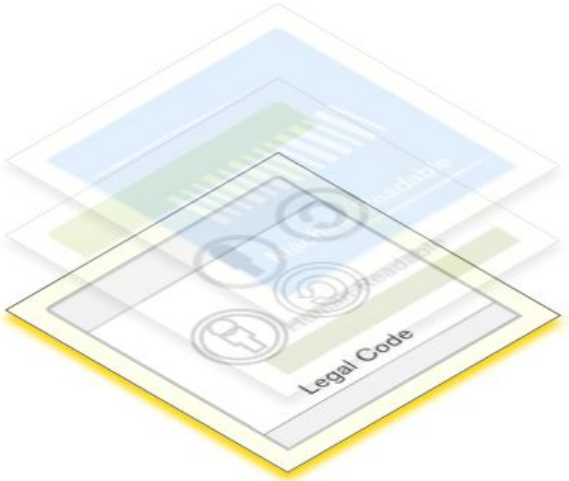


creative commons

“open”

creative commons

Legal code, harmonized for national/international law



creative commons

Legal code, harmonized for national/international law

Creative Commons **- What is licensed?** Human readable, understandable text

Nathan Yergler, Alex Roberts **- Who is to be attributed?**

Licensed to the public under CC BY 3.0 Unported **- Which license?**

creative commons

Legal code, harmonized for national/international law

Human readable, understandable text

Machine readable html tag attachable to metadata

Challenges

Multiple Attributions for several sources (license stacking)

Challenges

Challenges

Multiple Attributions for several sources (license stacking)

Multiple incompatible source licenses

Challenges

Multiple Attributions for several sources (license stacking)

Multiple incompatible source licenses

Legal commercial definition (e.g. use by journals)

Challenges

Multiple Attributions for several sources (license stacking)

Multiple incompatible source licenses

Legal commercial definition (e.g. use by journals)

Unintentional restrictive

Open Data Commons

Databases are different to simple data (e.g. EU-copyright)

Open Data Commons

Databases are different to simple data (e.g. EU-copyright)

Open Data Commons Public Domain Dedication and License (PDDL)

Open Data Commons

Databases are different to simple data (e.g. EU-copyright)

Open Data Commons Public Domain Dedication and License (PDDL)

Open Data Commons Attribution License (ODC-By)

Open Data Commons

Databases are different to simple data (e.g. EU-copyright)

Open Data Commons Public Domain Dedication and License (PDDL)

Open Data Commons Attribution License (ODC-By)

Open Data Commons Open Database License (ODbL)

Repository specific regulations

Individuals submitting data to the international sequence databases managed collaboratively by DDBJ, EMBL, and GenBank should be aware of the following:

https://www.ebi.ac.uk/ena/standards-and-policies

Repository specific regulations

ENA
European Nucleotide Archive

Individuals submitting data to the international sequence databases managed collaboratively by DDBJ, EMBL, and GenBank should be aware of the following:

The INSDC has a uniform policy of **free and unrestricted access** to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilising published scientific literature.

The INSDC will **not attach statements to records that restrict access to the data**, limit the use of the information in these records, or prohibit certain types of publications based on these records. Specifically, no use restrictions or licensing requirements will be included in any sequence data records, and no restrictions or licensing fees will be placed on the redistribution or use of the database by any party.

https://www.ebi.ac.uk/ena/standards-and-policies

Norwegian Licence for Open Government Data (NLOD) 2.0

Digitaliseringsdirektoratet

data.norge.no

https://data.norge.no/norwegian-licence-open-government-data-nlod-

Norwegian Licence for Open Government Data (NLOD) 2.0

Digitaliseringsdirektoratet
data.norge.no

A licence compatible by contract shall mean the following licences:

for all information: Open Government Licence (version 1.0, 2.0 and 3.0), **Creative Commons Attribution Licence (international version 4.0 and norwegian version 4.0)**

for those parts of the information which do not constitute databases: **Creative Commons Attribution Licence (generic version 1.0, 2.0, 2.5 and unported version 3.0) and Creative Commons Navngivelse 3.0 Norge**

for those parts of the information which constitute databases: **Open Data Commons Attribution License (version 1.0)**.

https://data.norge.no/norwegian-licence-open-government-data-nlod-

Norwegian Licence for Open Government Data (NLOD) 2.0

Digitaliseringsdirektoratet
data.norge.no

A licence compatible by contract shall mean the following licences:

for all information: Open Government Licence (version 1.0, 2.0 and 3.0), **Creative Commons Attribution Licence (international version 4.0 and norwegian version 4.0)**

for those parts of the information which do not constitute databases: **Creative Commons Attribution Licence (generic version 1.0, 2.0, 2.5 and unported version 3.0) and Creative Commons Navngivelse 3.0 Norge**

for those parts of the information which constitute databases: **Open Data Commons Attribution License (version 1.0)**.

https://data.norge.no/norwegian-licence-open-government-data-nlod-

Open Source Software licenses

Special considerations for Software

Liability

Warranty

Modifications

Network use = Distribution?

