2021 RCD CM Community Data report

Technical Report: RCDNexus-TR-2022.3, May 13, 2022 [DOI: 10.5281/zenodo.6502962](https://doi.org/10.5281/zenodo.6502962)

Prepared by Patrick Schmitz, with input and review by the Research Computing and Data Capabilities Model Working Group members: Claire Mizumoto, Dana Brunson, Doug Jennewein, Galen Collier, John Hicks, Scotty Strachan, and Tom Cheatham.

Table of Contents:

Executive Summary

Research is increasingly dependent upon Cyberinfrastructure (CI), from instruments and sensors to Research Computing and Data (RCD) infrastructure and services. RCD is being used in new domains and is expanding beyond High Performance Computing (HPC) into secure computing; big data management; AI/machine learning; and into heterogeneous compute models, edge computing, and cloud-based computing. The rapid evolution and diversification of RCD poses significant challenges to academic institutions as they try to effectively assess and plan for the necessary resources required to keep pace with the growing needs of researchers. Many would also like to assess their capabilities in comparison to peers. The Research Computing and Data Capabilities Model (RCD CM) [3] allows organizations to self-evaluate across a range of RCD services and capabilities for supporting research, leveraging a shared vocabulary to describe RCD support. The Model supports a range of stakeholders and provides structured input to guide strategic planning and enable benchmarking relative to peer institutions.

Forty-one institutions completed assessments using the RCD CM and contributed these to the 2020 Community Dataset [4]. Twenty-four institutions completed an assessment in 2021, however 14 of these were repeat/updated assessments from 2020, and so the 2020/2021 community dataset includes a total of 51 institutional assessments, representing 32 states and U.S. Territories (and two Canadian provinces). They are a mix of public and private institutions, and while the majority are R1 institutions (Doctoral Universities with “Very high research activity”), the dataset includes a number of R2 institutions and other Carnegie Classifications [1].

The Capabilities Model presents roughly 150 capabilities (in the form of questions) structured around the five *Facings* that are increasingly used as a means of characterizing the roles of people who support RCD: *Researcher-Facing, Data-Facing, Software-Facing, System-Facing,* and *Strategy and Policy-Facing*[1]. The Assessment Tool also allows institutions to mark specific capabilities as *priorities*. The resulting dataset provides important insights into the state of support for RCD at both a summary and granular level. The Dataset also clearly shows the different levels of RCD support among certain sub-communities.

In many cases, the patterns in the data confirm common perceptions about RCD support across the community, particularly about relative levels of support (and gaps) among sub-segments of the community. While these conclusions may be unsurprising to some, it is important to provide quantitative data so we have a baseline for understanding RCD support broadly and in sub-communities. In several areas, the data made clear that differences among certain groups are even more profound than many may have expected. This type of insight may allow RCD leadership and others to refine their understanding of which areas merit the most attention. In a few instances, the data indicate patterns that may not conform to expectations (e.g., in some comparisons of smaller institutions to the

[1] For more on the Facings, see https://carcc.org/facings.

set of R2 institutions). It may be that the relatively small sample size (there were only five smaller institutions) produced an anomalous result, or that other factors were involved. It should be noted that few results presented herein are statistically significant, due both to the sample size as well as the wide variation among institutions. Nevertheless, the main patterns described resonate with many members of the community who have reviewed the results.

Among the significant themes that emerged in our analysis of the data are the following (these are consistent with patterns seen in the 2020 dataset):

- There is wide variation in support levels, and in areas of stronger and weaker coverage, across institutions.
- There is generally stronger support for *Researcher*, *System*, and *Strategy and Policy-Facing* areas, than for *Data* and *Software-Facing* capabilities.
- Private institutions have higher levels of coverage than public institutions, although the difference varies by Facing.
- R1 institutions have much higher levels of coverage than other Carnegie Classifications, particularly in *Researcher*, *Data*, and *Strategy and Policy-Facing* areas.
- EPSCoR institutions[2] have significant gaps in capabilities coverage relative to institutions in other states, including dramatic gaps in certain areas of *Data-Facing* support.

The 2020 Community Dataset provided an initial snapshot of RCD support, and this combined 2020/2021 dataset adds additional institutions and another year of data. While many of the core patterns this year are consistent with the 2020 findings, we are beginning to see some longitudinal changes, especially among institutions that repeated (updated) their assessment in 2021. Some notable changes were seen in the 2021 data, compared to 2020:

- Overall, capabilities coverage values were up slightly in 2021 compared to 2020, across all groups[3]. However, coverage values for the 14 institutions that repeated their assessments increased by an average of 12%, with the largest gains among EPSCoR institutions (18%) and minority-serving institutions (23%).
- A group of EPSCoR institutions has been using the RCD CM as the basis for understanding shared gaps and challenges in their community, as part of strategic planning efforts. Among the five EPSCoR institutions that repeated (updated) their assessments, average *Strategy and Policy-Facing* coverage rose from 44% to 60% – a remarkable 35% increase in one year.

In addition to the capabilities assessment data, the aggregated priorities data provides insight into which areas institutions plan to emphasize, devote resources, etc. Priorities were spread widely across the capability areas, although compared to 2020, priorities in this dataset show a considerable shift from *Strategy and Policy-Facing* topics to other areas, mostly in *Researcher-Facing* areas. Several themes emerged in our analysis of the priorities data:

- The top 25 priorities for all institutions are heavily dominated by topics in the *Researcher-Facing* and *Data-Facing* areas.
- The top 25 priorities for private institutions focus largely on *System-Facing* and *Researcher-Facing* topics, and while priorities for public institutions are more in *Researcher-Facing* and *Data-Facing* areas, they are distributed somewhat more evenly across the Facings.
- The top 25 priorities for EPSCoR institutions focus in large part on *Data-Facing, Researcher-Facing*, and *Software-Facing* topics, while priorities for non-EPSCoR institutions are more in *Researcher-Facing* and *System-Facing* areas, with *none* in *Software-Facing* topics.

[2] An EPSCoR jurisdiction is defined as a state, U.S. territory or U.S. commonwealth that receives less than or equal to 0.75 percent of NSF research funding. The program mission states: “EPSCoR enhances research competitiveness of targeted jurisdictions. . .by strengthening STEM capacity and capability.” See, e.g., https://www.nsf.gov/od/oia/programs/epscor/

[3] Compared using the new 2021 weighting model for both years; see also section 1.3 for details.

1. Introduction

This report describes the second Research Computing and Data Capabilities Model Community Dataset, aggregating the assessments of 51 higher education institutions. Thirty-one of these assessments were completed using version 1.0 of the RCD CM over a period of several months in the Spring and Summer of 2020, with the remainder completed using version 1.1 in the first 9 months of 2021[4]. These data provide insight into the current state of support for RCD across the community and in a number of key sub-communities. The report is intended to be useful to:

- The Higher Education research community (including campus leadership, funding agencies, and others) who are interested in RCD support for research;
- RCD program leaders who are considering the use of the RCD CM Assessment Tool for their strategic planning work; and
- RCD leadership at institutions who contributed to the 2020/2021 Community Dataset and would like additional context on the individualized benchmarking reports they receive for their institutions.

The RCD Capabilities Model is described in [3]. The rest of this paper presents the structure of the Community Dataset; visualization and analysis of significant patterns and themes in the capabilities coverage data; and an exploration of the priorities identified by institutions. We close with conclusions and future work.

More detailed visualizations of the capabilities coverage and priority data are provided in Appendices A and B, which are restricted to community contributors (institutions that completed a Capabilities Model assessment and contributed the resulting data to this effort).

1.1. The Research Computing and Data Capabilities Model

The Research Computing and Data Capabilities Model allows institutions to assess their support for computationally- and data-intensive research, to identify potential areas for improvement, and to understand how the broader community views Research Computing and Data support. The Model was developed by a diverse group of institutions with a range of support models, in a collaboration among Internet2, CaRCC, and EDUCAUSE. This Assessment Tool is designed for use by a range of roles at each institution, from front-line support through campus leadership, and is intended to be inclusive across small and large, and public and private institutions. The RCD Capabilities Model is described more fully in [3].

The 1.0 version became publicly available in January 2020, and an updated 1.1 version was released in January 2021. The assessment tool has been downloaded by over 154 unique institutions across 48 US states, the District of Columbia and two U.S. Territories, four Canadian Provinces, and several other countries. Downloaders include both public and private institutions, a range of Carnegie classifications, and many EPSCoR and minority-serving institutions.

1.2. Structure of the RCD CM Community Dataset

The Community Dataset structure mimics that of the RCD Capabilities Model, and our analyses are presented in terms of the major organizing principles of the Model:

- The primary structure is based upon the *Facings* that organize the various different roles staff and faculty may fill in supporting Research Computing and Data[5].
- Within each Facing, questions are grouped into areas or themes.

[4] See Section 1.3 Notes on Analysis Methodology for a discussion of the differences between versions 1.0 and 1.1
[5] For more on the Facings, see https://carcc.org/facings.

