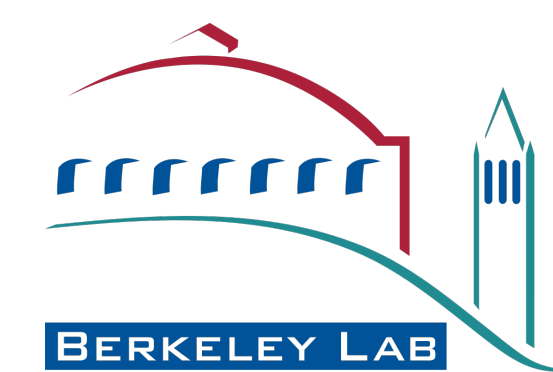


# Evaluating scientific data licensing with the (Re)usable Data Project

Lilly Winfree<sup>1</sup>, Robin Champieux<sup>1</sup>, Letisha Wyatt<sup>1</sup>, Julie McMurry<sup>1</sup>, Melissa Haendel<sup>1,2</sup>, Seth Carbon<sup>3</sup>

<sup>1</sup>Oregon Health & Science University, Portland, OR, USA;

<sup>2</sup>Oregon State University, Corvallis, OR, USA; <sup>3</sup>Lawrence Berkeley National Laboratory, Berkeley, CA, USA



## Reusable Data Project (RDP)

The RDP was created to systematically describe the current data licensing landscape from the perspective of data aggregation, reuse, and redistribution of publicly funded biological & biomedical data resources. We hope to engage the scientific community in discussion regarding the legal use and reuse of scientific data.



Most seemingly “open data” does not provide legal permissions for reuse and redistribution.



Technical, logistical, and legal barriers impede data reusability and interoperability, and are a significant burden and expense.



Being unable to integrate and redistribute our collective data resources blocks innovation and development of life-improving diagnostic & drug selection tools.



To help biomedical research & research support communities understand & navigate the data licensing landscape, we created a rubric to determine the reusability of data resources and have applied it to 56 scientific data resources (e.g. databases) to date.

See the full rubric at:

<http://reusabledata.org>

See the preprint at:

<https://doi.org/10.1101/282830>

Community contributions welcome!

<https://github.com/reusabledata/reusabledata>

## Acknowledgements

This work is supported by NIH Grants OT3TR002019, U24TR002306, R24OD011883, and the U.S. Department of Energy Contract No. DE-AC02-05CH11231. We would like to thank Arvin Paranjpe, Senior Technology Development Manager, OHSU, for his thoughtful discussion.

## Contact

Lilly Winfree: [winfree@ohsu.edu](mailto:winfree@ohsu.edu)

Seth Carbon: [sjcarbon@lbl.gov](mailto:sjcarbon@lbl.gov)

Twitter: @lilscientista @MonarchInit #ReusableData

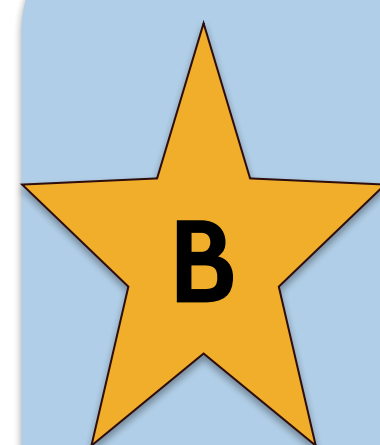
## Star Rubric in Summary

### Public, discoverable, and standard



- Is the license easily findable?
- Is there one, unambiguous license?
- Is the license standard?

### Complete scope



- Are terms of reuse clearly defined without need for negotiation with data creators/curators?
- Is all data covered by the license?

### Easily accessible data



- Is the data in a reasonable place?
- Is the data easily accessible in bulk?

### Few restrictions on *kind* of reuse



- Are all types of reuse (copying, editing, building upon, remixing, distributing) allowable?

### Few restrictions on *who* can reuse



- Can any type of user group reuse the data?

## Resources can score from 0 to 5 stars



**5-star example:** The license unambiguously allows reuse and redistribution of the data.



**4-star example:** The license uses custom terms that allow reuse & redistribution of data but only for certain groups.



**3-star example:** The license is clearly stated & standard, & has clear data access, but has terms that greatly limit reuse & redistribution.



