Harmonising Access Procedures for Sensitive Data

Wim Hugo
Jorik van Kemenade

EOSC Conference - 21 September 2023

ODISSEI

Open Data Infrastructure for Social Science and Economic Innovations

The ODISSEI Roadmap Project is funded by the Dutch Research Council (NWO) under grant number 184.035.014.

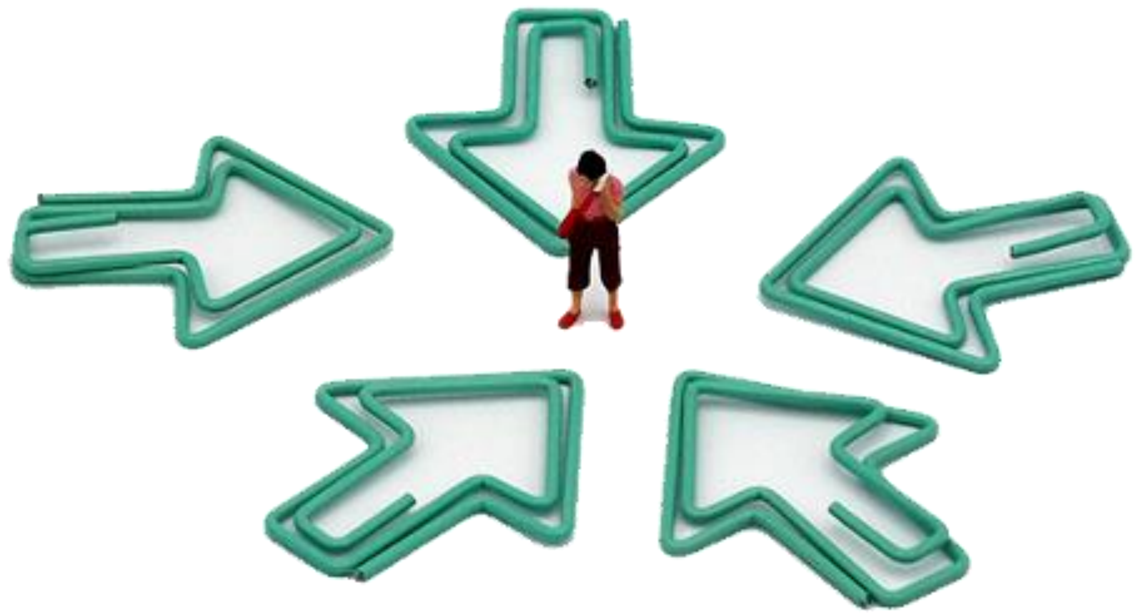
As open as possible as closed as necessary

We recommend Open Access and CC0 or CC BY where possible

Why would that not be possible?

- Personal information
- Sensitive information

Aim: protect the interests of the subjects of research



Images CC-BY-SA https://dmeg.cessda.eu/

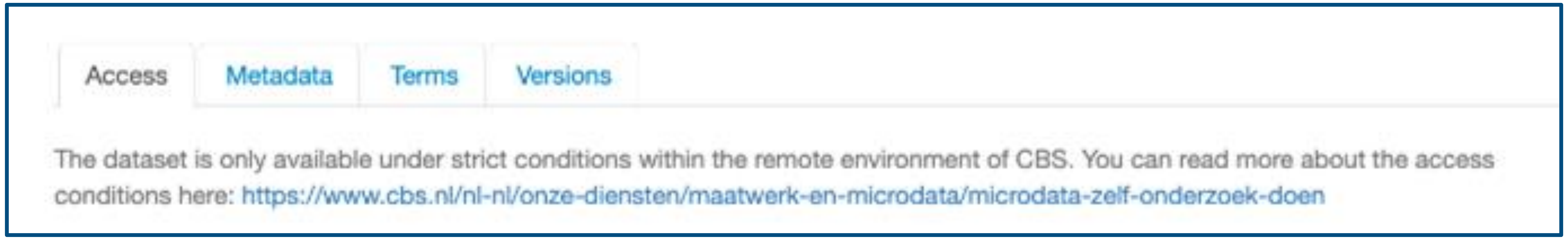
ODISSEI

The challenges of sensitive data

Sensitive data can often not be made available Open Access

→ Access needs to be managed

→ Managed Access = no standardized licence or access procedures



The dataset is only available under strict conditions within the remote environment of CBS. You can read more about the access conditions here: https://www.cbs.nl/nl-nl/onze-diensten/maatwerk-en-microdata/microdata-zelf-onderzoek-doen

?