- Each question is considered through several lenses (columns in the assessment tool):
    - *Deployment at Institution* (how broadly available is a topic or service)
    - *Multi-Institutional Collaboration* (whether service providers engaged with community or collaborators)
    - *Service Operating Level* (how robust, resilient, and sustainable is each topic or service)

For each question, an assessment team provides answers for the three lenses, and the answers are combined to produce a numerical **coverage value** for that aspect in the model. The calculated coverage values are combined to produce a summary coverage value for the thematic groupings, and are then aggregated into a coverage value for each Facing. Our analyses compare average coverage values across Facings and down to the theme level, for the community as a whole as well as different sub-groups. This year, we also compared the average values for several of the individual lens (column) values.

A subset of contributing institutions indicated priorities in their assessments (marking topics as either *Medium Priority* or *High Priority).* These do not contribute to the coverage values and are generally used for local strategic planning work; however, we also consider patterns among these values, independent of the rest of the Model.

For institutions that request a copy of the RCD CM Assessment Tool, we pull institutional metadata from the Carnegie Classification dataset [1] including the public/private status, the Carnegie Classification, minority-serving status, and whether they are an EPSCoR institution.

Figures 1 to 5 below illustrate the demographic characteristics of the 51 institutions that contributed assessments to the 2020/2021 Community Dataset. Some points on these contributing institutions include:

- Thirty-two US states and Territories are represented in the data. While we look forward to more complete coverage of US states, we note that these 32 states include nearly 80% of the R1 and R2 Universities in the US (as reported by Carnegie[6]), and so should be fairly representative.
- Roughly 78% of our reporting institutions are public, while 22% are private. Carnegie reports that about 70% (185 of 266) of R1 and R2 institutions are public, and 30% (80 of 266) are private, so the proportions in our data are roughly comparable to the broader US, with public institutions somewhat over-represented.
- About 22% of our reporting institutions are designated as minority-serving. This is slightly higher than the proportion of R1 and R2 institutions that Carnegie lists as minority-serving (42 of 266, or about 16%), and is very close to the roughly 20% of Doctorate and Master's institutions that are minority-serving (based upon NCES data[7]). However, we note our Dataset includes no Historically or Predominantly Black Colleges and Universities (HBCU/PBIs), nor American Indian-serving (Tribal Colleges and Universities/TCU) institutions (we continue efforts to address this gap through targeted outreach and support).
- Just under 30% of our reporting institutions are in EPSCoR states, representing 11 of the 28 EPSCoR jurisdictions. This is a good deal higher than the ~23% (61 out of 266) of R1 and R2 institutions Carnegie reports in these states and Territories[8]. This may in part reflect work related to an NSF-funded workshop[9] to gather EPSCoR institutions and consider their shared challenges, leveraging the RCD Capabilities Model a means of analysis.

[6] The Carnegie Classification of Institutions of Higher Education, https://carnegieclassifications.iu.edu/. Accessed on 11/30/2020.
[7] National Center for Education Statistics, DataLab Tables Library, https://nces.ed.gov/Datalab/TablesLibrary/TableDetails/3995. Accessed on 11/30/2020.
[8] And also higher than the 24% (275 of 1136) of all Doctoral and Masters institutions in the US.
[9] National Science Foundation (OIA) Award 2033483, Award 2033519, and Award 2033514, Collaborative Research: "Building Research Cyberinfrastructure in EPSCoR Jurisdictions: Assessment, Planning and Partnerships."

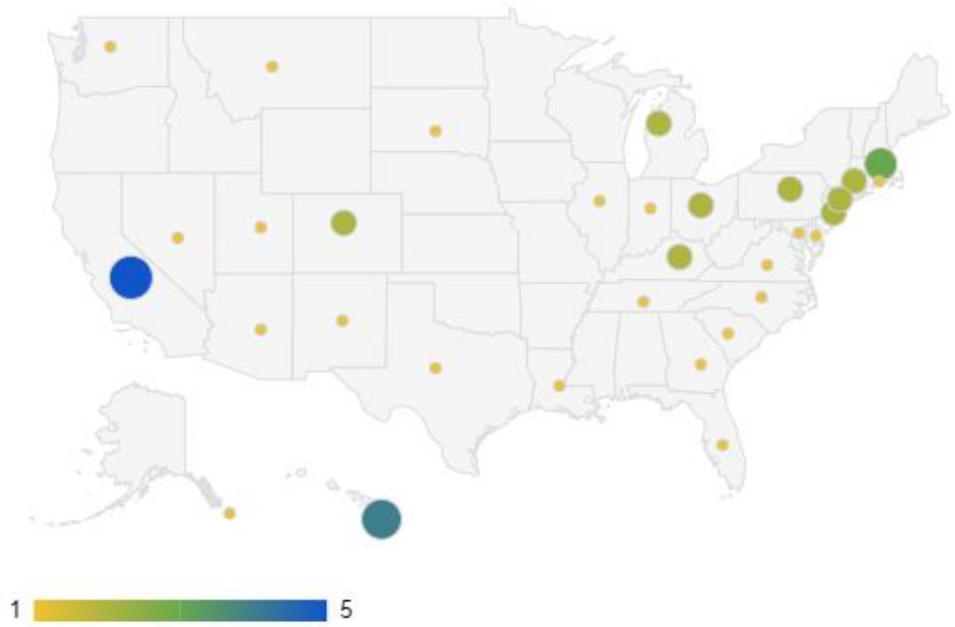

Figure 1: RCD Capabilities Model 2020/2021 Dataset contributing institutions by State

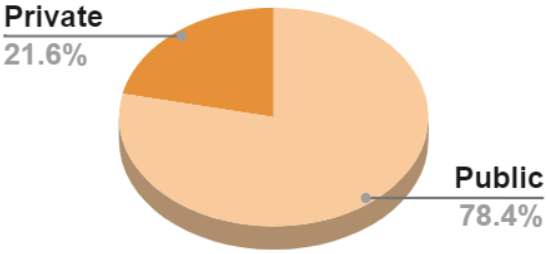

Figure 2: Contributing institutions by type of control

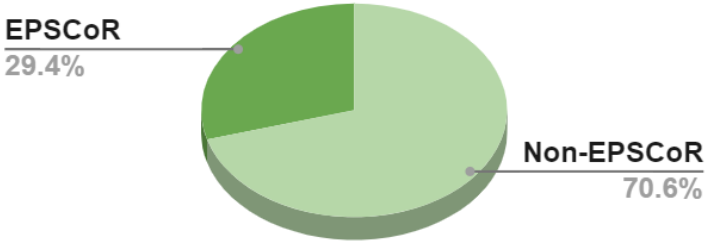

Figure 3: EPSCoR status of contributing institutions

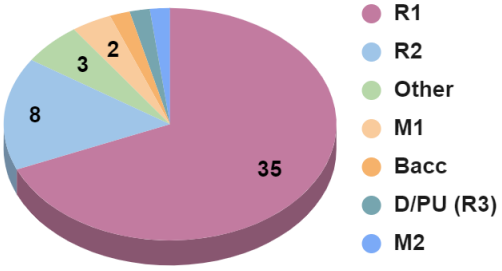

Figure 4: Contributing institutions by Carnegie Classification

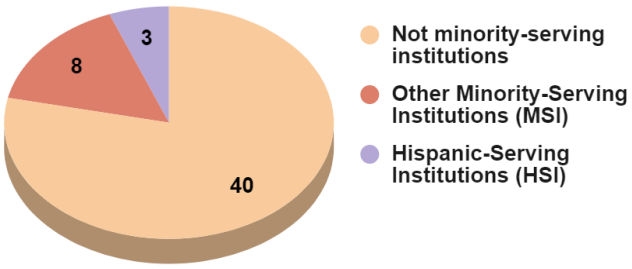

Figure 5: Contributing institutions by Minority-Serving status

1.3. Notes on Analysis Methodology

The 2020 Community Dataset leveraged Google Sheets functionality to aggregate the data with additional tools for graphing, however these tools did not scale well with the additional 24 assessments. For 2021 we created a CSV file of all assessment data, and developed a Jupyter notebook for analysis and graphing of data. It is our hope that for the 2022 assessment, a new data portal (now in development) will be available and will replace at least some of the content in this report with an interactive exploration interface (especially for the appendices).

In response to considerable community feedback, version 1.1 of the tool used by contributors in 2021 included two changes to the calculation of the coverage value for each row/topic in the model:

1. We adjusted weighting of the three main scoring columns in the assessment tool (*Deployment at Institution*, *Multi-Institutional Collaboration*, and *Service Operating Level*) to reduce the impact of the collaboration column, not only to reduce the computed coverage when there was no collaboration at all, but also to give a slight boost for *Leading multi-institutional collaboration*.

2. The version 1.0 "*Local Relevance*" column was removed and incorporated into the choices for *Deployment at Institution*. In version 1.1 of the assessment tool, when a row/topic is marked as "*Not relevant or applicable*" this row was simply ignored in the summary values (e.g., for that theme in the Facing, and for the overall Facing coverage). In the version 1.0 used in 2020, a *Local Relevance* of zero yielded a 100% coverage for that row/topic, which had the effect of slightly boosting the summary computed values (not a reasonable result).

