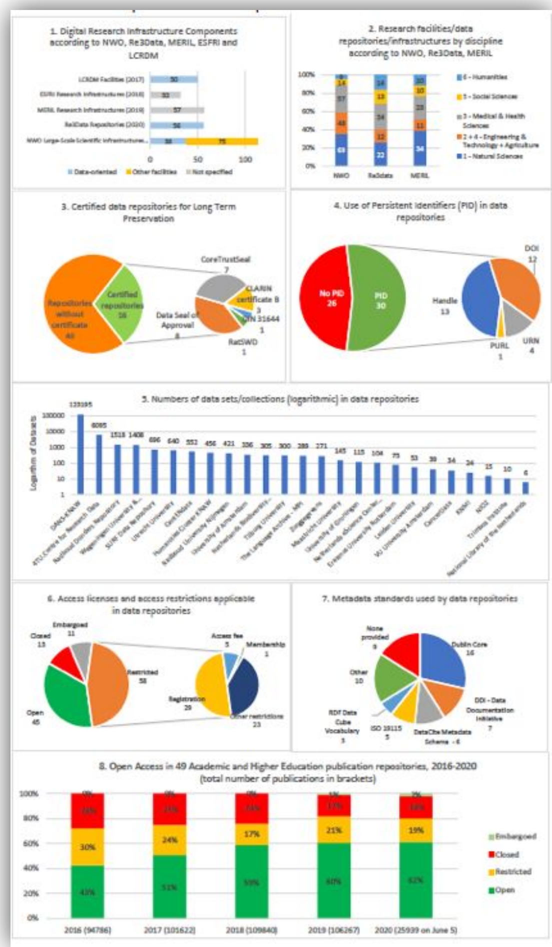
Data Archiving and Networked Services

DANS

EOSC SYNERGY

Dutch Open Science Dashboard 2020-2021

The Dutch Landscape Monitored in 10 Figures

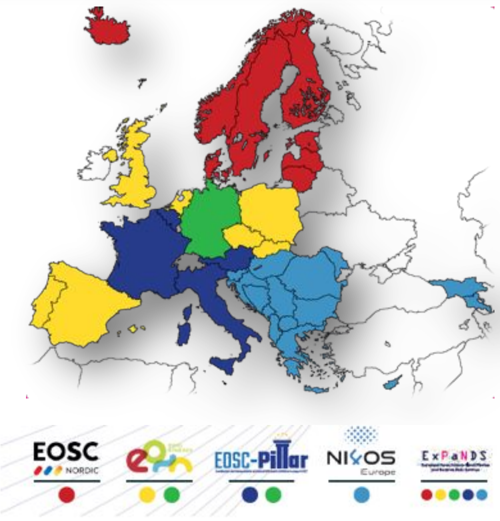

1. Digital Research Infrastructure Components according to NWO, Re3Data, MERIL, ESFRI and LCRDM
2. Research facilities/data repositories/infrastructures by discipline according to NWO, Re3Data, MERIL
3. Certified data repositories for Long Term Preservation
4. Use of Persistent Identifiers (PID) in data repositories
5. Numbers of data sets/collections (logarithmic) in data repositories
6. Access licenses and access restrictions applicable in data repositories
7. Metadata standards used by data repositories
8. Open Access in 49 Academic and Higher Education publication repositories, 2016-2020 (total number of publications in brackets)

Peter Doorn

@DANSKNAW @PKDoorn

Draft 3, 27 September 2021

Need for metrics on Open Science



In Collaboration with FAIRSFAIR EOSC-hub

EUROPEAN OPEN SCIENCE CLOUD

Working Proposal for Living Indicators to Monitor MS Progresses towards EOSC Readiness

Dimension	Main indicators	Sub indicators	Additional indicators
1: Architecture	3	6	6
2: Organisation & Governance	3	3	5
3: Policies	2	10	1
4: infrastructure	5	5	4
5: Training & skills	3	5	0
All 5 dimensions	16	29	16

First proposal: July 2020
Second proposal: January 2021

Present situation: 43 indicators selected, PoC for tool available;
However: - few metrics available at European level
- national metrics vary and are hard to compare.