Image CC-BY-SA https://dmeg.cessda.eu/

Context: Some background on licences

- A licence describes the ***rights and obligations of the owner***, and which of these rights and obligations are afforded/ transferred to the end user
- ***Two extremes***: 'All Rights Reserved' (for the owner), or 'Public Domain' (all rights waived)
- ***Embargo does not change the licence***, and is not part of the licence - it is a time period associated with workflow (publication) states of an object
- A ***licence can specify but not control*** the future availability of data outside of the repository - for example a requirement to destroy records after a certain period has elapsed
- Free culture licences (e.g. Creative Commons) do not allow ***restrictions***, but for sensitive data it is important. It may be partly included at present in ***terms of use*** that are provided outside the licence
- ***Terms of use should ideally be part of the licence,*** because the licence should be acted on independent of software platform and its business rules

Licence Elements

- Permissions
- Obligations/ Duties
- Limitations

- Restrictions

Ideal Licence Characteristics

- As uniform and simple as possible
- Machine-actionable
- As few as possible
- Apply in most jurisdictions

Problem to Solve

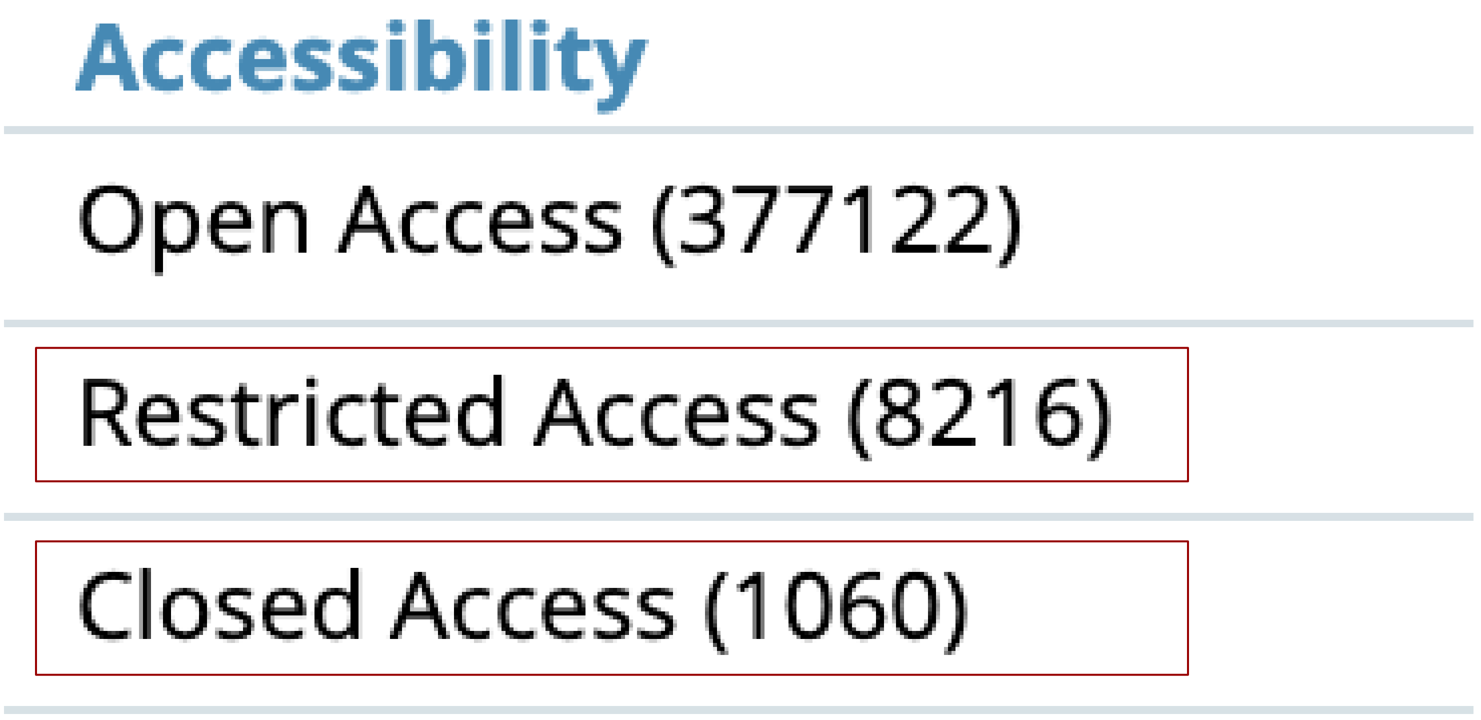
- Non-machine-readable licences, and licences that allow **arbitrary formulated conditions** and restrictions cannot be processed as automatically as possible, and as a result, infrastructure is not maximally scalable.
- Moreover, such licences require specific code to be created in each application that wants to evaluate compliance.
- Arbitrary licences are **not guaranteed to be valid** in any or more than one jurisdiction.
- Arbitrary conditions are **contrary to the spirit of Open Science** - it often discriminates on the basis of class, for example, by limiting access to a specific cadre of end users.
- In software platforms that allow additional access conditions to be specified, these are often divergent, and may **contradict the licence provisions**.