The first change affected nearly all assessments, but the second change affected only a relatively few assessments (those that answered "*Not relevant or applicable*" on more than just a few topics). As such, the net impact of these changes was a slight increase in the summary computed coverage values for most institutions. However, in a few isolated cases, this reduced the summary computed coverage values for an institution (i.e., where the second change had a greater impact).

The RCD Capabilities Model Assessment Tool includes features for assessing the breadth of support across different academic domains, and these are used to compute a *domain-weighted coverage value* at the summary level. The analyses in this report all use the unweighted capabilities coverage values.

For the combined 2020/2021 dataset, we retrofitted the revised weighting to all 2020 assessments, and disregarded all 2020 assessments for which there was a 2021 assessment from the same institution (i.e., we used the newer data for repeating institutions). This resulted in a total of 51 unique contributing institutions.

Where we group by Carnegie classification, we simplified the groupings to R1, R2, and "Other Academic" institutions. The "Other Academic" group includes only five contributing institutions (a mix of R3, Masters, and Baccalaureate institutions), and so must be considered with care, but some interesting patterns emerged in the data (as discussed below). Note that "Other Academic" does not include the Carnegie classification "Other," as the contributing institutions with that classification are dedicated research laboratories or centers; inclusion of them with the R3, masters, etc. institutions would make interpretation of data for a combined group much less meaningful. As there were only three institutions with the "Other" classification, it is unfortunately too small a group for analysis.

The RCD CM Working Group is working with the EPSCoR institutions that contributed to the 2020/2021 Dataset and we have produced an analysis of EPSCoR results compared to the broader community in [7]; we do not duplicate that analysis here.

Note that some caution must be exercised in comparing graphs in this report to those in the 2020 report as the 2021 graphs use the new weighting scheme. Where we mention relative percentages comparing 2020 and 2021, we are using average coverage values computed *with the new weights for both years*.

As an assessment team works through the tool, they may also identify specific aspects as an area of priority in their institutional planning and mark these as either *Medium Priority* or *High Priority* (these do not contribute to the coverage values and are just for local strategic planning work).

We updated the methodology for aggregating priorities this year to account for the fact that some institutions marked only a few priorities and others marked most or all rows as a priority. Each topic could be marked as either a *High Priority* or *Medium Priority*, and at the outset we assigned two priority points for each *High Priority* and one priority point for each *Medium Priority.* However, we allowed up to 10 topics across the assessment to be marked as a priority, and if more than 10 were marked, we scaled the priority points, multiplying them by 10 divided by the number of total topics marked as priorities. We then summed the (scaled) priority points for each question, and

sorted to get the top priorities. For additional background on the evolution of the model for aggregating marked priorities, see [4].

1.3.1. Supporting Institutional Benchmarking

Many of the institutions indicated a strong interest in being able to benchmark their assessments relative to the community and also to demographic slices. Just as for the 2020 Dataset, contributing institutions can request an individual benchmarking report that shows their results in the context of the larger dataset. For more information, see [4].

1.4. Observations and reflections on the analysis

While the 2021 analysis tools were a significant improvement over the 2020 workflow, it remains clear we need an interactive portal that would allow community members to explore the data themselves. Such a portal is currently under development as part of the RCD Nexus project[10], with the 2020 report and this one informing the use-cases we will support.

There are a few outliers in the dataset that lead to fairly wide divergence of mean and median values, at least for certain sections of the data (for more discussion of this, see [4]). We decided to simply use the mean (i.e., average) in this report, but will support tools in the new portal that let users explore where the mean and median diverge (e.g., to identify significant gaps across the community in specific areas of support). In addition, as we gather more data over time, we will observe this phenomenon and consider whether a trimmed mean is appropriate for the analyses.

2. Significant Themes for Capabilities Coverage

2.1. Community-wide patterns

One of the most striking aspects of the 2020 Community Dataset was the significant variation in the data – and this is also true with the combined 2020/2021 Dataset. For the data as a whole, and for many of the subsets in the data (selecting a Facing, a theme, etc. and by different demographic slice), the standard deviation is often a very large proportion of the mean value. In a few cases, error bars in the graphs (at one standard deviation) extend above or below the full range of the axis. The scatter graph in Figure 9 below illustrates the range and variation of assessed RCD capabilities coverage for the institutions represented in the 2020/2021 Dataset. Each vertical stripe represents a given institution (in no particular order) with five colored dots indicating the summary coverage for the five Facings.

Several features are worth noting in the scatter graph visualization:

1. The coverage values are literally all over the map, from very low to very high.
2. Only a few institutions have coverage values that are consistent across Facings. Most have fairly different levels of coverage in each Facing, or at least one Facing for which coverage is quite different.
3. There is little commonality to the relative ranking of Facing coverage across institutions. That is, different institutions have strengths and weaknesses in different areas.

The last point may be of interest to sub-communities, regional groups, and other potential collaborators. The variation in coverage across Facings indicates potential opportunities for collaboration with partners who may have complementary areas of strength, with the potential to share leading practices in different areas.

[10] "Advancing Research Computing and Data: Strategic Tools, Practices, and Professional Development," an NSF Cyberinfrastructure Centers of Excellence (CI CoE) pilot, PI Dana Brunson, OAC-2100003. See also http://rcd-nexus.org/.

Note also that due to the new weighting, the data in figure 9 have shifted up slightly compared to the equivalent figure in the 2020 report (on average, a 3.4% higher value, which is a 6.4% relative increase). This shift and the contribution of specific factors is explored in Appendix C.

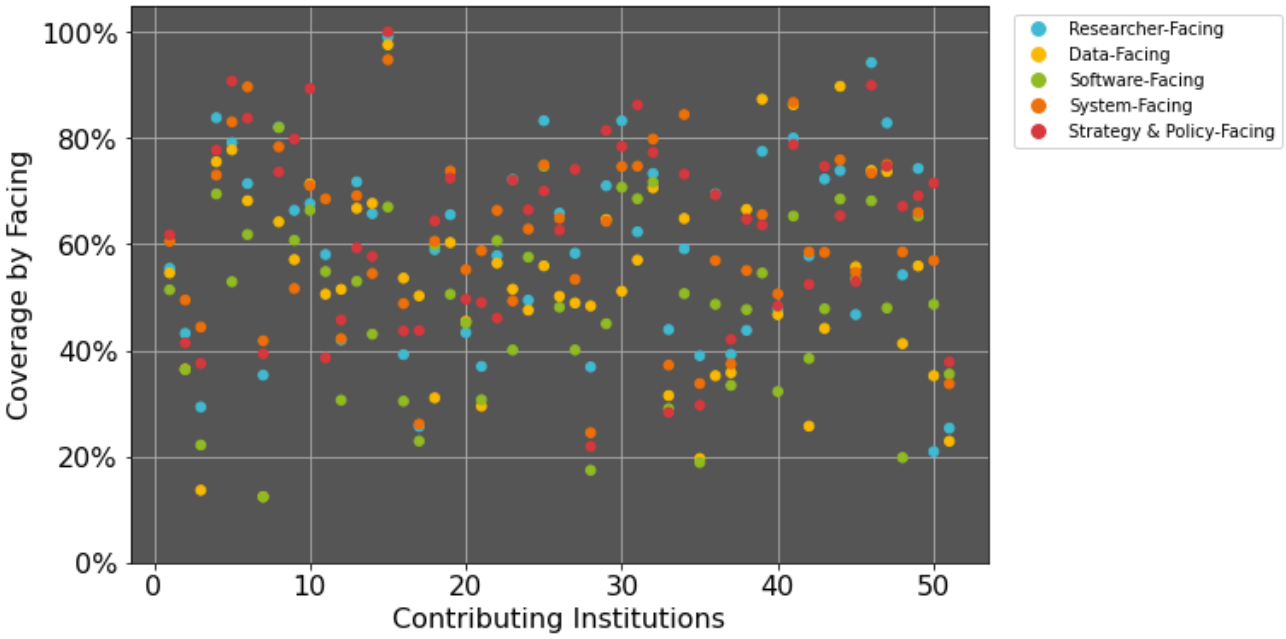

Figure 9: Scatter graph of capabilities coverage by Facing for all 51 institutions

Figure 10 below illustrates the average capabilities coverage for each Facing across all institutions. On average across the community, the broadest coverage is in the Strategy and Policy-Facing, System-Facing, and Researcher-Facing areas, with somewhat less coverage in both Data and Software-Facing areas. The error bars provide another indicator of the considerable variation among institutions. While the variance is slightly smaller for the System-Facing category, all Facings show considerable variation across contributing institutions.