**2-star example:** The license is custom, does not apply to all the data, and has terms that greatly limit reuse and redistribution.

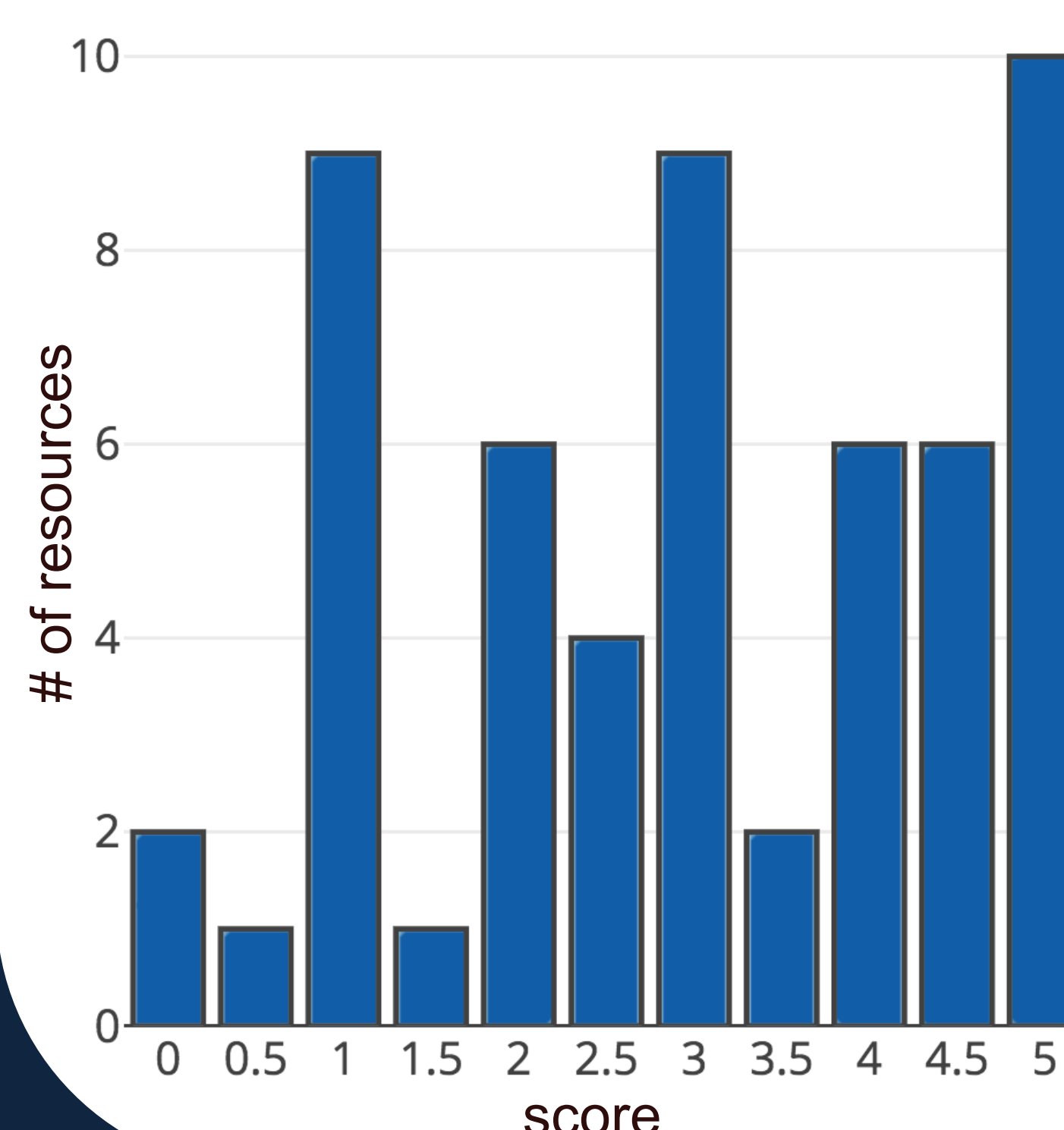


**1-star example:** No license is found, but data is accessible.



**0-star example:** No license is found, and data is not accessible.

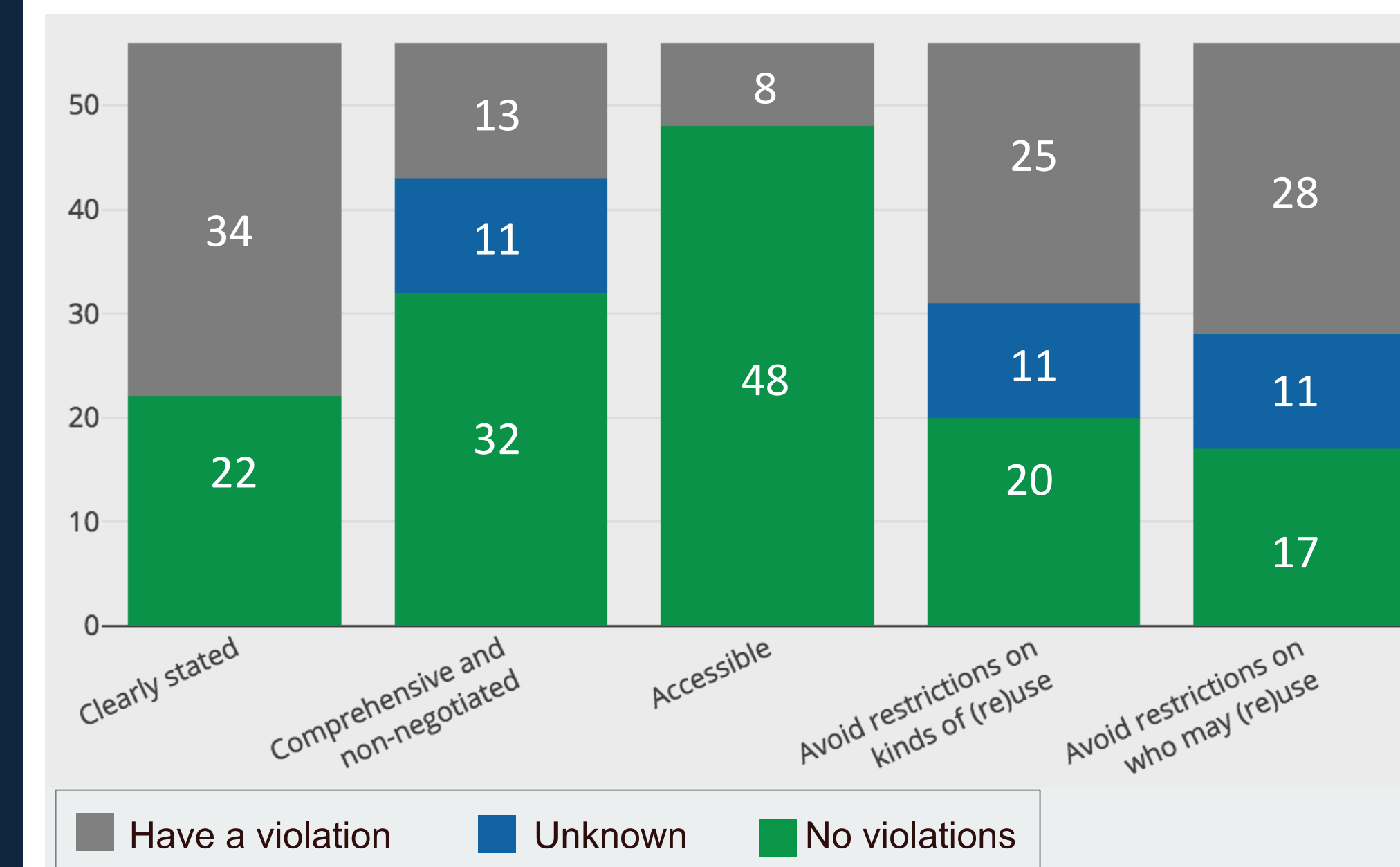
## 41% of resources scored less than 3 stars



**Star score distribution:** 3 sources scored below 1 star, 10 scored between 1-1.5 stars, 10 resources scored 2-2.5 stars, 11 resources scored 3-3.5 stars, 12 resources scored 4-4.5 stars, and 10 resources scored 5 stars.