Open Source Software licenses

Open Source Software licenses

MIT license

GNU
AGPLv3

Open Source Software licenses

BSD licenses

MIT license

GNU AGPLv3

GNU GPLv3

GNU LGPLv3

Apache 2.0

Open Source Software licenses

open source initiative
Approved License®

BSD licenses

GNU

AGPLv3

MIT license

GNU GPLv3

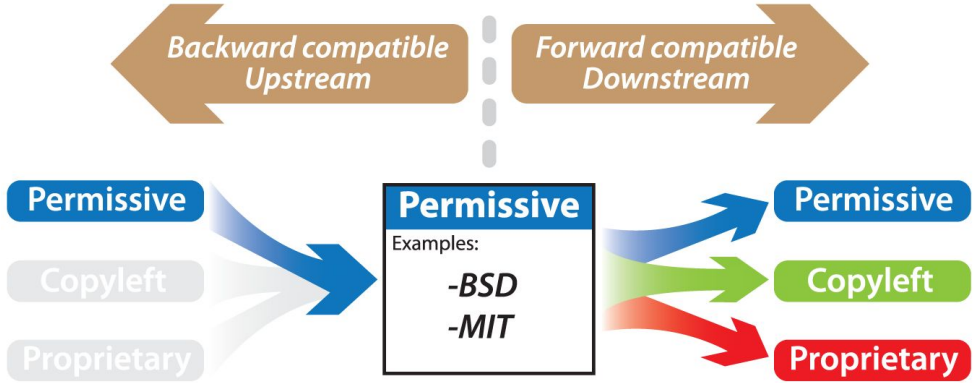
GNU LGPLv3

Apache 2.0

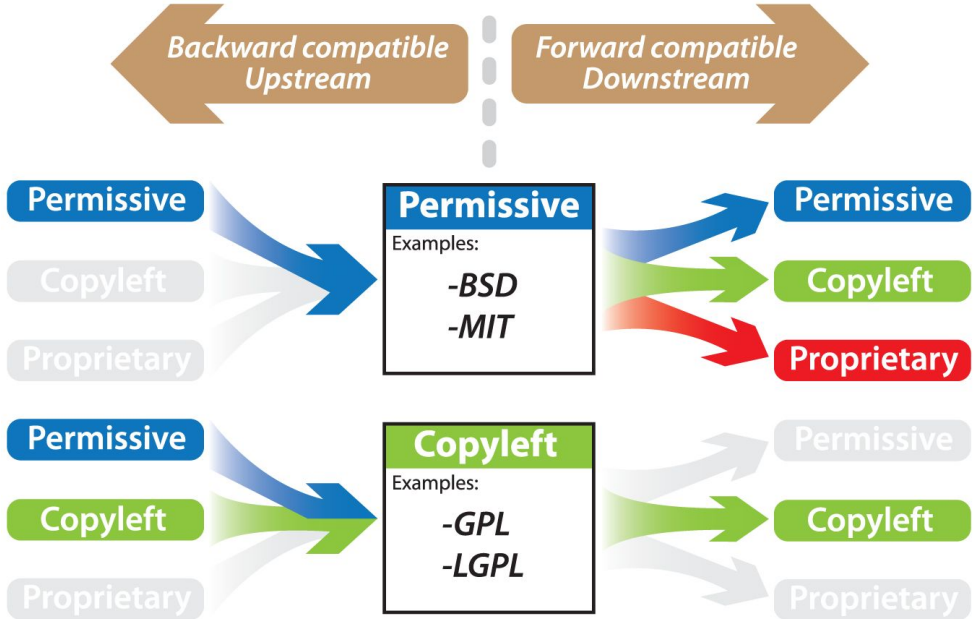
https://opensource.org/licenses
https://choosealicense.com/