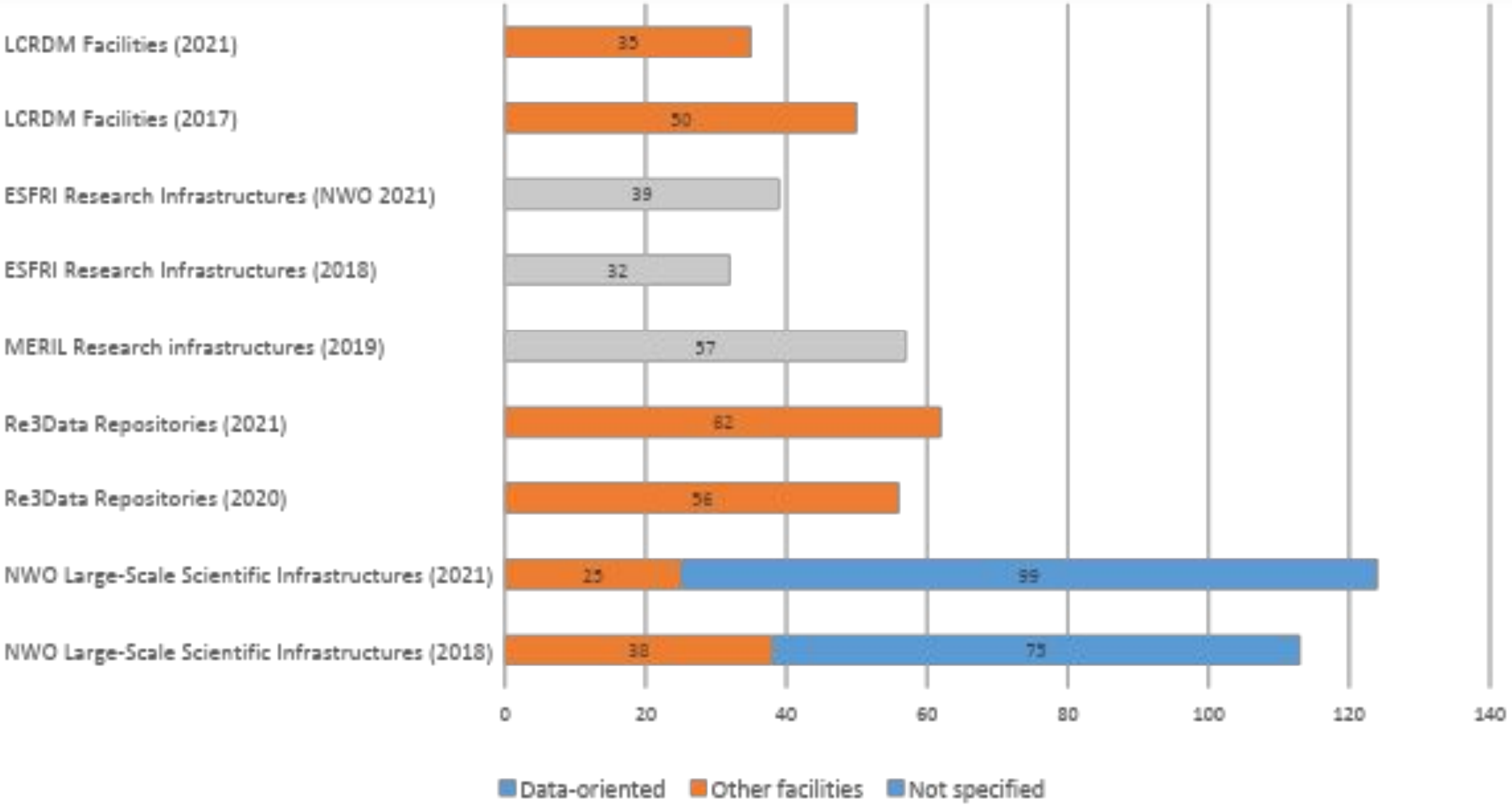
DANS Approach

- Based on indicators selected in EOSC Synergy project (Landscape & Gap analysis)
- Focus on quantitative metrics (not on narratives)
- Focus on DANS mission: related to Data Servics and Infrastructure
- Focus on deeds, not words (only implemented policies)
- First: manual & annual data collection, later automatic & real-time
- Focus on metrics that are available
- Limited number of indicators: presently 10

Metrics

1. Digital (Research) Infrastructures according to various sources and definitions
2. Research facilities/data repositories/infrastructures by discipline
3. Certified Repositories
4. PID systems implemented
5. Numbers of datasets in repositories (selection, harvested by NARCIS)
6. Datasets in University repositories
7. Access licenses and access restrictions
8. Metadata standards used
9. Content types present in repositories
10. Open Access in publication repositories

Fig. 1. Sources & Definitions: Digital Research Infrastructure Components according to NWO, Re3Data, MERIL, ESFRI and LCRDM



Sources: a, b, c, d, e (see last slide for descriptions and links)

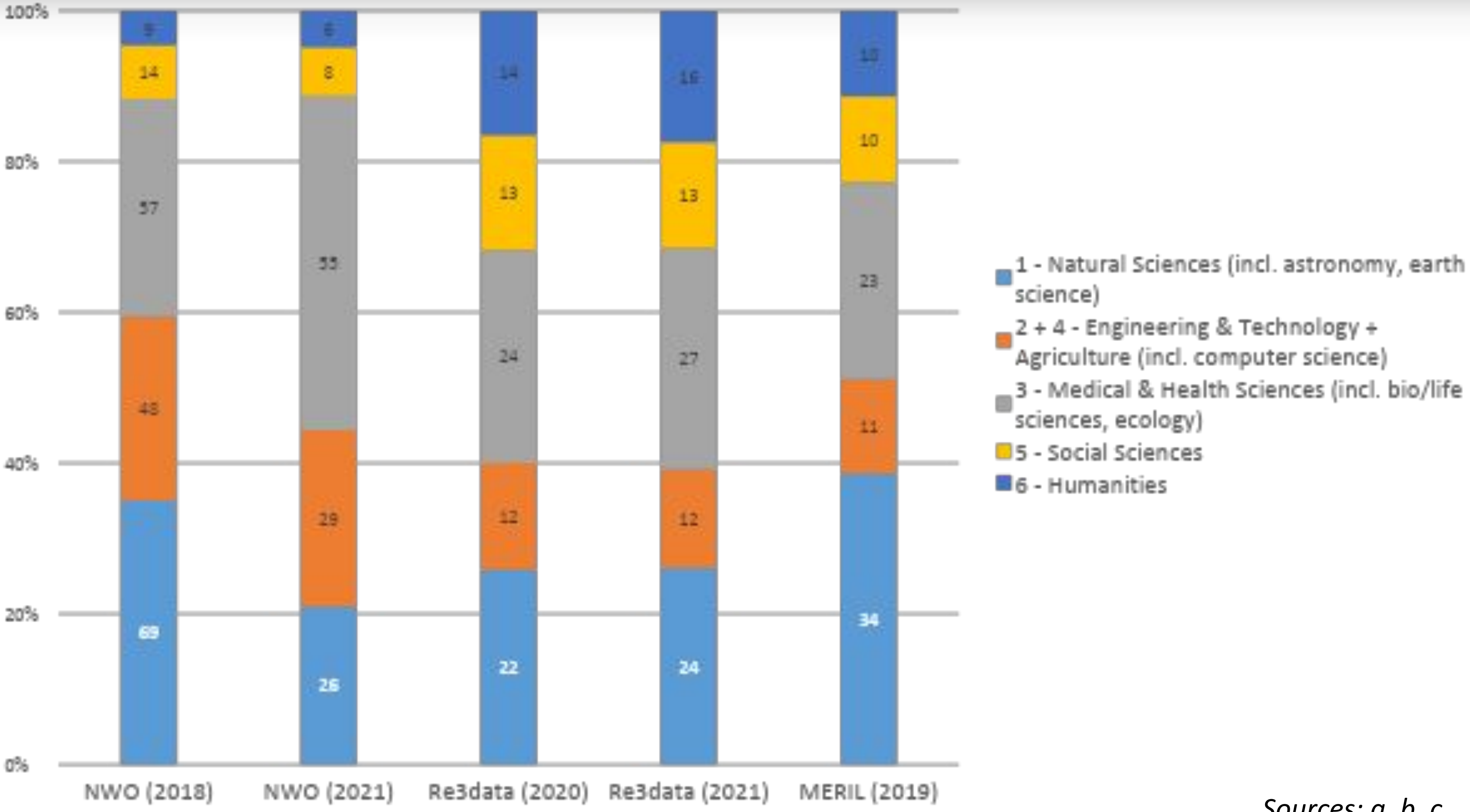
1. Sources & Definitions: Digital Research Infrastructure components according to NWO, Re3Data, MERIL, ESFRI and LCRDM

Different registers and overviews define digital research infrastructure differently.

