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

Nicole Vasilevsky[1], Seth Carbon[2], Robin Champieux[3], Julie A. McMurry[1,5], Lilly Winfree[1], Letisha R. Wyatt[4], Melissa A. Haendel[1,5]

1. Oregon Clinical & Translational Research Institute, Oregon Health & Science University, Portland, Oregon, USA
2. Berkeley Bioinformatics Open-source Projects, Lawrence Berkeley National Laboratory, Berkeley, California, USA
3. OHSU Library, Oregon Health & Science University, Portland, Oregon, USA
4. Neurology, Oregon Health & Science University, Portland, Oregon, USA
5. Linus Pauling Institute, Oregon State University, Corvallis, Oregon, USA

ABSTRACT

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.

Technical, logistical, and legal barriers impede data reusability and interoperability, and are 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.

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

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.

RUBRIC AT A GLANCE

Public, discoverable, & 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 the data easily accessible in bulk?

Easily accessible data
- Is all data covered by the license?
- 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?

EXAMPLE SCORES

★★★★★ The license unambiguously allows reuse and redistribution of the data.

★★★★☆ The license uses custom terms that allow reuse & redistribution of data but only for certain groups.

★★★☆☆ The license is custom, does not apply to all the data, and has terms that greatly limit reuse and redistribution.
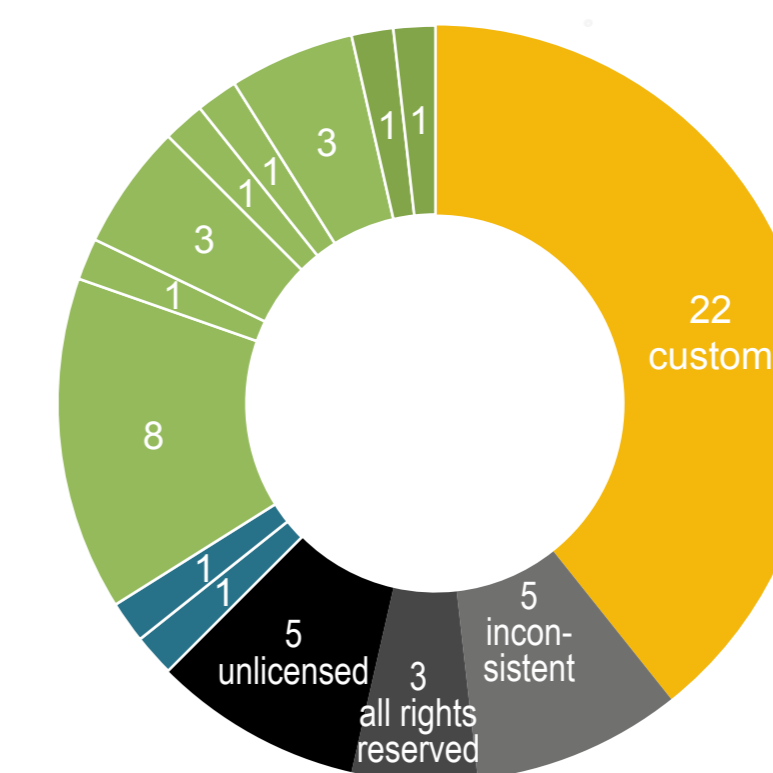
★★☆☆☆ The license is clearly stated & standard, & has clear data access, but has terms that greatly limit reuse & redistribution.

★☆☆☆☆ No license is found, but data is accessible.

☆☆☆☆☆ No license is found, and data is not accessible.

SCORE OVERVIEW

★★★☆☆ **56 sources**

3 out of 5 stars

Stars	Percent
5 star	18%
4 star	21%
3 star	20%
2 star	18%
1 star	18%
0 star	5%

SCORING BY CATEGORY

Legend: Have a violation / Unknown / No violations

Categories: Clearly stated, Accessible, Comprehensive & non-negotiated, Avoids restrictions on kinds of (re) use, Avoids restrictions on who may (re) use