State of Play

Data Archiving and Networked Services

DANS

NARCIS



Accessibility

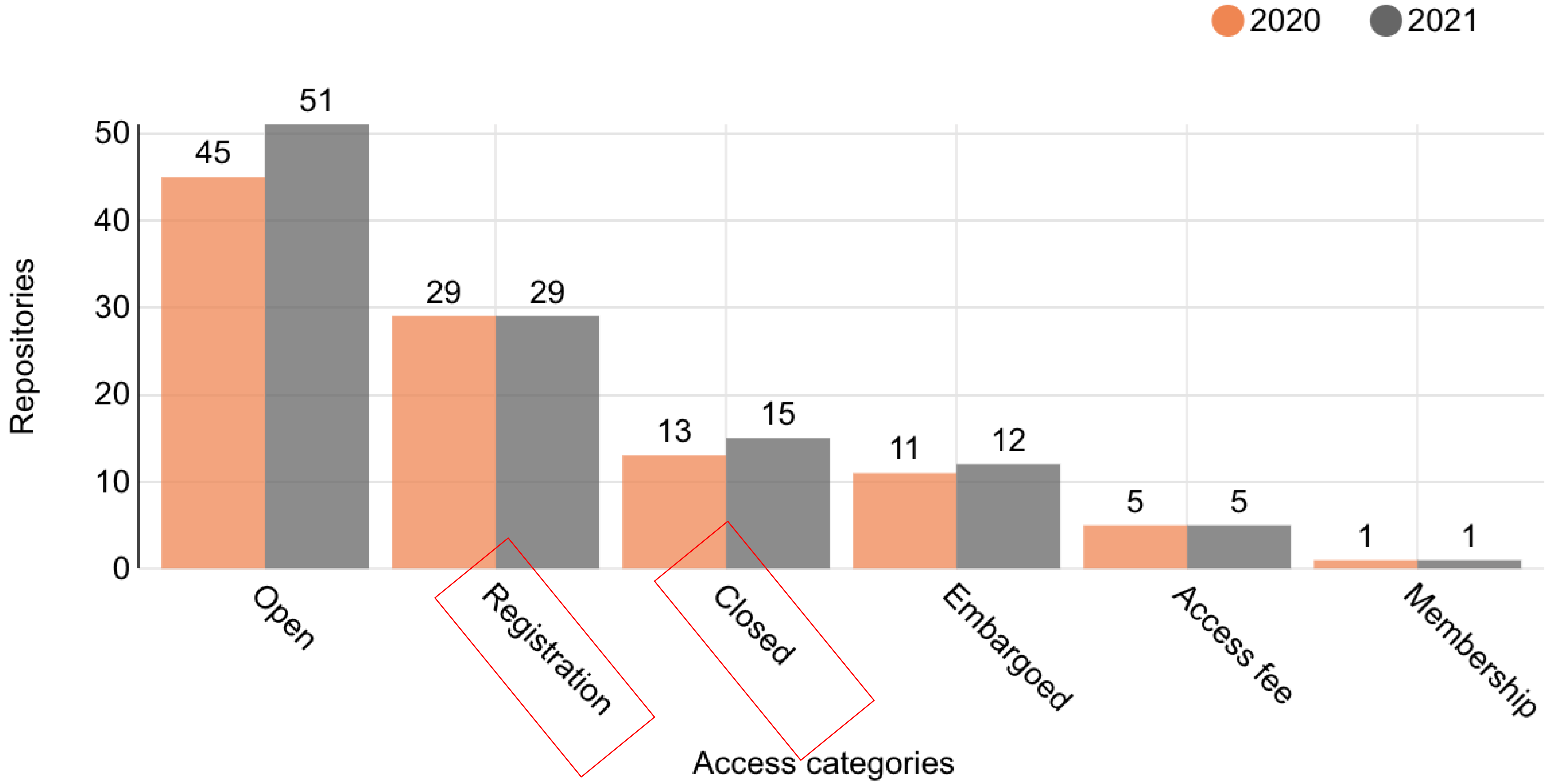
Open Access (377122)