- NWO updated its 2018-list of “large-scale scientific infrastructures”, also called “facilities” in 2021. Many of them qualify as equipment rather than as data facilities.
- Re3Data is an international registry of “repositories”, which include information systems and databanks of various nature
- MERIL and ESFRI provide overviews of research infrastructures:
  - MERIL is no longer maintained since 2019 and is being succeeded by CATRIS, which is however incomplete
  - ESFRI maintains a Roadmap, distinguishing R.I. “projects” and “landmarks”
- the overview of LCRDM “RDM facilities” was updated last year. The 2017 list was rather heterogeneous, including repositories, training courses, RDM services, etc. From the 2020/21 update we selected “repositories” and “infrastructures”.

Conclusion: the figures presented depend on what is measured and how. Keep in mind that registrations are seldom complete, and get quickly outdated. Still, we can safely assume the Dutch data infrastructure consist of > 100 components of some substance and recognition.

Fig. 2. Research facilities/data repositories/infrastructures by discipline according to NWO, Re3Data and MERIL



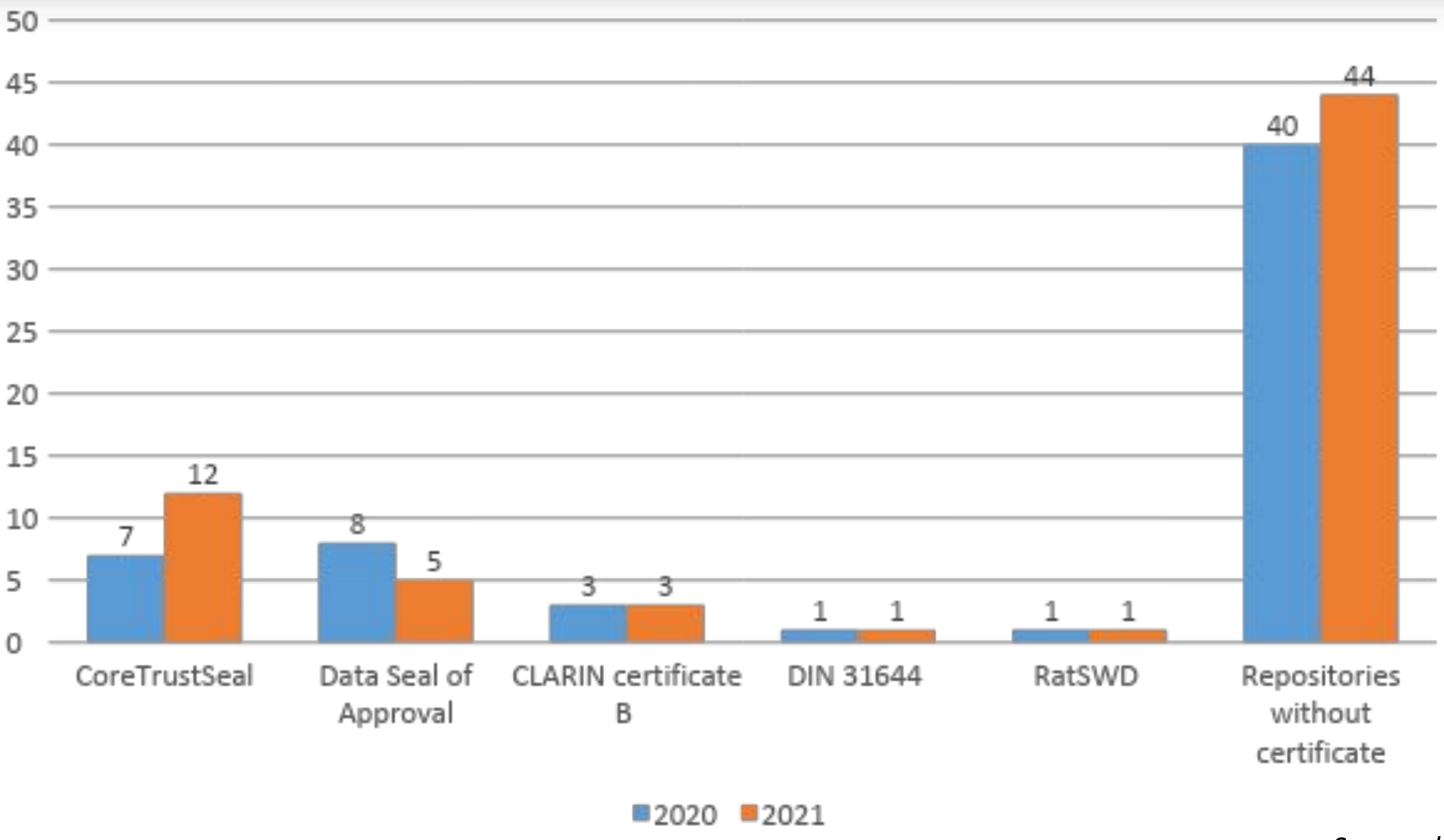
Sources: a, b, c

2. Research facilities/data repositories/infrastructures by discipline according to NWO, Re3Data, MERIL

Three sources make it possible to compare facilities according to discipline. The distribution over disciplines depends on the definitions used.

- The 2018 NWO-list provides multiple disciplines, in 2021 we classified them according to the primary discipline. No matter the classifciation, the humanities and social sciences (HSS) are strongly underrepresented in both 2018 and 2021 (12%).
- According to Re3Data, the repositories are more evenly spread across scientific fields.
- In terms of MERIL's general RIs, the share of the HSS is in between the figures of NWO and Re3Data (22%).