LICENSE HETEROGENEITY

License landscape is very diverse

- 22 custom
- 5 unlicensed
- 3 all rights reserved
- 5 inconsistent
- 1 GPL-3.0
- 1 MIT
- 8 CC-BY-4.0
- 1 CC-BY-NC-4.0
- 3 CC-BY-ND-3.0
- 1 CC-BY-SA-3.0
- 1 CC-BY-SA-4.0
- 3 CC0-1.0
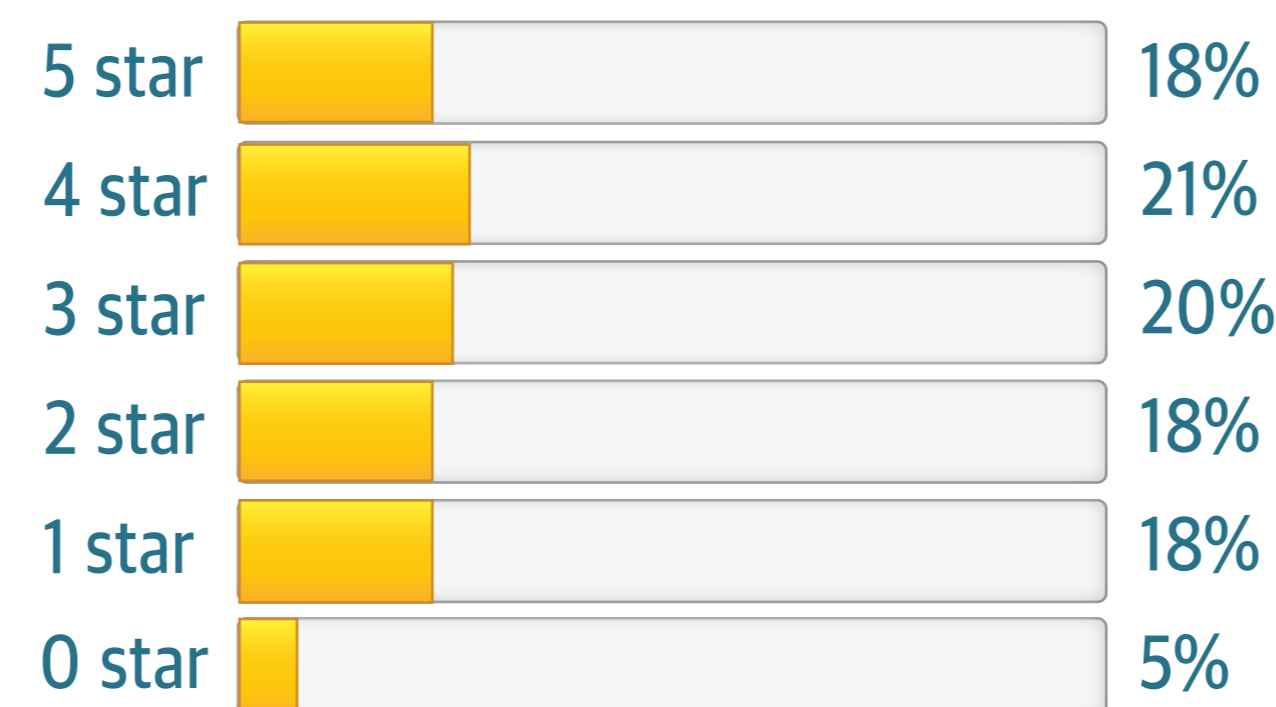- 1 ODbL-1.0