Restricted Access (8216)

Closed Access (1060)

https://www.narcis.nl/search/coll/dataset/Language/en

Fig. 7. Access categories applicable in data repositories, 2020-2021

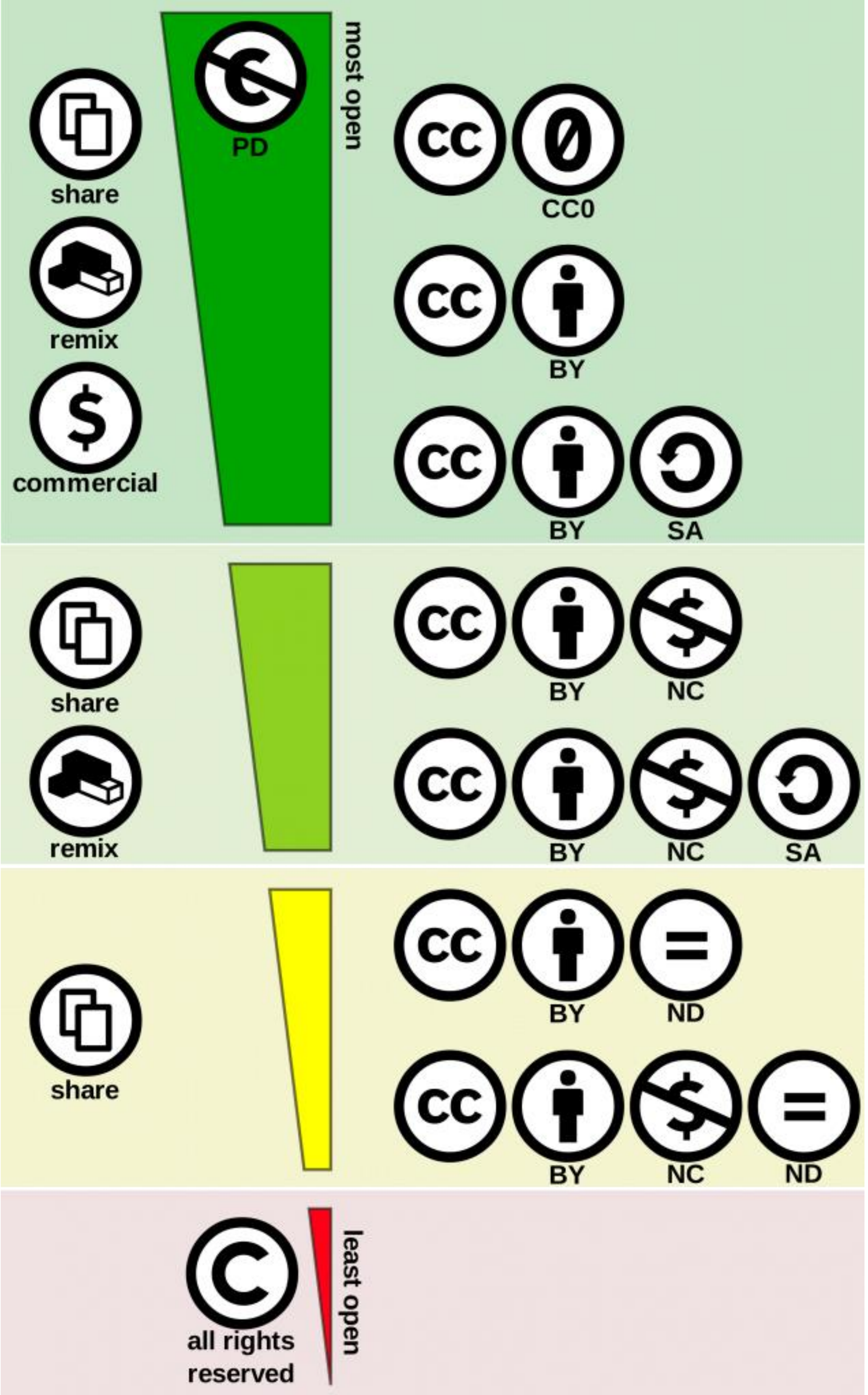


https://dans.knaw.nl/en/data-expertise/monitoring-and-analysis/

ODISSEI

Main considerations

Free Culture



Open Science

“As Open as Possible,
As Closed as Necessary”