Fig. 3. Certified Repositories 2020-21



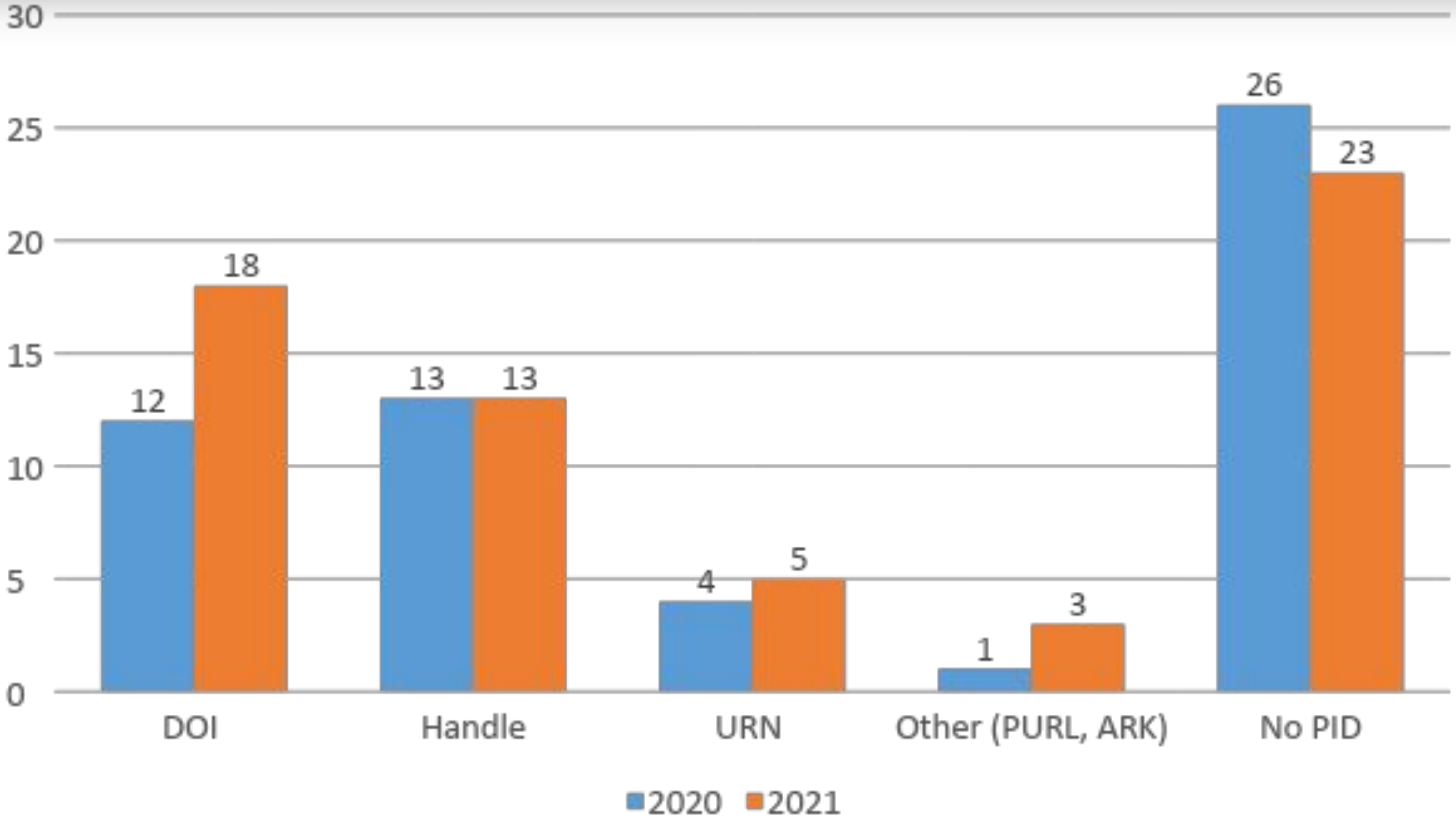
Source: b

3. Certified data repositories for Long Term Preservation

In 2020, 16 out of 56 data repositories in The Netherlands complied with some certificate for trustworthy long-term preservation; in 2021, the numbers increased to 18 out of 62.

The total number of certificates grew from 20 to 22 (4 repositories have more than one certificate). The Data Seal of Approval (DSA) is gradually superseded by the Core Trust Seal (CTS). 17 repositories comply with DSA or CTS in 2021 (in 2020: 15). These Seals make sure data is preserved and shared in a FAIR way.