GitHub

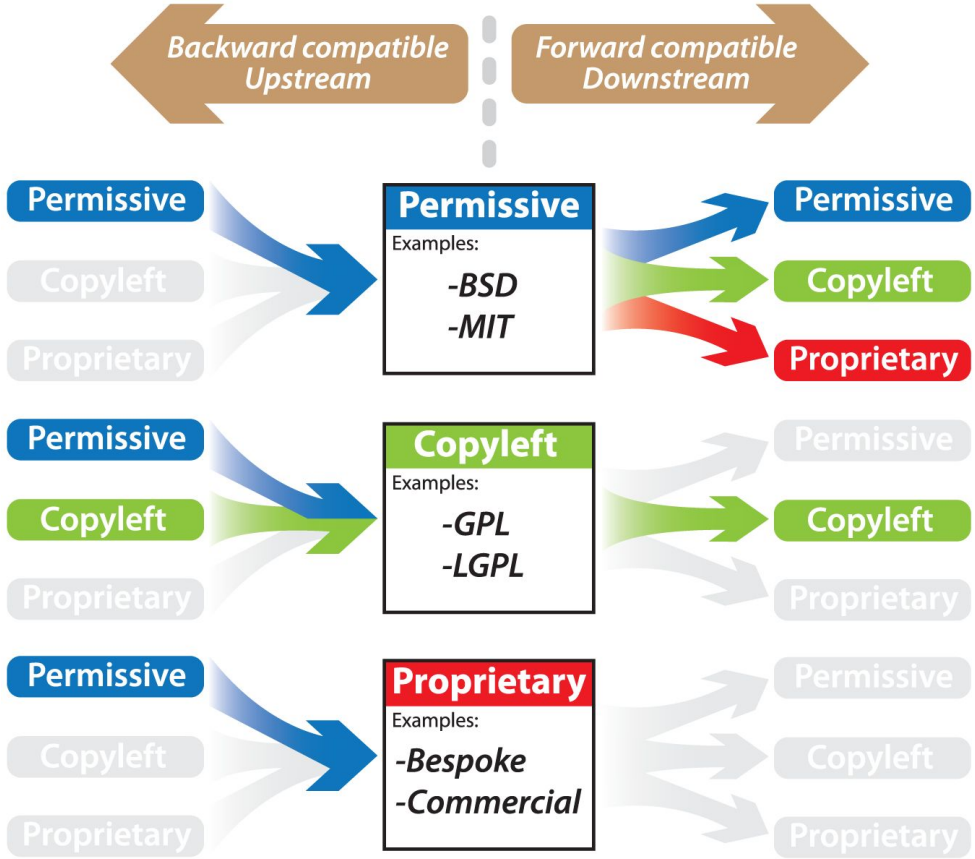
Open Source Software licenses



Open Source Software licenses



Open Source Software licenses



http://software-carpentry.org/

Material Transfer Agreements

Used for e.g. reagents, cell lines, plasmids, mice, …

Can safeguard your commercial interest, while allowing others to use your material for research

Can ensure attribution

Can enforce remain in Public Domain

Uniform Biological Material Transfer Agreement

Benefits to the Recipient:

Permission to use material for research or teaching purposes

Rights to all research results, modifications, and invention

Patent applications on modifications or inventions

Publishing without editorial comment or review by provider

Limited liability

https://autm.net/surveys-and-tools/agreements/material-transfer-agreements/mta-toolkit/uniform-biological-material-transfer-agreement
https://blog.addgene.org/why-do-i-need-an-mta-anyway

Data availability statement

“Data available upon reasonable request”

… so it’s not?

Gabelica, M., Bojčić, R. & Puljak, L. *Journal of Clinical Epidemiology* **0**, (2022). https://doi.org/10.1016/j.jclinepi.2022.05.019

Tedersoo, L. *et al. Sci Data* **8**, 192 (2021). https://doi.org/10.1038/s41597-021-00981-0

Data availability statement

RNA-Seq data are available in

the ArrayExpress database

http://www.ebi.ac.uk/arrayexpress/experiments/E-MEXP-12345/

under accession number E-MEXP-12345

https://www.ebi.ac.uk/arrayexpress/about.html

Data availability statement

RNA-Seq data are available in **- which dataset?**

the ArrayExpress database **- which repository?**

http://www.ebi.ac.uk/arrayexpress/experiments/E-MEXP-12345/ **- Link!**

under accession number E-MEXP-12345 **- ID**

https://www.ebi.ac.uk/arrayexpress/about.html

Data availability statement

RNA-Seq data are available in **- which dataset?**

the ArrayExpress database **- which repository?**

http://www.ebi.ac.uk/arrayexpress/experiments/E-MEXP-12345/ **- Link!**

under accession number E-MEXP-12345 **- ID**

(+access conditions & license)

-> Check repository and journal guidelines!

https://www.ebi.ac.uk/arrayexpress/about.html

Useful Resources

- Federated EGA Norway node, https://ega.elixir.no
- TSD (UiO) - https://www.uio.no/tsd
- SAFE (UiB) - https://www.uib.no/safe (English link top right)
- Hunt Cloud (NTNU) - https://www.ntnu.edu/mh/huntcloud
- Data Sharing https://rdmkit.elixir-europe.org/sharing
- Data Licensing https://rdmkit.elixir-europe.org/licensing
- Data Brokering https://rdmkit.elixir-europe.org/data_brokering
- EUDAT license chooser for data and software

Learning Activity

Data Sharing Quiz Link

Thank you!

elixir-norway.org

@elixirnorway

contact@bioinfo.no