According to the H2020 Program Guidelines on FAIR Data, data should be “as open as possible and as closed as necessary”, “open” in order to foster the reusability and to accelerate research, but at the same time they should be “closed” to safeguard the privacy of the subjects

Avoid Arbitrary Restrictions
Transparent Conditions

FAIR

Accessible
Once the user finds the required data, they need to know how they can be accessed, possibly including authentication and authorisation.

Reusable
R1.1. (Meta)data are released with a clear and accessible data usage licence

Specific Use Cases Covered

- We need to **identify the end user** and be able to **communicate**, since
  - consent to use a dataset may be withdrawn, the dataset may not be retained beyond a certain date, in which case the repository needs to inform the end user.
  - data exchange outside of [a jurisdiction] or countries approved by the [jurisdiction] may not be allowed, in which case the end user location needs to be verified.

- We need to **verify the end use** of the dataset, since it may harm the interests of the subject(s): Individuals, commercial entities, ecosystems, communities, cultural or ethnographic groups, …

- The dataset is so **sensitive** that **any leak will have severe adverse consequences**

Specific Use Cases *Not* Covered

“Academic Use Only” or “Dutch Institutions Only” - not covered!

- We need to **identify the end user** and be able to **communicate**, since
  - consent to use a dataset may be withdrawn, the dataset may not be retained beyond a certain date, in which case the repository needs to inform the end user.
  - data exchange outside of [a jurisdiction] or countries approved by the [jurisdiction] may not be allowed, in which case the end user location needs to be verified.
- We need to **verify the end use** of the dataset, since it may harm the interests of the subject(s): Individuals, commercial entities, ecosystems, communities, cultural or ethnographic groups, …
- The dataset is so **sensitive** that **any leak will have severe adverse consequences**

Does not specify workflow - this is a different concern. Can involve pre-release and/ or pre-publication checks and checks on the bona fides of the researcher

No Harm Provisions: Managed Access

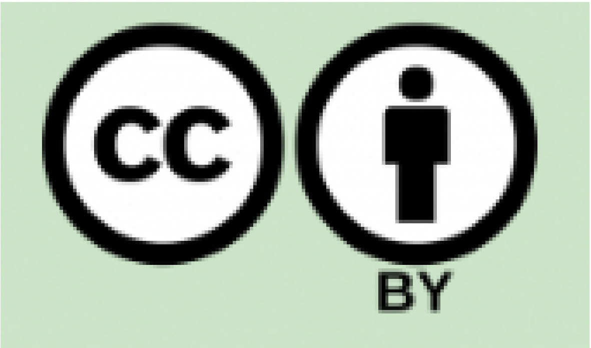
The principles of no harm provisions are:

1. They are needed to ***protect the subjects of research***, and not to arbitrarily or unnecessarily restrict access.
2. As such, one should view the limitations as ‘managed access’ and not ‘restricted access’.
3. Metadata is openly discoverable.
4. ***Anyone*** meeting the provisions (qualifications) of access will be granted access.
5. The basis of access adjudication should be known ***before access is requested*** and cannot be arbitrarily imposed.
6. Structured and responsible implementation of ‘As open as possible, as closed as necessary’.
7. Managed access ***does not mean unlimited access*** to a known individual. Future applications of a dataset by such an individual may not protect the subject’s rights.
8. Each case must be ***evaluated on merit*** because one must demonstrate active protection of subject rights.