Fig. 4. PID systems used in repositories, 2020-21



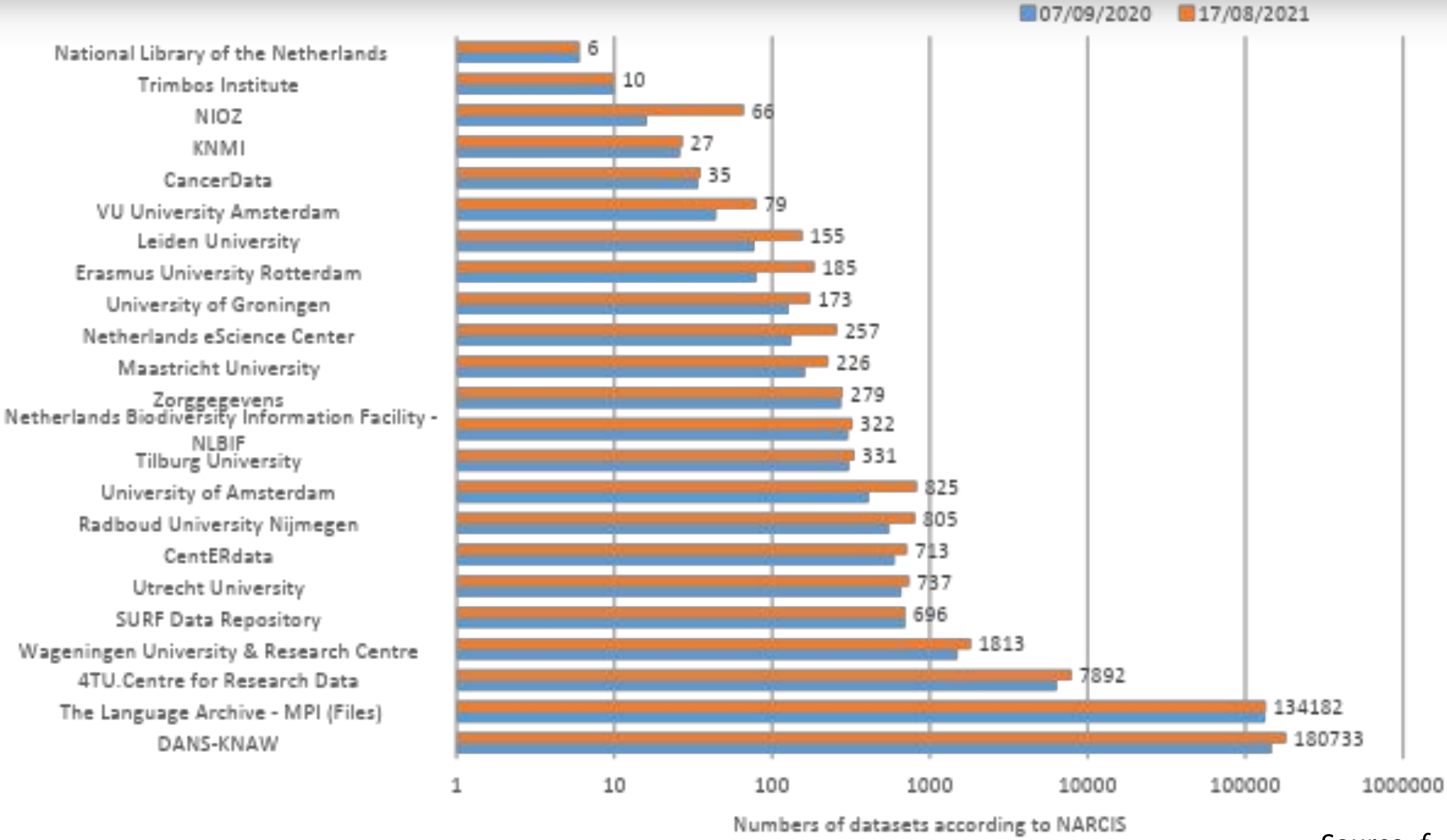
Source: b

4. Use of Persistent Identifiers (PID) in data repositories

The use of persistent identifiers is an important element of FAIR data. In 2020, 54% of the Dutch data repositories supplied PIDs, mostly Handle (23%) or DOI (21%), sometimes a URN (8%).

Over the past year, the numbers improved: now, 63% supplies a PID, and especially the share of DOI grew (29%), while Handle remained stable.

Fig. 5. Numbers of datasets (logarithm) in repositories harvested by NARCIS, 2020-2021



Source: f

5. Numbers of datasets in repositories harvested by NARCIS

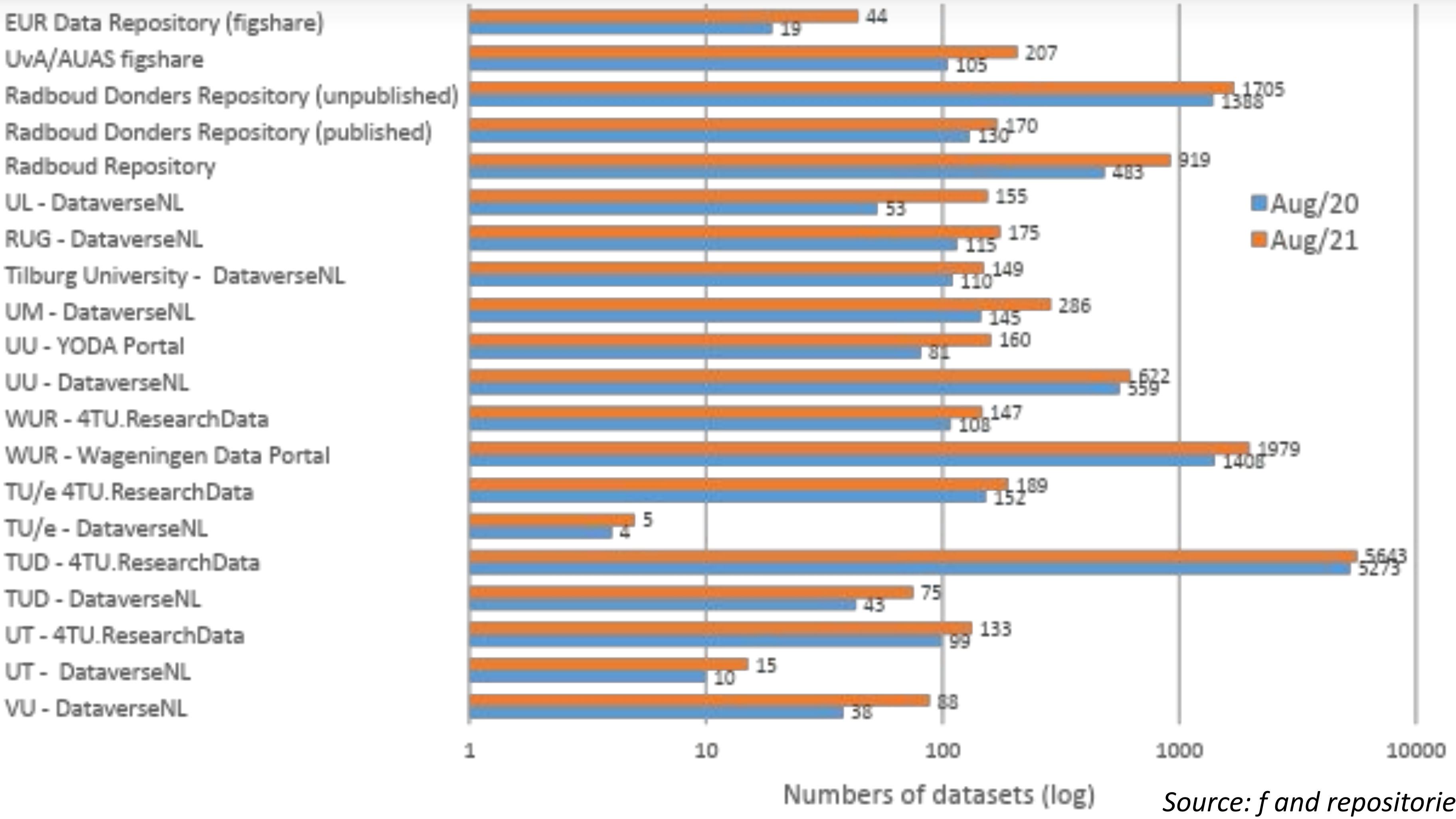
There is no complete overview of the research data in all Dutch repositories, let alone of data that are not stored in a repository.