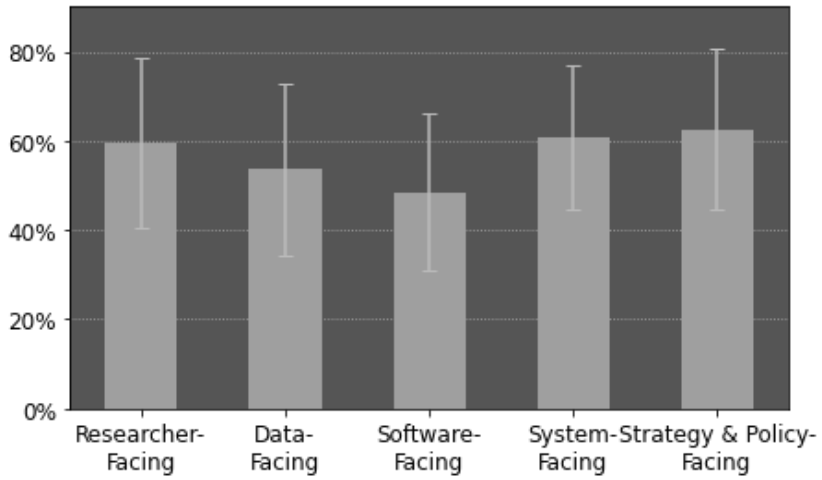

Figure 10 - Capabilities coverage for all institutions, for each Facing

2.2. Demographic commonalities and differences

Figure 11 presents summary capabilities coverage for different demographic slices. Note the differences between public and private institutions, R1 versus other institutions, and between institutions in EPSCoR states as opposed to those in other states. These patterns are consistent with the 2020 Community Dataset, and may conform to

expectations of experienced academic community members. The data clearly support popular conceptions of relative capabilities, and show these differences are often substantial.

In Figure 12 we can see significant variation in certain Facing areas when we filter the data by the Carnegie Classification of the institutions. Institutions are fairly comparable in the System-Facing and Software capabilities, perhaps reflecting the longer history of established good practices for systems definition, administration, and maintenance, as well as software management. However, there is considerable variation in capabilities coverage in Researcher-Facing, Data-Facing, and Strategy and Policy-Facing areas, where roles and good practices have more recently emerged and/or are rapidly expanding and evolving. One somewhat surprising result is the relatively strong coverage of Data-Facing capabilities for the “Other Academic” group (a mix of R3, Masters, and Baccalaureate institutions). This is a small group (n=5), and while three had coverage values below the R2 group average, two institutions report very high coverage values (well above the average for the R1 group), and so it is not meaningful to draw general conclusions.

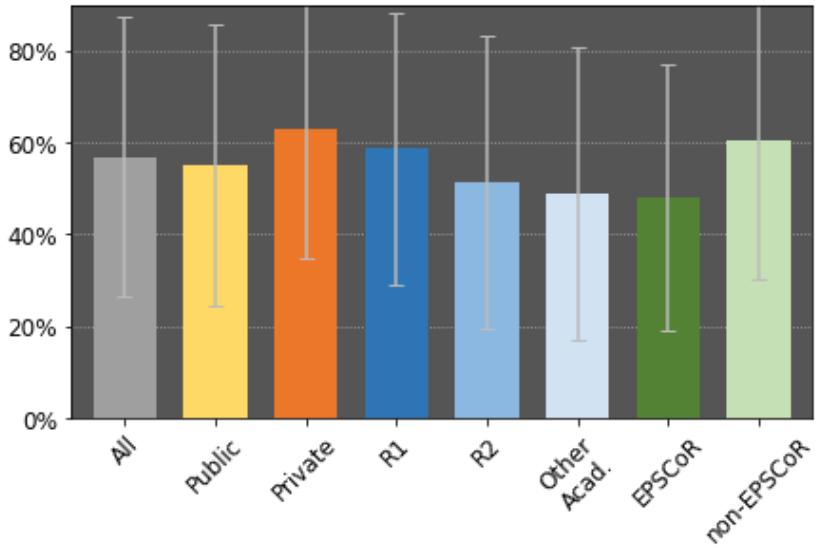

Figure 11 - Total RCD Capabilities coverage by key institutional demographics

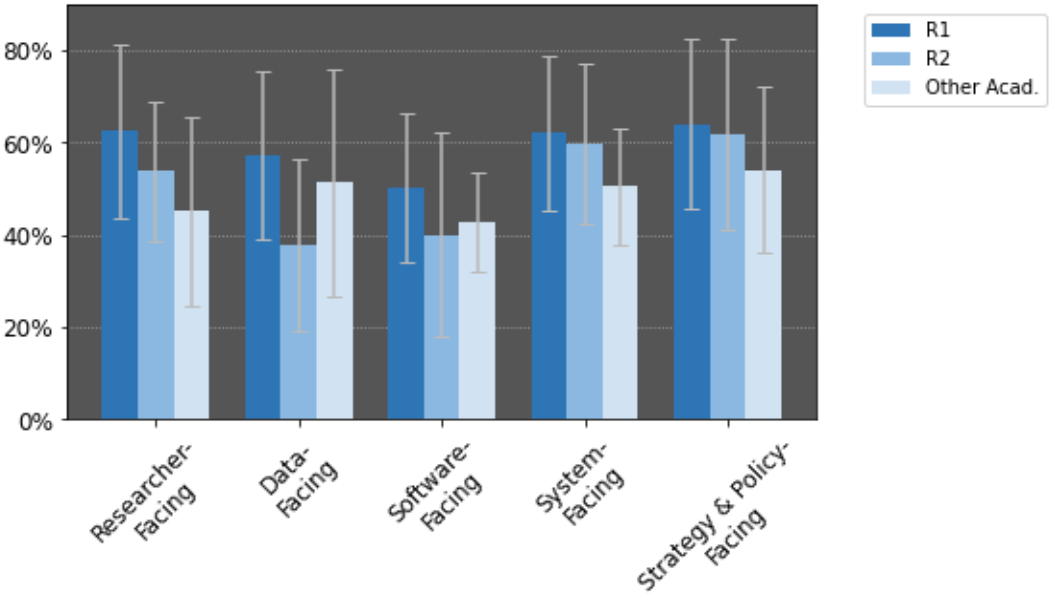

Figure 12 - Average Capabilities coverage across Facings by Carnegie Classification

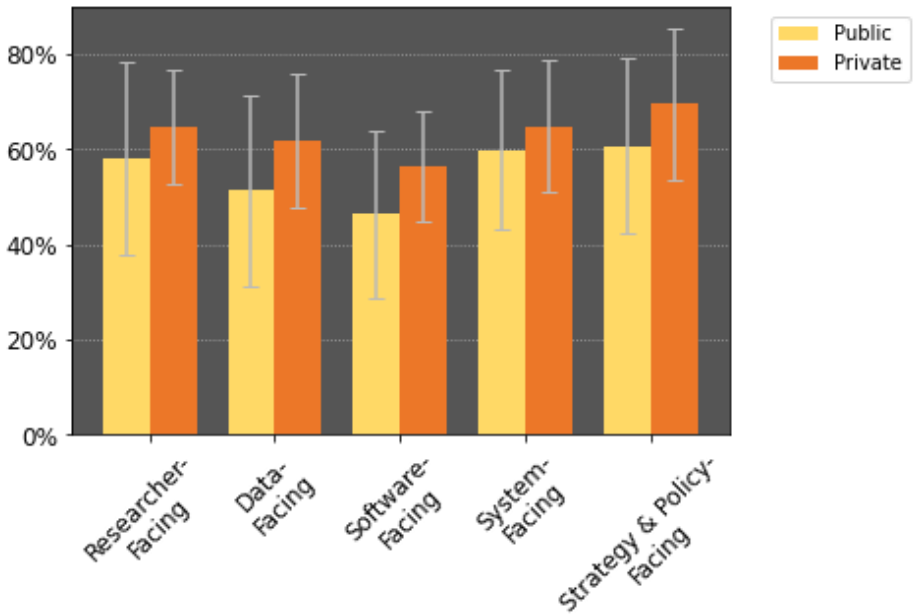

Figure 13 - Capabilities coverage across Facings, public vs. private institutions

Figures 13, 14, and 16 present similar comparisons by Facing, for public versus private institutions, for institutions in EPSCoR states versus other states, and by Minority-serving status. Just as in 2020, the current data show that the private institutions consistently have higher capabilities coverage than the public institutions (again, unsurprising to many observers, but confirmed here by assessment data). It is interesting to note that the *variation* among public institutions is considerably greater than that of the private institutions, except for System-Facing capabilities. Looking deeper into the Facings, certain areas show wide gaps between public and private institutions (detailed graphs are in Appendix A). Points of note include:

- In the 2020 dataset, public institutions reported higher coverage in two of the four Researcher-Facing themes, but in the 2021 dataset, private institutions have higher coverage in all topics, with the widest margin in ***RCD Staffing***.
- In the Data-Facing themes for ***Data Analysis*** and ***Data Visualization***, private institutions average more than 40% higher coverage than public institutions. This is a slightly wider gap than seen in 2020.
- Private institutions average more than 20% higher coverage of capabilities for ***Data Security/Sensitive Data Support***, but this has narrowed from 2020 when the average for private institutions was nearly twice that of public institutions.
- Private institutions reported much higher coverage values than public institutions in the Software-Facing areas of ***Research Software Development***, ***Software Optimization***, ***Workflow Engineering***; and ***Software Portability, Containers, Cloud***. While generally consistent with the 2020 dataset, the gap widened substantially for the first and last topics in this set.
- Public and private institutions are much more comparable in the System-Facing topics. In 2020, there were four topics for which public institutions had higher coverage values than private institutions, but in 2021 there was only one such case.
- In the 2020 dataset, private institutions reported much higher average coverage of capabilities in the Strategy and Policy-Facing themes of ***Institutional Culture for Research Support***; ***Funding***; ***Partnerships / Engagement with External Communities****; and* ***Diversity, Equity, and Inclusion***. While the 2021 dataset again shows private institutions having consistently higher than public institutions in the Strategy and Policy-Facing topics, the gaps have narrowed somewhat.