- 1 public domain
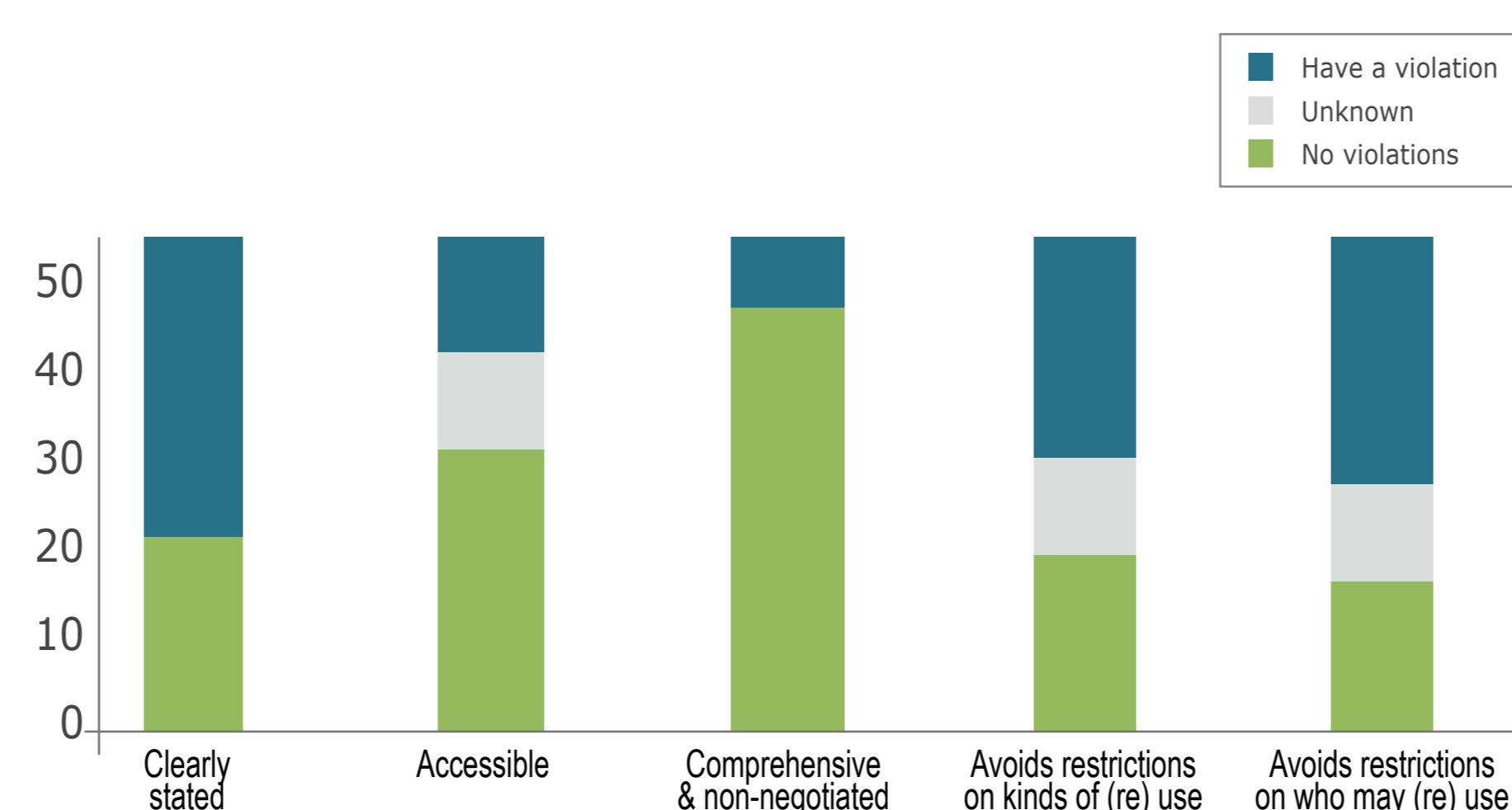
Non-permissive licenses are the largest group
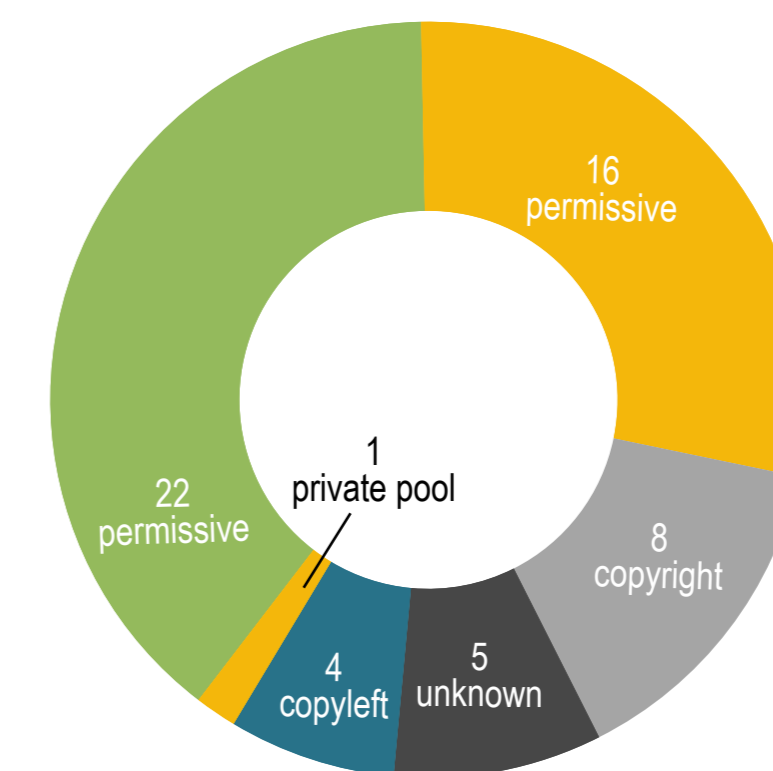
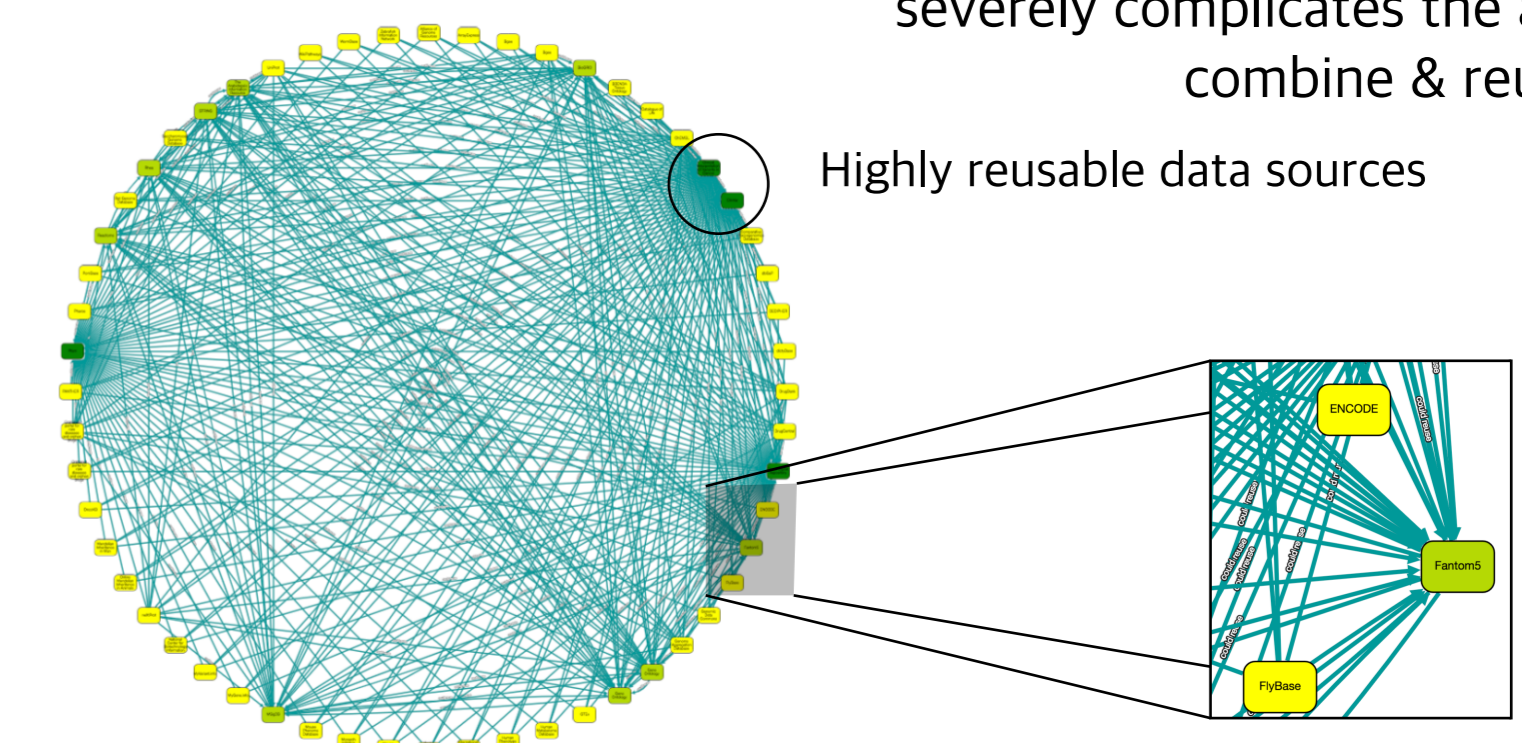- 22 permissive
- 16 restrictive
- 8 copyright
- 5 unknown
- 4 copyleft
- 1 private pool

IMPLICATIONS

This heterogeneity and restrictiveness severely complicates the ability to combine & reuse data.

Highly reusable data sources

ACKNOWLEDGEMENTS & INFO

✉ info@monarchinitiative.org

🐦 @monarchinit

✔ 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