The types of the data units (files, sets, collections) vary over (and within) repositories. One data set may consist of just one file, or of thousands of files. One file may be a small table or consist of millions of records. Such differences are partially dependent upon the discipline or community. E.g., The Language Archive of the Max Planck Institute for Psycholinguistics registers individual language files or bundles, that are parts of projects, archives or collections.

NARCIS aggregates the information on research data sets stored in 23 repositories. The total number of datasets in these repositories grew from 291,697 to 330,547 in the past year, an increase of 38,850 or 13%. 55% of these datasets are stored at DANS (in 2020 this was 50%).

More research into what data are stored under controlled conditions in repositories, and what data are not, is urgently required.

Fig. 6. Datasets in University repositories



Source: f and repositories

Data Archiving and Networked Services

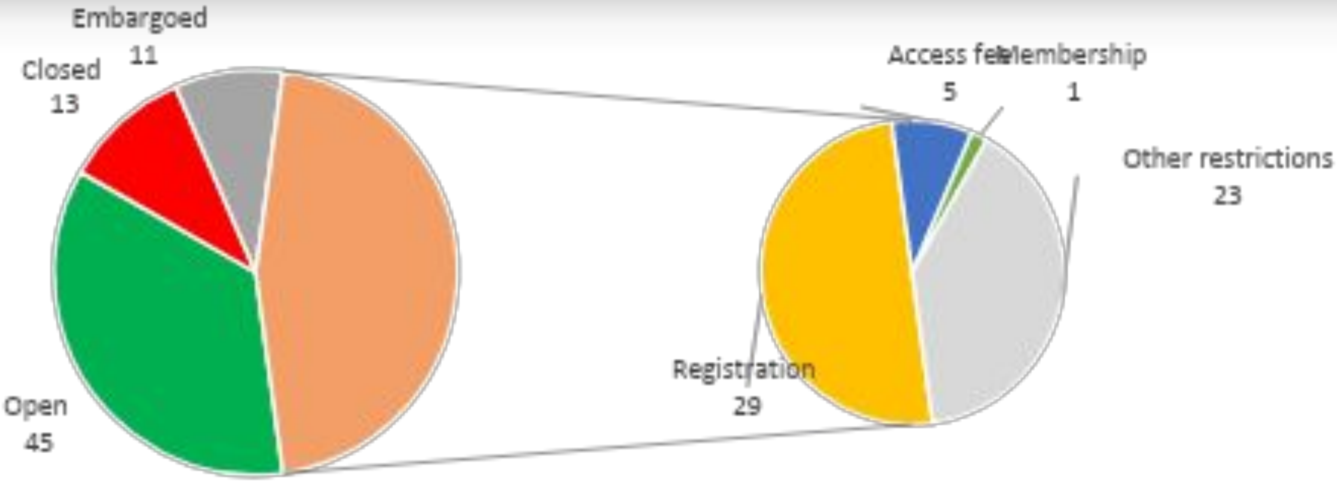
Fig. 6. Datasets in University repositories

Another view is obtained by looking just at the university repositories. 9 universities use Dataverse.NL hosted by DANS; the 3 TUs and Wageningen use 4TU.ResearchData (a figshare implementation). 2 More universities (UVA and EUR) also use figshare.

Nijmegen uses the home-grown repository system built by Donders Institute. Utrecht uses YODA next to Dataverse. Several universities use more than one data repository, next to their publication repositories.

As yet, the total contents of these data repositories is modest: 10,323 datasets in 2020, 12,866 now (August 2021). Yet, the increase of 2,543 datasets or 25% in one year demonstrates the potential.

Fig. 7. Access licenses and access restrictions applicable in data repositories, 2020-2021



2020

2021

98% of datasets in 23 repositories is openly accessible since 2017 (source: NARCIS)

Source: b

Fig. 7. Access licenses and access restrictions applicable in data repositories, 2020-2021

The repositories registered in Re3Data use a variety of licenses, ranging from fully open to several kinds of restrictions. As one can expect, this situation is quite stable over time.

In practice, the overwhelming majority (98%) of datasets in the 23 repositories harvested by NARCIS, appears to be openly accessible, with variations in the degree of openness.

NARCIS paints this situation perhaps too rosy: drilling deeper into the repositories, more restrictions apply than seems at first glance. E.g., The Language Archive distinguishes four access levels:

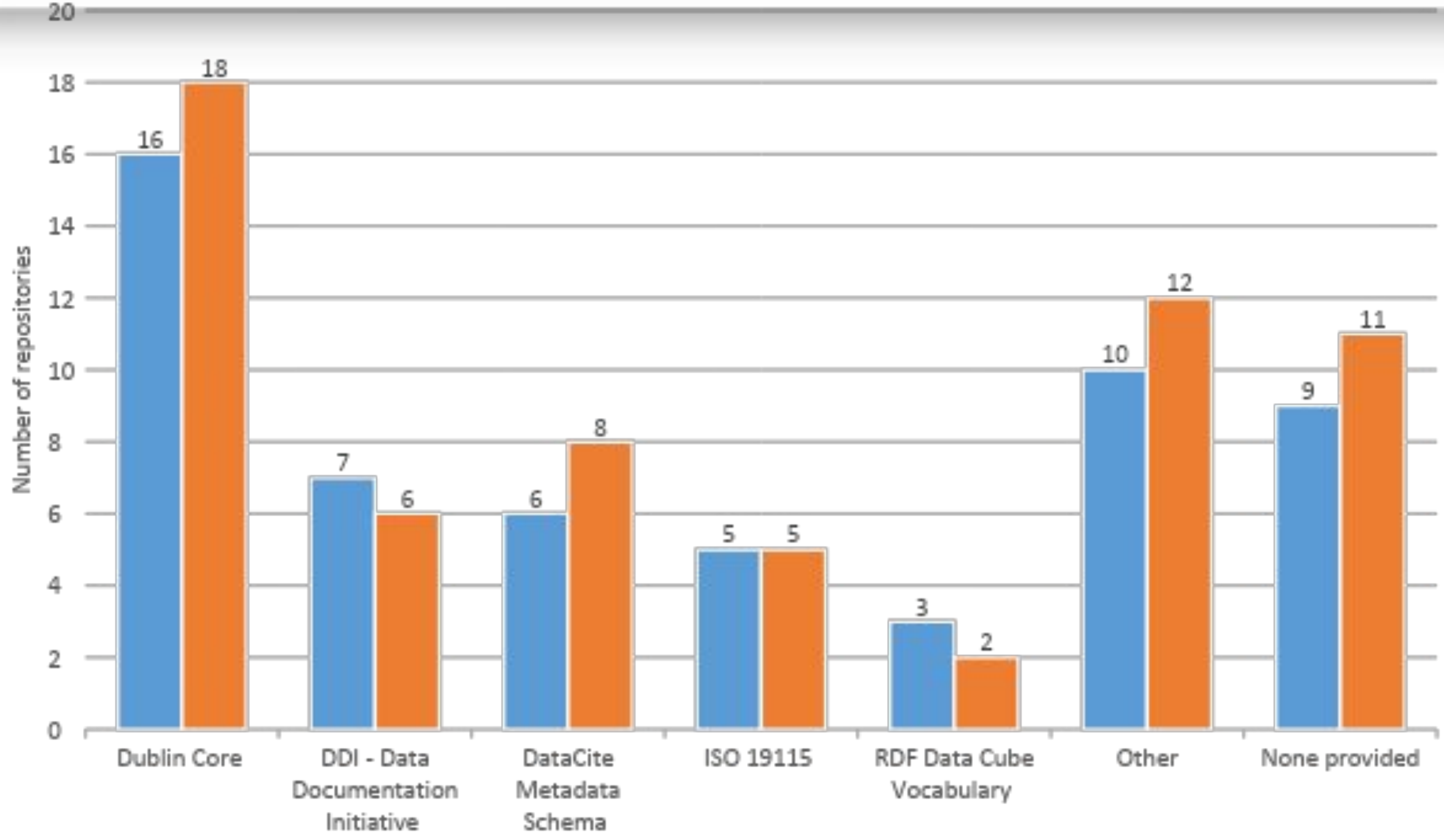
Open (27% of datasets): accessible to anyone (without registration).

Registered (9%): accessible for registered users.

Academic (2%): accessible for academic users.

Restricted (67%): accessible on request.

Fig. 8. Metadata standards used by data repositories, 2020-2021



Source: b

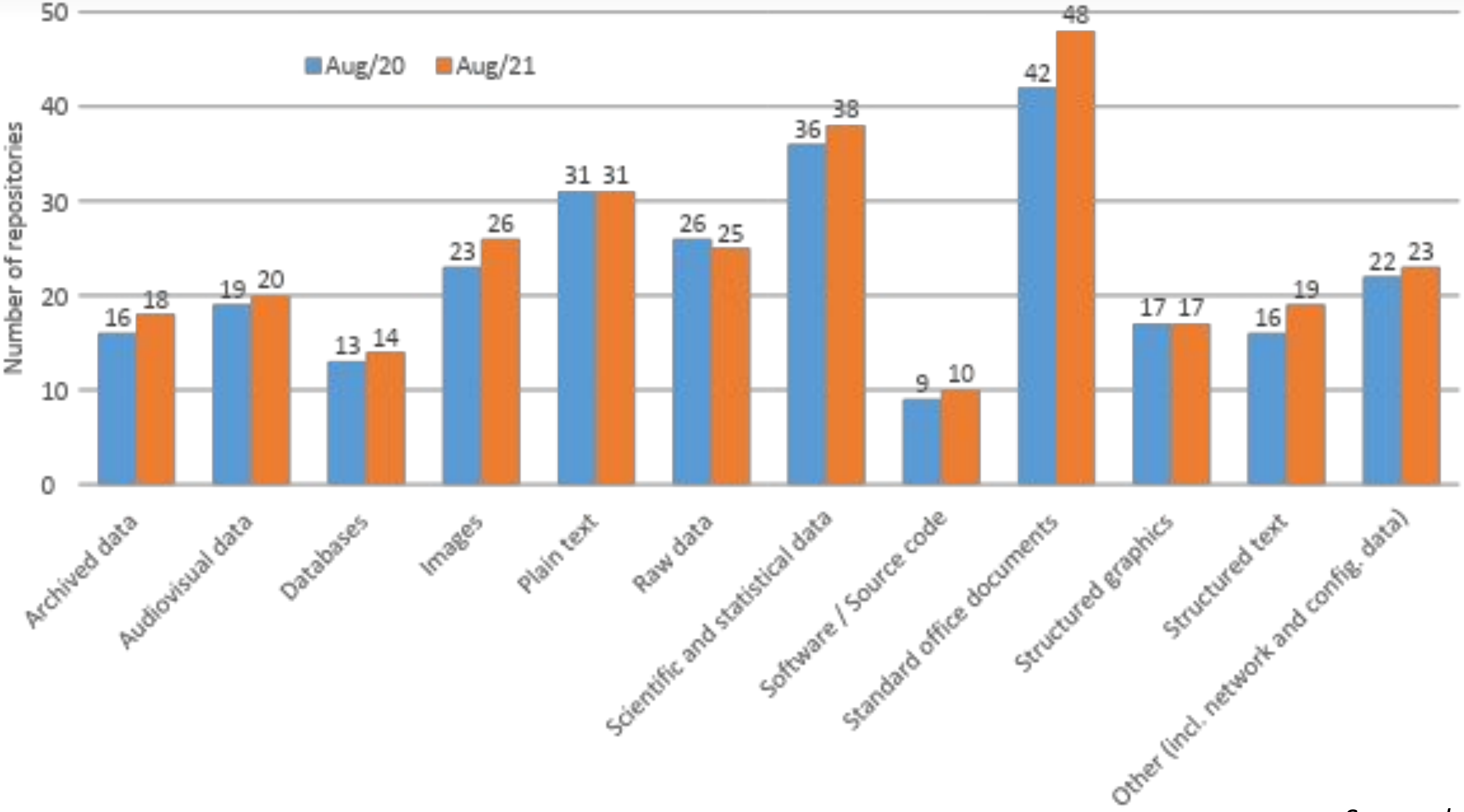
Fig. 8. Metadata standards used by data repositories, 2020-2021

The majority of repositories describes datasets according to a metadatastandard, among which Dublin Core, DDI and DataCite are most frequent.