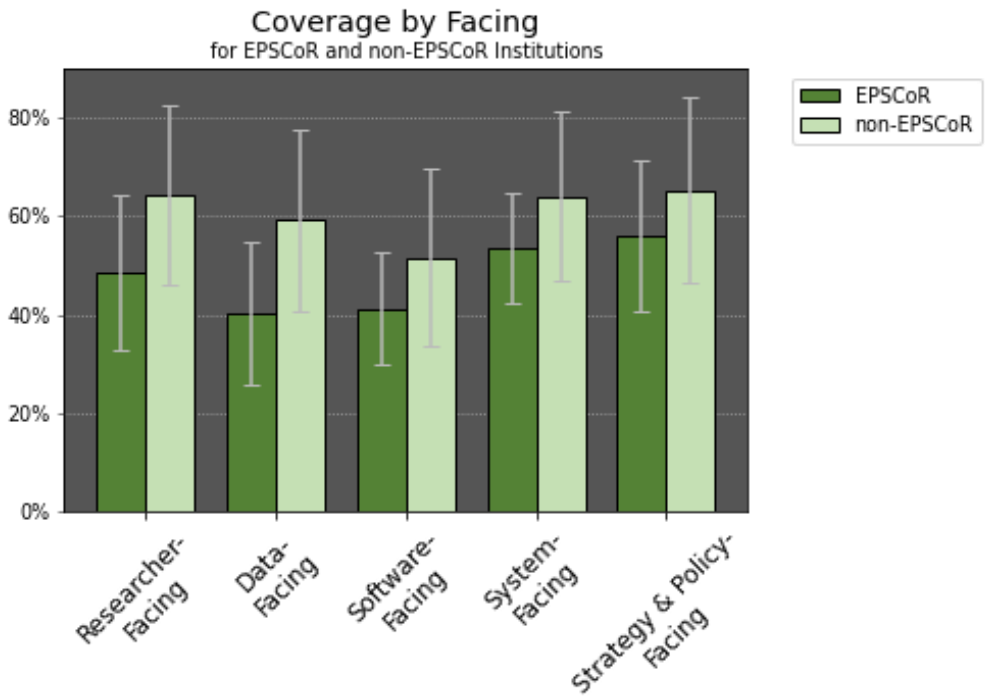

Figure 14 - Average capabilities coverage across Facings by EPSCoR status

A similar pattern emerges for the EPSCoR versus. non-EPSCoR institutions, but with even starker differences. The EPSCoR institutions experience wide gaps in Researcher-Facing and Data-Facing (see Figs. 14 and 15). The more detailed views (in Appendix A) show significant gaps and a few areas of parity[11]; highlights include:

- Gaps exist in the areas of ***RCD Staffing*** and ***RCD Outreach***.
- As in 2020, some of the widest gaps between EPSCoR and non-EPSCoR institutions are seen in the Data-Facing Areas, with much lower average coverage in EPSCoR institutions for: ***Data Discovery and Collection***; ***Data Analysis***; ***Data Visualization***; and particularly, support for ***Security/Sensitive Data***.
- EPSCoR institutions reported significantly lower coverage in the Software-Facing areas of ***Research Software Development***; ***Workflow Engineering***; ***Software Portability, Containers, Cloud***; and ***Securing Access to Software***.
- System-Facing capabilities coverages are broadly lower for EPSCoR institutions, with wide gaps in the themes of ***Compute Infrastructure***, ***Storage Infrastructure***, and ***Documentation***. However, several large gaps seen in the 2020 dataset have narrowed, particularly for ***Storage Infrastructure*** and ***Security Practices for Secure Environments***.
- EPSCoR institutions show much lower assessed coverage values in the areas of ***Institutional Culture for Research Support*** and ***Diversity, Equity, and Inclusion***, but these gaps are somewhat smaller than in 2020.

It is interesting that from 2020 to 2021, the gap between EPSCoR and non-EPSCoR institutions in Strategy and Policy-Facing topics has narrowed. While the non-EPSCoR average rose slightly (1.1 percentage points) from 2020 to 2021, the EPSCoR average rose more sharply, from 50.2% to 57.4%. This is due, in part, to the addition of new EPSCoR institutions (which averaged 52% in Strategy and Policy-Facing coverage). However, it should be noted that for the five EPSCoR institutions that repeated assessments in 2021, the average Strategy and Policy-Facing coverage rose from 44% to 60% – a remarkable 35% increase. Could this year-over-year progress in topics related to strategy and policy be correlated to engagement with the RCD CM and using it in their strategic planning practices? Perhaps these institutions had resolved to make improvements in this area and so were motivated to use the RCD CM as part of that, and perhaps the exercise of conducting the assessment helped them structure their Strategy and Policy-Facing work. In any case, it is a notable data point. As a point of comparison, the average

[11] A more detailed analysis of the patterns among EPSCoR institutions is presented in [7].

Strategy and Policy-Facing coverage for non-EPSCoR institutions that repeated an assessment in 2021 also rose, from 64% to 72%. For all institutions, the average Strategy and Policy-Facing coverage for repeating institutions is far above that of the new 2021 institutions. This is an interesting trend to watch over time.

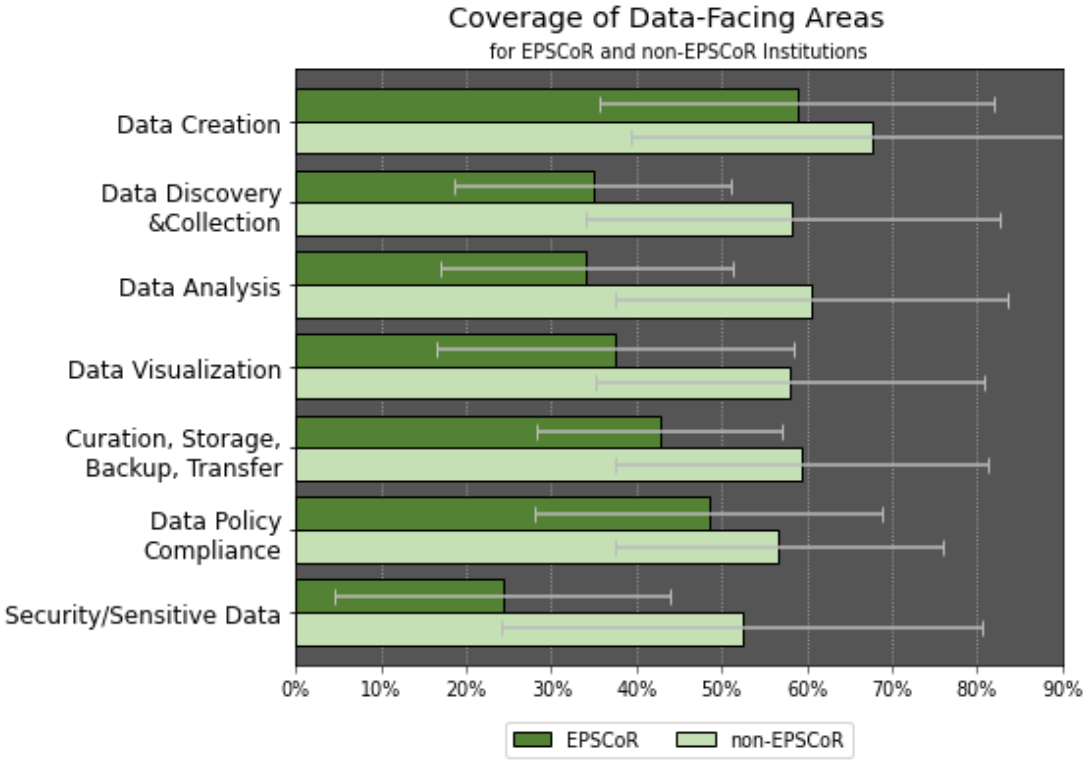

Figure 15 - Average Data-Facing capabilities coverage by EPSCoR status

Minority-serving institutions (Fig. 16) show a similar stark pattern of gaps relative to institutions that are not minority-serving. These gaps exist across the spectrum of Facings, with significant differences in some of the same areas described for the other demographic comparisons, above (especially in Researcher-Facing and Data-Facing topics). In 2020, there was a more pronounced difference between average and median values in the System-Facing and Software-Facing areas, but with the addition of more institutions in 2021 this is no longer the case.