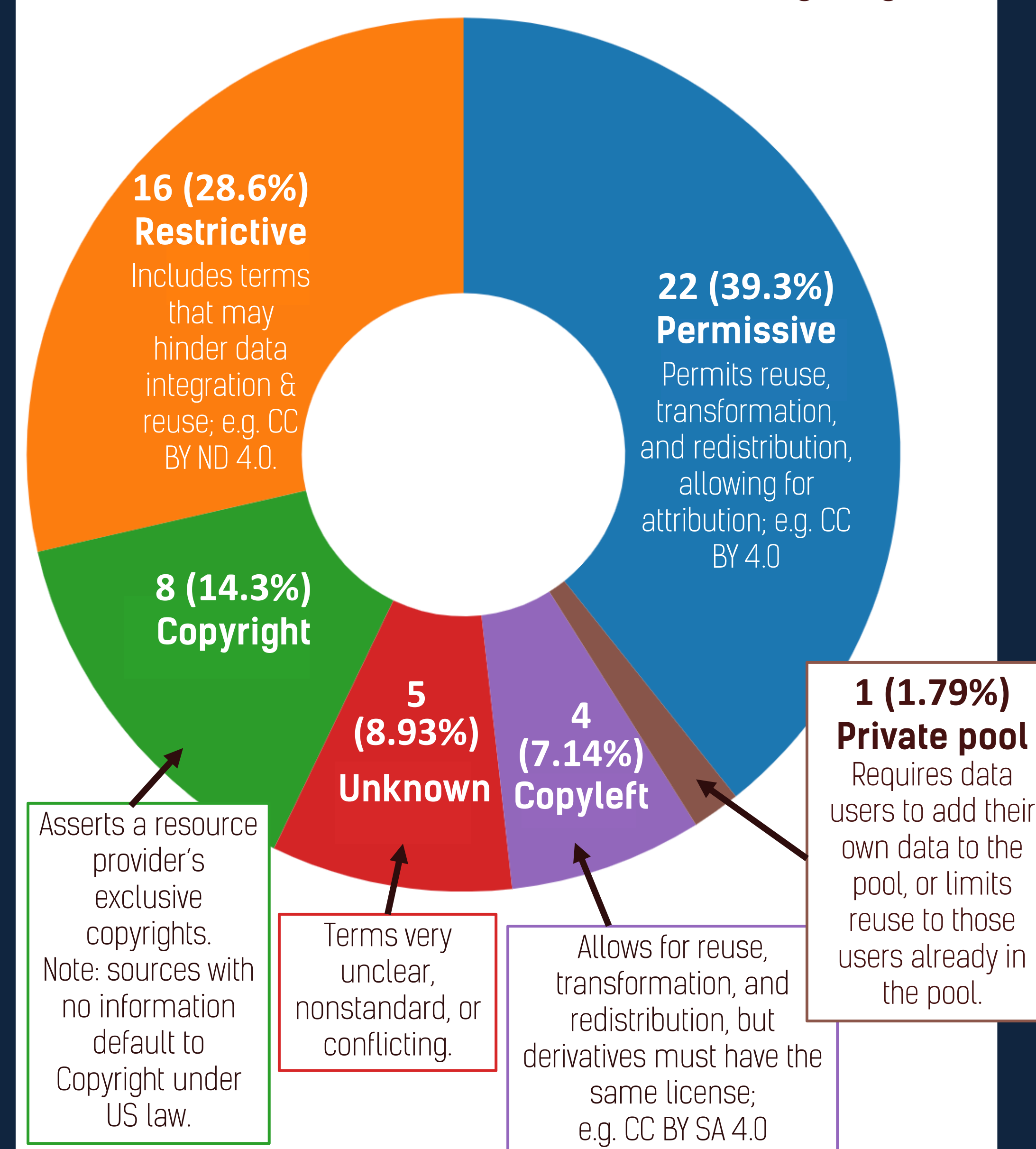
## Findings

### Most resources scored poorly over the 5 categories

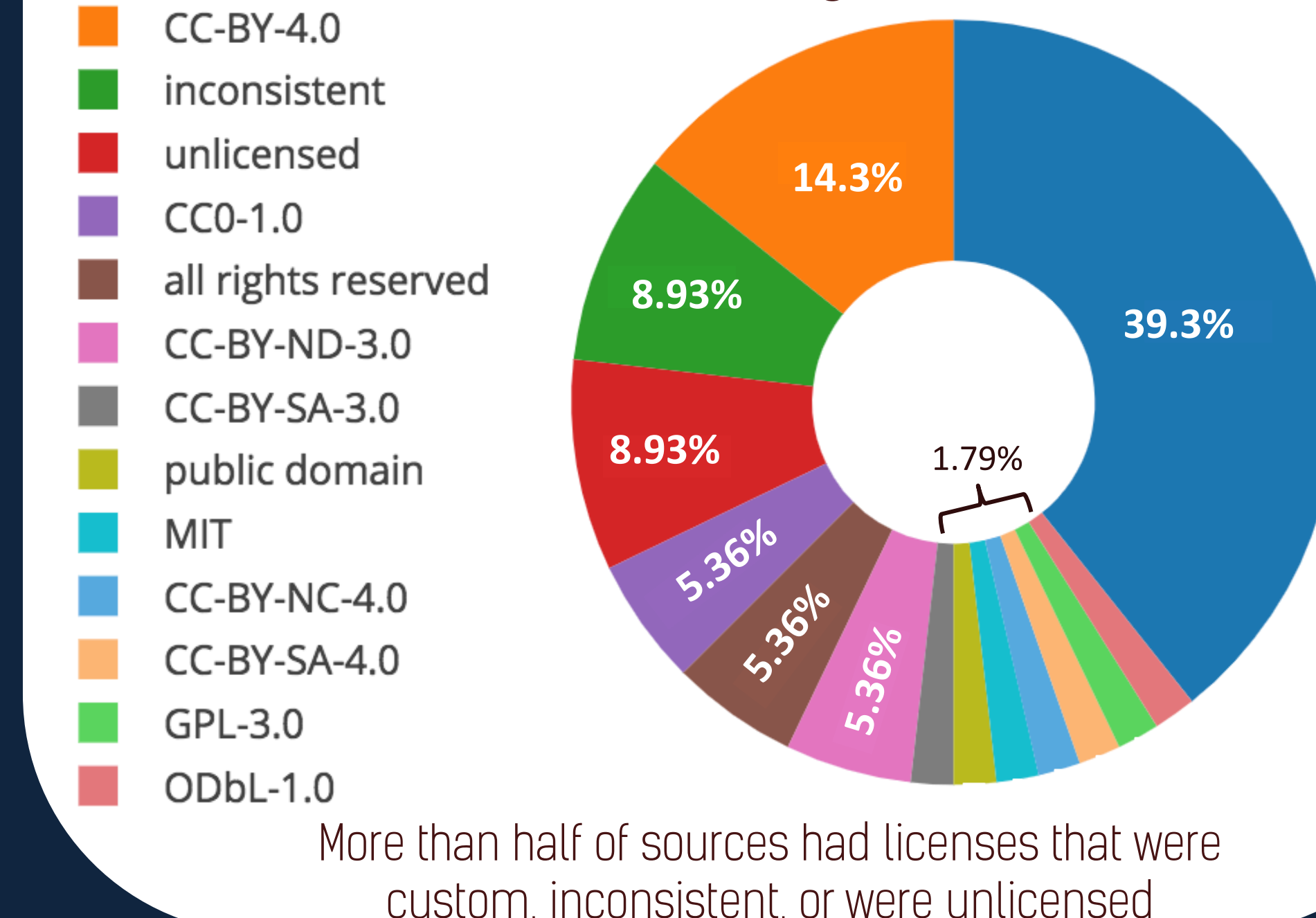


**Summary of curated data resources:** 39% of sources have a license that is found & clearly stated. 57% have a license that covers all the data without negotiation. 86% have data that is easily accessible. 36% have no limits on type of reuse. 30% have no limits on who can reuse.

### Overall, non-permissive licenses are the largest group



### Custom licensing dominates



More than half of sources had licenses that were custom, inconsistent, or were unlicensed

## Conclusions

**Data is often accessible, but frequently cannot be reused or redistributed.** While 48 (86%) of the resources we evaluated provided easy & actionable data access, only 10 (18%) received a full 5-star rating & 32 (57%) resources received 3 stars or less, indicating that there were serious barriers to reuse.