The changes in the graph are caused by the increase of the number of repositories registered in Re3Data.

11 out of the 62 repositories registered in 2021 did not report to use a standard to describe their holdings.

Fig. 9. Content types present in repositories



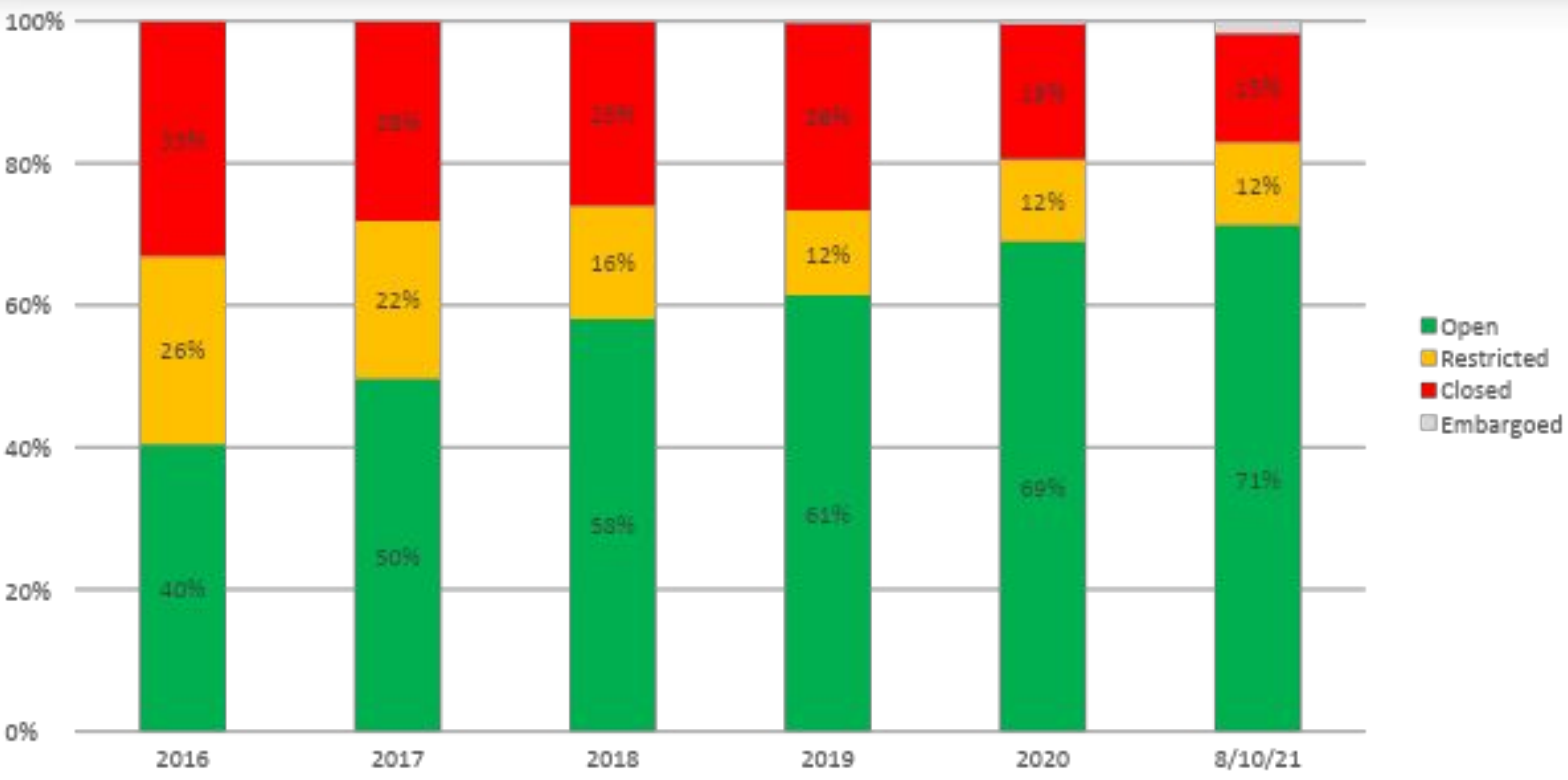
Source: b

Fig. 9. Content types present in repositories

Most repositories accept a variety of data types or formats for preservation and/or sharing. The main changes over the past year are caused by the increase of the number of repositories registered (from 56 to 62).

- Unsurprisingly, data types such as scientific/statistical (38x), raw (25x), archived (18x), databases (14x), and other data (23x) are most frequently mentioned by the 62 repositories registered in Re3Data.
- Text files are also frequently accepted: office documents (48x), plain (31x) and structured texts (19x) occur frequently PDF-documents are not separately recorded, but are likely to be an important category as well.
- Images (26x), A/V data (20x) and graphics (17x) are the third most important category.
- Software (source code) is an upcoming category, present in 10 repositories.

Fig. 10. Open Access in 38 Academic and Higher Education publication repositories, 2016-2021



Source: g

Fig. 10. Open Access in 38 Academic and Higher Education publication repositories, 2016-2021

This final graph monitors the progress in Open Access to scientific publications (not research data!) in 38 publication repositories.

The percentage of OA publications is clearly on the rise, from 40% in 2016 to slightly 71% in August 2021).

This is close to the original target set by the vice-minister for science in 2014, but still below the later and more ambitious target of 100% OA in 2020.

Sources

a. NWO Large-Scale Scientific Infrastructures (2018): http://www.onderzoeksfaciliteiten.nl
b. Re3Data Repositories (2020): https://www.re3data.org/search?query=&countries[]=NLD
c. MERIL Research infrastructures (2019): https://portal.meril.eu/meril/
d. ESFRI Research Infrastructures (2018): http://roadmap2018.esfri.eu/media/1049/roadmap18-part3.pdf
e. LCRDM Facilities (2017): https://www.lcrdm.nl/en/rdm-in-the-netherlands
f. DANS NARCIS (datasets): https://www.narcis.nl/search/coll/dataset/Language/en
g. DANS NARCIS (publications): https://www.narcis.nl/search/coll/publication/Language/en

For a more extensive report on 2020 see: *Landscaping Country Report: The Netherlands*. EOSC-Synergy 14-6-2020. https://doi.org/10.17026/dans-2by-ereu.

To be continued…

Thank you for your attention