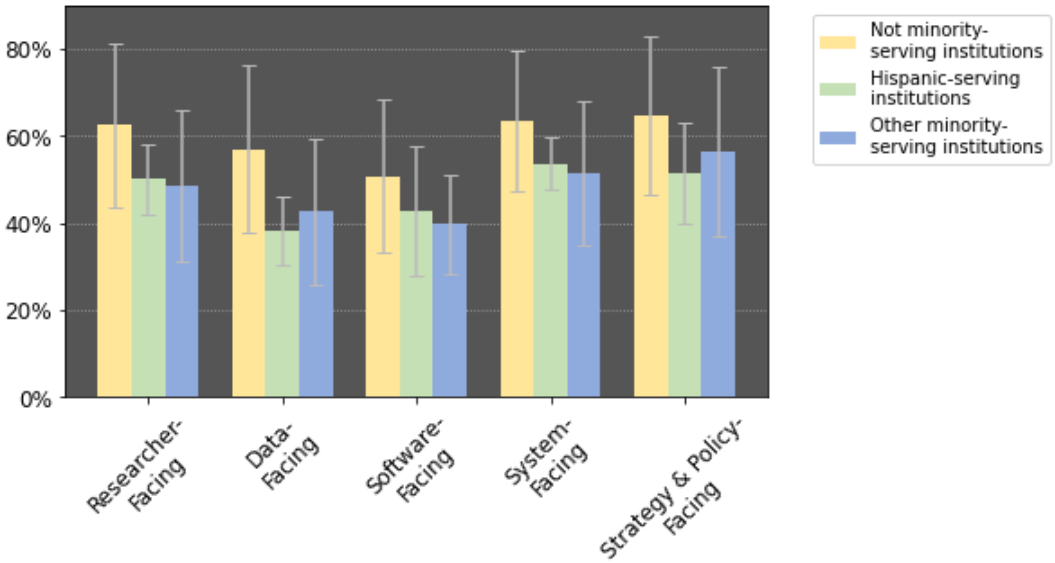

Figure 16 - Average capabilities coverage across Facings by minority-serving status.

3. Significant Themes for Priorities

We are seeing increased use of the assessment tool's priorities feature. In 2020, 22 of the 41 contributing institutions marked priorities (54%). In 2021, 17 of the 24 contributed assessments[12] included priorities (74%). In the combined dataset, 33 of the 51 unique institutions marked priorities (65%). Of note is that just over half of first-time assessments marked priorities, but nearly 80% of those repeating the assessment marked priorities, indicating they are further leveraging the tool as an input to strategic planning. We summed the priority points for each question[13] and sorted to get the top priorities. We present both the overall top list, as well as the top five in each Facing in Tables 1-6 below. A more extensive presentation of priorities data is available (to contributors) in Appendix B.

#	Priority Points	Topic/Question	In Facing
1	5.38	Do researchers have access to introductory user support and training related to the use of research computing and data resources available at local, regional, and national level?	Researcher-Facing
2	4.08	Does your Research Computing and Data (RCD) team/group have a strategic plan? Is this strategic plan updated on a regular basis (e.g., annually, semi-annually)?	Strat. & Pol.-Facing
3	3.39	Do researchers have access to dedicated resources (e.g., staff) who can perform data wrangling/manipulation and data analysis?	Data-Facing
4	3.29	Is there an institutional practice to proactively reach out to researchers, new faculty, or prospective faculty (for example during interviewing or new faculty onboarding processes) to explain research support services and help with computing beyond the desktop?	Researcher-Facing
5	3.16	Can researcher-facing staff effectively serve as advocates for the research community to leadership and IT governance?	Researcher-Facing
6	3.16	Are researchers made aware of research computing and data related resources?	Researcher-Facing
7	3.15	Are Research Computing and Data services funded in a sustainable manner?	Strat. & Pol.-Facing
8	3.13	Do researchers have access to data archival and preservation services (e.g. tape, cloud)?	System-Facing
9	3.08	Do researchers have access to consulting and expertise on data visualization?	Data-Facing
10	3.04	Do researchers have access to compute and data environments to manage and use "notice triggering" data (e.g., PHI, HIPAA, Export Control, licensed data)?	Data-Facing

Table 1 - Overall top priorities for contributing institutions

3.1. Community-wide patterns in the summary data

Several aspects are worth noting in the ranking of priorities across the full community (Table 1). Seven of the top ten are Researcher-Facing and Data-Facing topics, and no Software-Facing topics appear. The top-ten proportions are consistent with the broader set as well: of the top 25 priorities, over two-thirds are in Researcher-Facing and Data-Facing topics, and only two of the top 25 are Software-Facing topics. Only five of these top 10 in 2021 were in the top list in 2020, but the top two were numbers three and one in 2020[14].

The high ranking for strategic planning in both years likely mirrors the engagement of programs with the RCD CM tools, and may reflect the progress made in Strategy and Policy-Facing topics discussed in section 2.3 above. It is

[12] Including those that were repeats of a 2020 assessment.
[13] See section 1.3 for details on how we computed priority points for each question.
[14] These comparisons use the new weighting and ranking for priorities, and so may not comport with the rankings in the 2020 report.

also interesting that many of the top ten are fairly fundamental to engaging with researchers, perhaps reflecting the often expressed need to balance technical infrastructure with trained RCD professionals [5]. The last three topics in the top ten were ranked much lower in 2020, and together with number three seem to indicate the growing importance of challenges managing data.

Table 2 shows the top five Researcher-Facing priorities for the whole community. All but the last are in the top ten overall, and the fifth is ranked number 11 overall. All five are in the themes of ***RCD Staffing*** and ***RCD Outreach*** (as are nine of the top ten!), underscoring the emphasis being placed on core researcher engagement and support, (although the fifth is also closely related to strategic planning).

Prio Pts	Researcher-Facing Topic/Question
5.38	Do researchers have access to introductory user support and training related to the use of research computing and data resources available at local, regional, and national level?
3.29	Is there an institutional practice to proactively reach out to researchers, new faculty, or prospective faculty (for example during interviewing or new faculty onboarding processes) to explain research support services and help with computing beyond the desktop?
3.16	Can researcher-facing staff effectively serve as advocates for the research community to leadership and IT governance?
3.16	Are researchers made aware of research computing and data related resources?
2.96	To what extent is there a clear vision, effective guidance, and strategy for the allocation and prioritization of support resources/personnel?

Table 2 - Top five Researcher-Facing priorities for contributing institutions

Table 3 shows the top 5 Data-Facing priorities for the whole community. The first three in the top ten overall, and the other two are ranked numbers 13 and 14 overall, consistent with the broad emphasis placed on data-related issues. However, it is interesting that the topics cover a broad range of themes within the Data-Facing area.

Prio Pts	Data-Facing Topic/Question
3.39	Do researchers have access to dedicated resources (e.g., staff) who can perform data wrangling/manipulation and data analysis?
3.08	Do researchers have access to consulting and expertise on data visualization?
3.04	Do researchers have access to compute and data environments to manage and use “notice triggering” data?
2.88	Do researchers have access to consulting and expertise on data wrangling/manipulation and data analysis?
2.80	Does your institution have research data governance processes in place to establish data policies for research data?

Table 3 - Top five Data-Facing priorities for contributing institutions

Prio Pts	Software-Facing Topic/Question
2.79	Do researchers have access to support for research software package compilation and installation?
2.48	Do researchers have access to support, facilitation or training on how to compile, install, and deploy research software (e.g. The Carpentries, documentation on how to install and deploy anaconda environment, etc.)?
2.34	Do researchers have access to resources (e.g., staff) who can develop software for wide usage?
2.13	Do researchers have access to resources (e.g., staff) who can develop research software?
1.80	Are processes defined and adhered to for educating, monitoring, and auditing Research Computing and Data staff, other IT professionals, and researchers, to comply with software license agreements?

Table 4 - Top five Software-Facing priorities for contributing institutions

Table 4 lists the top five Software-Facing priorities for the whole community, but only two of these are in the top 25 overall. It is worth noting who within the community marked Software-Facing priorities; as Figure 17 illustrates, the weight among Software-Facing priorities is dominated by EPSCoR and minority-serving institutions, many of whom are well behind in RCD capabilities compared to their counterparts. The Software-Facing priorities in the top five list may reflect the need to establish or solidify support that is a “given” in more established programs. For both private institutions and the group of non-minority-serving institutions, no Software-Facing topics appear in their respective top 25 priorities.