Four New Licence Provisions

VI	**“Verified Identity”**	The identity and email address of the end user (no other personal particulars) are recorded for purposes of future communication in respect of the dataset.	🟢
VA	**“Verified Application”**	The purpose and nature of the derived dataset is verified by the depositor or curator. This may be done prior to processing, or prior to publication, or both. *This step may include additional checks on the person or institution.*	🟡
CE	**“Controlled Environment”**	The data can only be accessed and processed in a controlled environment.	🟢
HD	**“Hidden Data”**	The data itself cannot be seen by the process - only the output is visible to the end user.	🟢

Main considerations

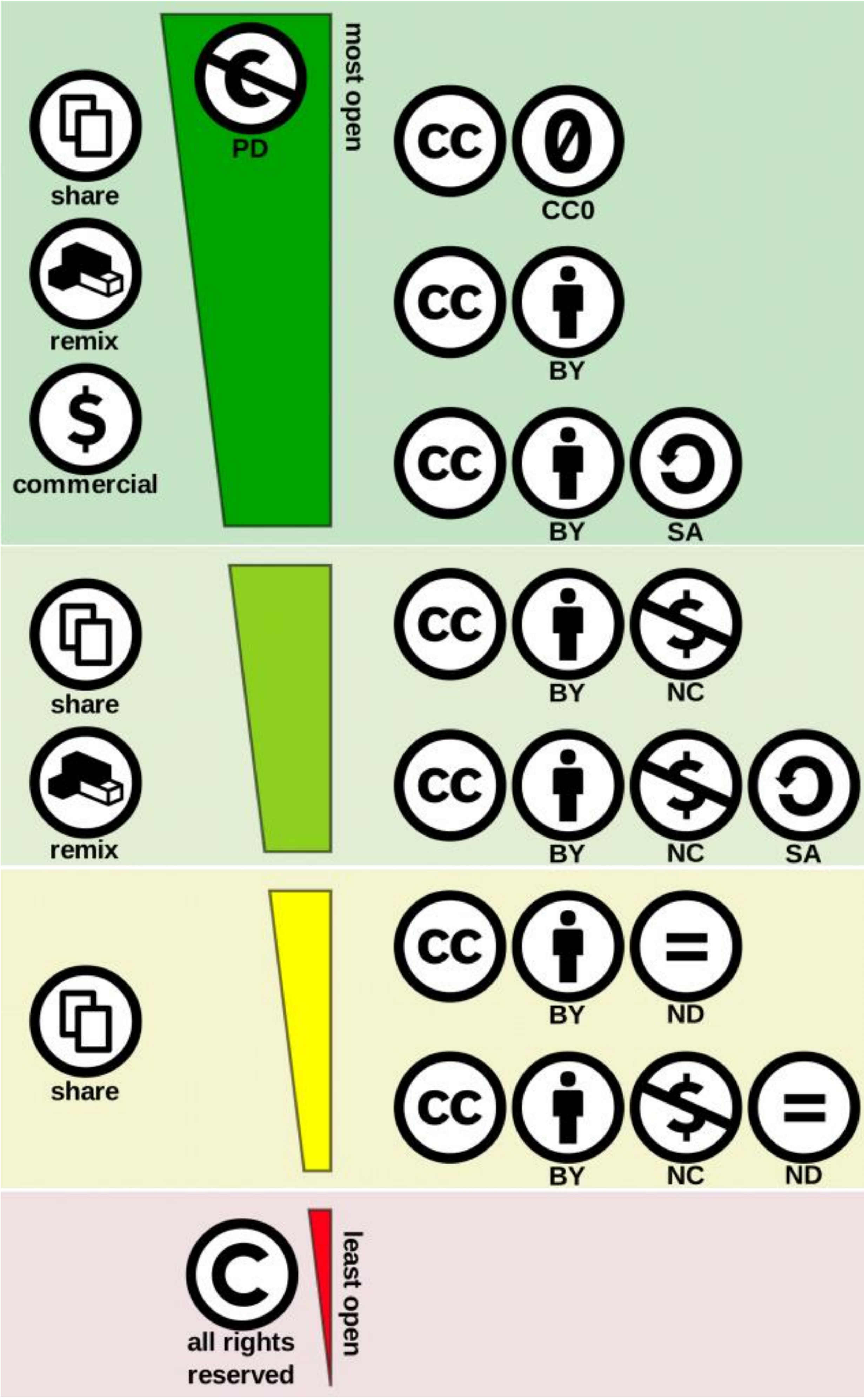


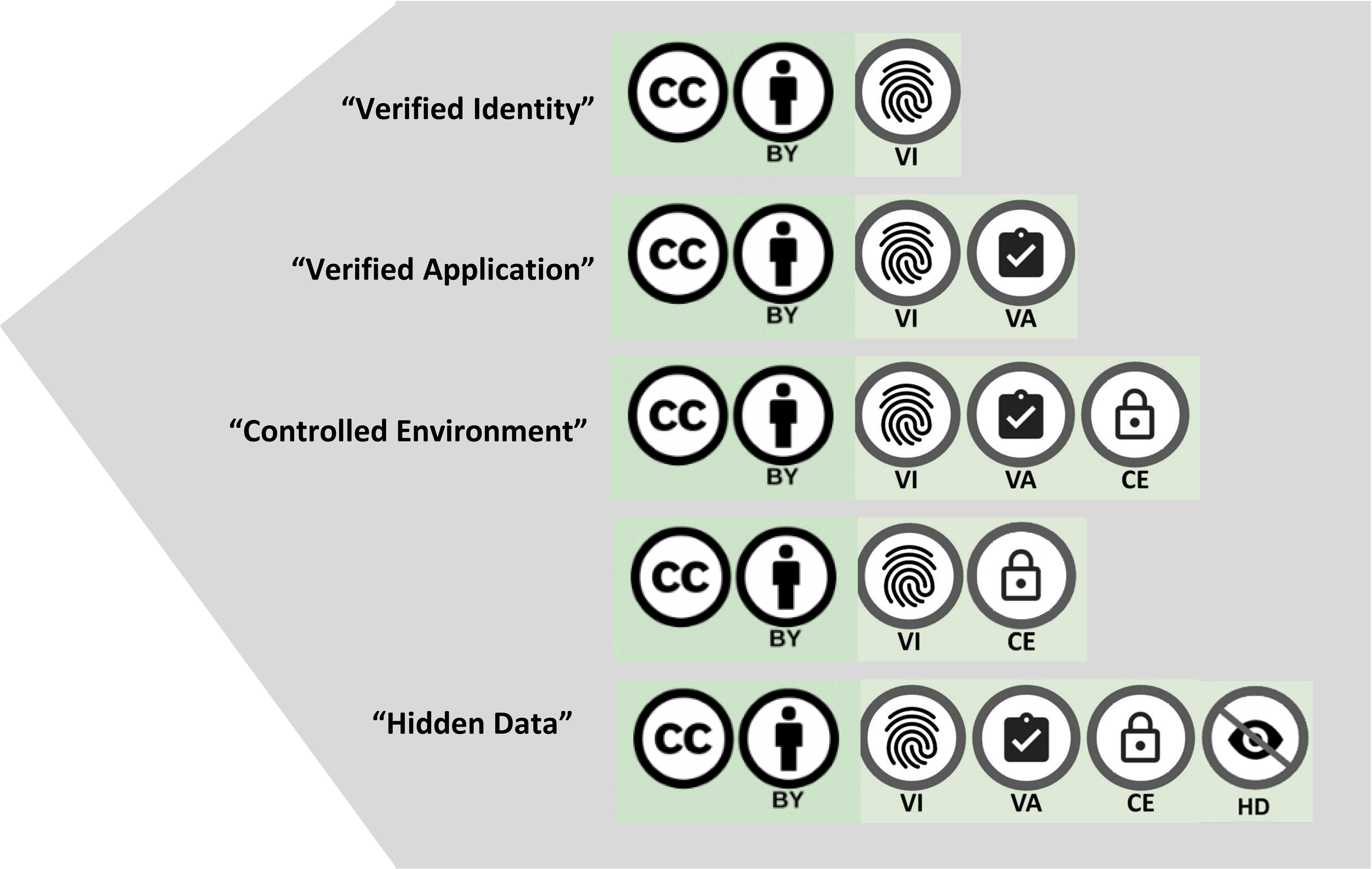
“Managed Access”