Table 5 lists the top five System-Facing priorities for the whole community, of which only the top one appears in the top ten overall, and only three are in the top 25 overall. However, private institutions actually placed the greatest weight of priorities in this area, and for non-EPSCoR institutions this was second only to Researcher-Facing in total priorities weight (see Fig. 17). It is also interesting that, of the System-Facing topics, many are related to non-traditional capabilities (e.g., cloud computing and interactive computing), and four of the next five (i.e., ranked sixth through tenth) are related to data management (see Appendix B).

Prio Pts	System-Facing Topic/Question
3.13	Do researchers have access to data archival and preservation services (e.g. tape, cloud)?
2.91	Do researchers have access to interactive computing services? E.g., support for VDI, Gateways, JupyterHub.
2.65	Are there institutional resources for leveraging commercial cloud services for research computing and researchers?
2.32	Are deployment, operations, and maintenance of your infrastructure automated (e.g. puppet, ansible, chef)?
2.20	Do researchers have access to high throughput computing (HTC)?

Table 5 - Top five System-Facing priorities for contributing institutions

Table 6 lists the top five Strategy and Policy-Facing priorities for the whole community, of which the top two appear in the top ten overall, and three are in the top 25 overall. The emphasis on strategic planning, strategic alignment, and funding echo themes in the 2020 dataset. Notably, private institutions, EPSCoR institutions, and minority-serving institutions placed much less weight here than the average overall.

Prio Pts	Strategy and Policy-Facing Topic/Question
4.08	Does your Research Computing and Data (RCD) team/group have a strategic plan? Is this strategic plan updated on a regular basis (e.g., annually, semi-annually)?
3.15	Are Research Computing and Data services funded in a sustainable manner?
2.36	Is there an understanding across the IT organization, research community, and institutional leadership of the distinction between Research Computing and Data services and standard (enterprise) IT services?
2.20	Is your Research Computing and Data (RCD) strategic plan aligned to campus plans?
1.58	Are researchers effectively informed and made aware of Research Computing and Data (RCD) resources and services?

Table 6 - Top five Strategy and Policy-Facing priorities for EPSCoR institutions

3.2. Distribution of priorities by demographic grouping

Figure 17 presents a distribution of the summed weight (i.e., the sum of the weighted priority points, and not just the count) of priorities across the five Facings. There are some remarkable outliers in this figure, particularly in the Software-Facing and System-Facing areas, where some groups put much more emphasis than their counterparts. We considered whether the numbers were somewhat skewed by the relatively small number of institutions in each group, but there were eleven minority-serving institutions in the dataset, and they had a median of eight priorities marked in their assessments, while other groups with outlier patterns had even larger institution counts. It is also

notable that while nine of the eleven minority-serving institutions are also EPSCoR institutions, the patterns of priorities for these two groups diverge in a number of cases.

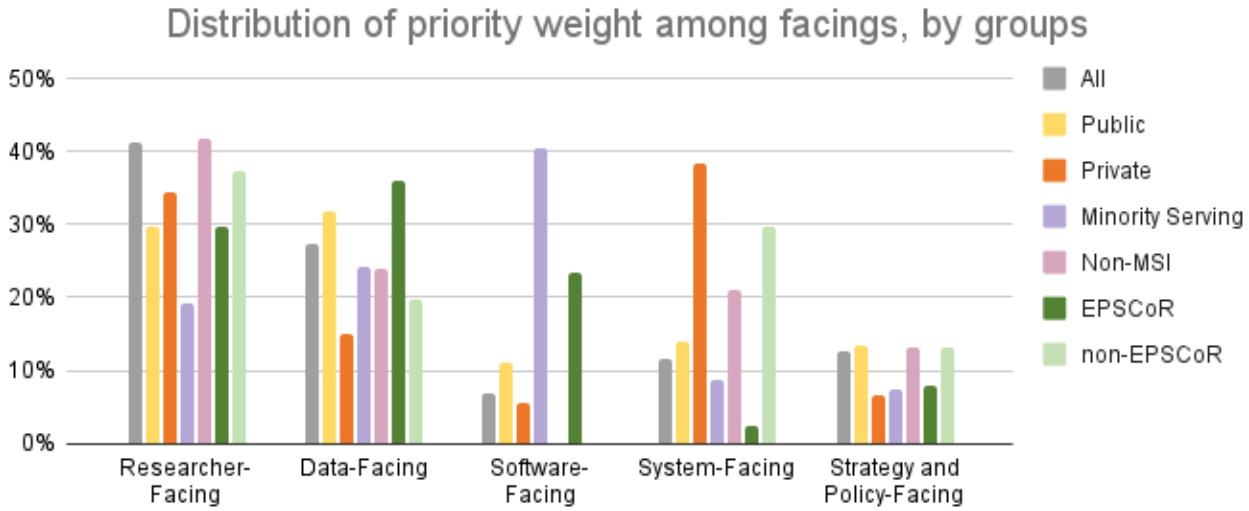

Figure 17 - Distribution of priority weight among Facings, by community sub-groups

The top priorities for R2 institutions diverged quite a bit from the overall set. Their top priority was the understanding across leadership of the distinction between RCD services and standard (enterprise) IT services, and the next five were all related to data management, but across the top 25, considerable emphasis (~40% of priority points) was placed on System-Facing topics. Among the other academic institutions (i.e., other than R1s and R2s), no Strategy and Policy-Facing topics were in the top 25 priorities, and nearly half of the priority points were on Researcher-Facing topics.

4. Conclusions and Looking Ahead

We have presented an analysis of the second Community Dataset that aggregates 51 institutional assessments using the Research Computing and Data Capabilities Model, including 14 repeated assessments from 2020. The 2020/2021 Community Dataset represents a diverse set of institutions and provides significant insights into the state of support programs for RCD at both the full community scope, as well as within different subsets of the community. The dataset provides an important complement to the model itself, allowing institutions to understand their relationship to the broader community, and providing various entities (including, e.g., funders) with the data to characterize RCD at a fine-grained level, and over time, to follow trends and track the impact of programs designed to advance RCD support.

The patterns in the combined 2020/2021 dataset are largely consistent with the initial 2020 dataset, however we are already seeing some changes and shifts. Some of the changes are likely due to the addition of new contributors, while others can clearly be attributed to changes over time. A particular example of this is the increase in coverage values in the Strategy and Policy-Facing topics among institutions that contributed an assessment in 2020, and then repeated the assessment exercise in 2021.

There are too few contributors, and the variation in assessed coverage is too great, to make statistically significant claims across much of the data. Nevertheless, the RCD Capabilities Model assessment data is the first source of structured information about the state of RCD support in the academic community, and provides the best view we have to date about where institutions excel, where they face challenges, and where they intend to put resources to sustain and improve their services.

4.1. Value to institutions

The CaRCC working group that coordinates this work also facilitates a series of outreach and support activities to assist the community in learning about the model and completing assessments, as well as training in strategic planning practices that can leverage the RCD CM as a key input (see, e.g., [2], [6]). Feedback from the community indicates growing interest in the Capabilities Model and access to the Community Dataset. The vast majority of those who have downloaded the model planned to use it for strategic planning and benchmarking work, as illustrated in Figure 18.

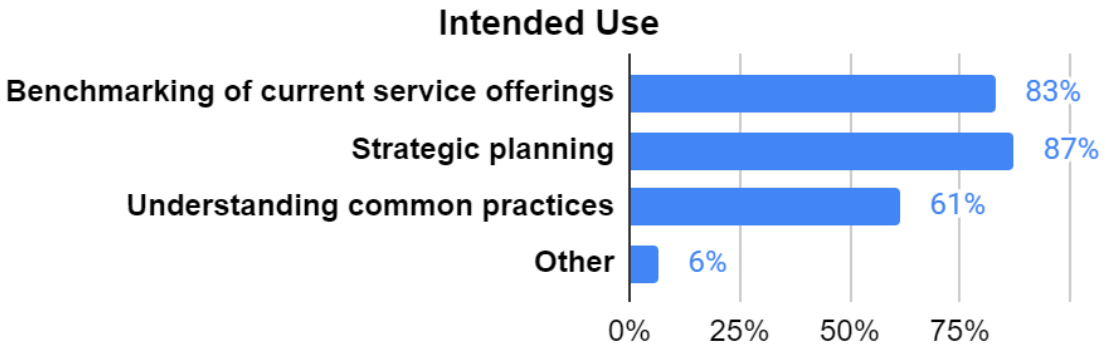

Figure 18 - Intended use of the RCD CM Assessment Tool for all institutions requesting a copy, as of April 2022

We asked institutions how they planned to use the data when they requested their 2021 benchmarking reports. All of them (100%) said it would be used as input to strategic planning, 78% said it would be used to justify campus funding in their program, and well over a third (36%) said it would be used to support a grant proposal. Respondents indicated a variety of uses for the raw data, from benchmarking individual topics to performing further analysis for additional insights.

We also asked how access to the detailed report and data was likely to impact their decision to complete a future assessment. One in four said it would make them just as likely to do so, and the rest (75%) said it made them *more likely* to complete a future assessment, clearly indicating the perceived value in the data as well as in longitudinal analysis. An even higher number (81%) indicated that the benchmarking reports made them more likely to complete a future assessment[15].