Retains CC BY provisions that are relevant

Extends with 4 new provisions

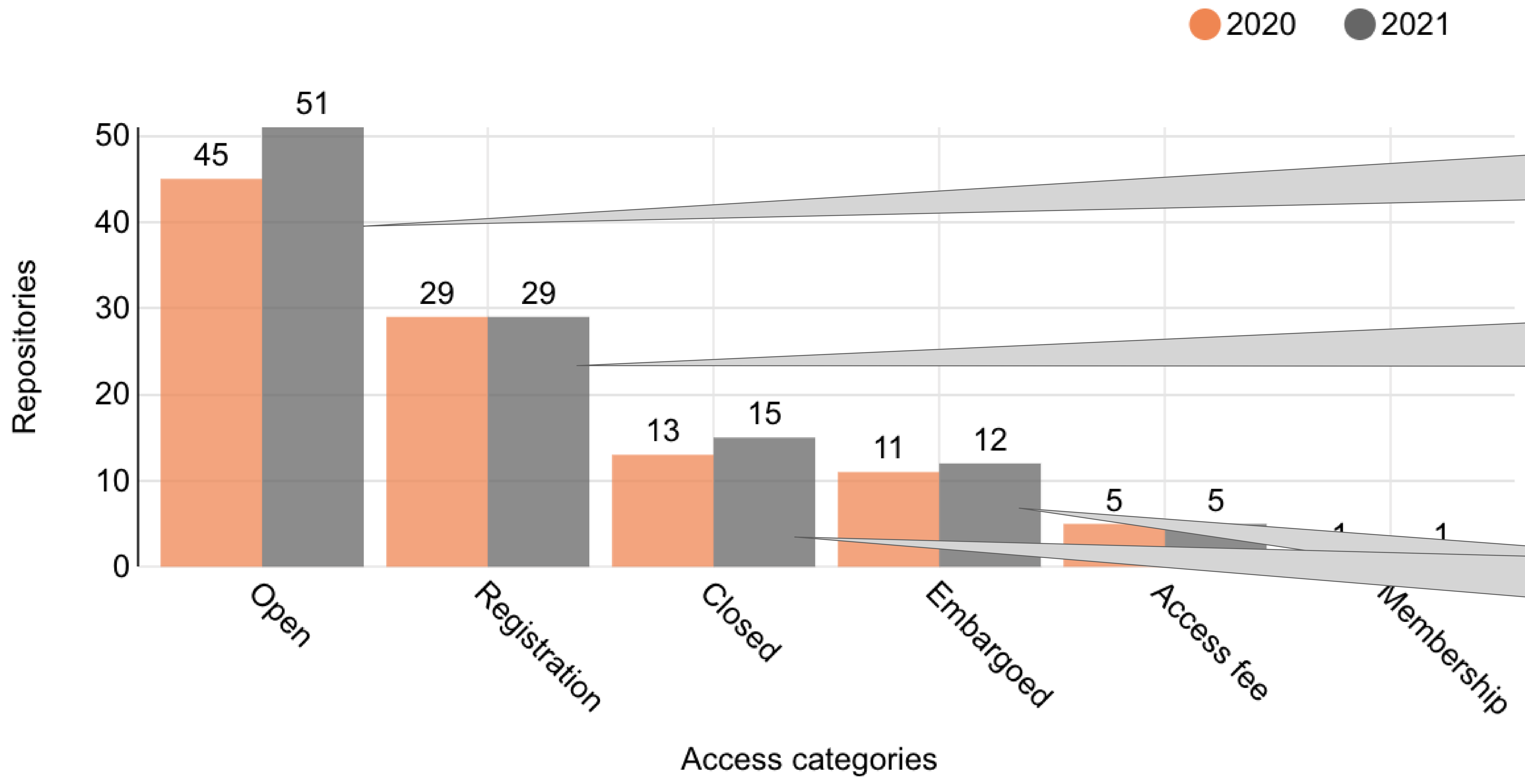
Free Culture





Impact of our proposal

Fig. 7. Access categories applicable in data repositories, 2020-2021



These are free culture licences and can be processed by machines

These can largely be covered by the 'Verified Identity' Managed Access Licence

Many of these can be covered by the other 2 Managed Access licences

Some known questions and considerations

1. In Europe, data copyright is constrained - but in practice, most datasets are published with CC-type licenses and CC BY licences are often recommended (implies copyright).

1. There are different approaches to ensuring compliance with access provisions - ranging from free selection on provisions from a standardisded portfolio, to a limited set of standardised licences. Allowing both approaches based on a set of standard provisions seems to be the most flexible.

Questions?

ODISSEI

Open Data Infrastructure for Social Science and Economic Innovations

Let’s stay in touch

Subscribe to our newsletter via https://odissei-data.nl
Follow us on Twitter @ODISSEI_NL

For questions:
info@odissei-data.nl

ODISSEI

Open Data Infrastructure for Social Science and Economic Innovations