Institutions reported that they needed an average of just over 32 person-hours to complete the assessment, with a median value of 25.5 hours. This represents a considerable investment of resources, and indicates the value they perceive in the assessment and access to the Community Dataset. Many institutions reported that they got significant value just in holding discussions with various campus stakeholders as they worked through the assessment.

4.2. The need for an Essentials version

Despite widespread enthusiasm for and increasing usage of the model, we also heard from many smaller and emerging programs that the current Capabilities Model was somewhat daunting, and/or they lacked available resources to complete the assessment (many of these smaller programs have teams of just a few people). In response to this feedback, the CaRCC working group is developing an *Essentials* version of the RCD CM. Focus groups in the Spring of 2022 will be followed by a workshop in the Summer to explore approaches to the Essentials version, and to develop an initial implementation for community review.

[15] No respondents (0%) to either question indicated it would make them *less likely* to complete a future assessment.

We defined a set of principles to guide the design of an Essentials version. Preliminary discussion concluded that the RCD CM Essentials version should:

- Be much shorter and simpler than the RCD CM full version.
- Closely involve candidate institutions in the definition and refinement of the Essentials version.
- Be a stepping stone to completing the full RCD CM. That is, it should not be completely orthogonal, and should provide a fairly natural basis for understanding the full model when and if an institution is ready to consider a richer set of capabilities.
- (Ideally) produce data that is in some manner comparable to the current community datasets.

The last principle reflects our desire to continue to build the community dataset and provide benchmarking support to users of the Essentials version, ideally leveraging data from the full version contributions. More information on the work to develop an Essentials version will be provided on the working group site (https://carcc.org/rcdcm/).

4.3. Refining the Assessment Tool and the data analysis platform

The RCD Nexus project[16] includes a workstream to develop a data exploration portal for RCD CM datasets, which will allow users much more flexibility in exploring the data, configuring benchmarking analyses, etc. We will continue to provide annual commentaries on the significant trends we see, but will refer readers to the portal for the data visualizations and further details. Just as with this report, public access will provide a high-level view of the data while contributors will have access to detailed data and fine-grained filtering tools. We hope this access policy will incent much broader participation in the assessment exercise and many more contributions to the community dataset. We are currently exploring whether the assessment itself can also migrate from the current spreadsheet implementation to a web application hosted on the RCD Nexus portal.

[16] "Advancing Research Computing and Data: Strategic Tools, Practices, and Professional Development," an NSF Cyberinfrastructure Centers of Excellence (CI CoE) pilot, PI Dana Brunson, OAC-2100003. See also http://rcd-nexus.org/.

5. Acknowledgements

This work is supported in part by National Science Foundation by an NSF RCN grant (OAC-1620695, PI: Alex Feltus, “RCN: Advancing Research and Education through a national network of campus research computing infrastructures – The CaRCC Consortium”), and by an NSF Cyberinfrastructure Centers of Excellence (CI CoE) pilot award (OAC-2100003, PI Dana Brunson, “Advancing Research Computing and Data: Strategic Tools, Practices, and Professional Development”). The EPSCoR related work was supported by NSF (OIA) Award 2033483, Award 2033519, and Award 2033514, Collaborative Research: “Building Research Cyberinfrastructure in EPSCoR Jurisdictions: Assessment, Planning and Partnerships.” Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

The Model, the assessment tool, and other associated resources were developed with the generous contributions of time and expertise from the 2018 workshop participants, and the working group members: Alex Feltus, Ana Hunsinger, Cathy Chaplin, Claire Mizumoto, Dana Brunson, Deborah Dent, Doug Jennewein, Gail Krovitz, Galen Collier, Jackie Milhans, James Deaton, Jen Leasure, Jill Gemmill, Jim Bottum, Joe Breen, Joel Cutcher-Gershenfeld, John Hicks, John Moore, Karen Wetzel, Mike Erickson, Patrick Schmitz, Preston Smith, Timothy Middelkoop, and Thomas Cheatham. In addition, individuals at a number of Universities provided valuable feedback on early versions of the model and assessment tool, and continue to provide feedback on aspects of the model and the assessment tool through office hours, etc., for which the working group is very grateful.

6. References

[1] Indiana University Center for Postsecondary Research. (n.d.). *The Carnegie Classification of Institutions of Higher Education, 2021 edition*. https://carnegieclassifications.iu.edu/

[2] Schmitz, P., Gail Krovitz, Dana Brunson, Thomas Cheatham, Galen Collier, John Hicks, Claire Mizumoto, Karen Wetzel, & Alex Feltus. (2019, July 29). *Leveraging a Research IT Maturity Model for Strategic Decision Making*. Practice and Experience in Advanced Research Computing (PEARC ’19).

[3] Schmitz, P., Mizumoto, C., Hicks, J., Brunson, D., Krovitz, G., Bottum, J. R., Cutcher-Gershenfeld, J., Wetzel, K., & Cheatham, T. (2020). A Research Computing and Data Capabilities Model for Strategic Decision-Making. *Practice and Experience in Advanced Research Computing (PEARC ’20)*. PEARC ’20: Practice and Experience in Advanced Research Computing, Portland, OR, USA. https://doi.org/10.1145/3311790.3396643

[4] Schmitz, P. (2021). Assessing the Landscape of Research Computing and Data Support: The 2020 RCD Capabilities Model Community Dataset. *Practice and Experience in Advanced Research Computing*, 1–8. https://doi.org/10.1145/3437359.3465580

[5] Schmitz, P., Yockel, S., Mizumoto, C., Cheatham, T., & Brunson, D. (2021). Advancing the Workforce That Supports Computationally and Data Intensive Research. Computing in Science Engineering, 23(5), 19–27. https://doi.org/10.1109/MCSE.2021.3098421

[6] Schmitz, P., Brunson, D., Mizumoto, C., Jennewein, D., & Strachan, S. (2022). *Towards a National Best Practices Resource for Research Computing and Data Strategic Planning* (RCDNexus-TR-2022.1). RCD Nexus. https://doi.org/10.5281/zenodo.6394989

[7] Schmitz, P., Bayrd, V., Strachan, S., & Jacobs, G. (2022). *A Baseline of EPSCoR Research CI Capabilities* (RCDNexus-TR-2022.2). RCD Nexus. https://doi.org/10.5281/zenodo.6395204

Appendix A: Detailed Graphs by Demographics

Detailed data for each of the facings is only available to contributing institutions; it is not included in the public report.

Appendix B: Extended Priorities data

Detailed priorities data is only available to contributing institutions; it is not included in the public report.

Appendix C: Impacts of reweighting the columns in 2021

As noted in the report, version 1.1 of the tool used by contributors in 2021 included two changes to the calculation of coverage value for each row/topic in the model:

1. We adjusted weighting of the three main scoring columns in the assessment tool (*Deployment at Institution*, *Multi-Institutional Collaboration*, and *Service Operating Level*) to reduce the impact of the collaboration column, not only to reduce the computed coverage when there was no collaboration at all, but also to give a slight boost for *Leading multi-institutional collaboration*.

2. The version 1.0 "*Local Relevance*" column was removed, and this was incorporated into the choices for *Deployment at Institution*. In the version 1.1 version of the assessment tool, when a row/topic is marked as "*Not relevant or applicable*" this row was simply ignored in the summary values (e.g., for that theme in the facing, and for the overall facing coverage). In the version 1.0 used in 2020, a *Local Relevance* of zero yielded a 100% coverage for that row/topic, which had the effect of slightly boosting the summary computed values (not a reasonable result).

The first change affected nearly all assessments, but the second change affected only a relatively few assessments (those that answered "*Not relevant or applicable*" on more than just a few topics). As such, the net impact of these changes was a slight increase in the summary computed coverage values for most institutions. However, in a few isolated cases, this reduced the summary computed coverage values for an institution (i.e., where the second change had a greater impact).

In comparing Figure 9 to the equivalent figure in the 2020 report, the data has shifted up slightly (on average, a 3.4% higher value, which is a 6.4% relative increase) due to the new weighting. Several points related to the reweighting are worth noting here:

- The reweighting resulted in a larger increase in average values for R1 institutions than for smaller programs (6.9% vs 4.6% respectively).

- A similar but smaller effect was seen for non-EPSCoR institutions vs. EPSCoR institutions (6.7% vs 5.2% respectively).

- A greater disparity was seen between institutions not designated as minority serving, compared to minority-serving institutions (7.1% vs. 2.3%).

We considered that in the 2020 data these discrepancies might be due to greater use of the "Local Relevance" column (i.e., marking something as less or not at all relevant) by non-R1s, EPSCoR, and minority-serving institutions (since this can reduce the coverage values with the reweighting). We did see that non-R1 institutions accounted for 51.9% of all uses of *Local Relevance* (as not fully relevant) but are only 23% of contributing institutions and this could account for the discrepancy. However, the EPSCoR and minority-serving proportions of *Local Relevance* use were both proportionately low, and so this is unlikely to explain the discrepancies:

- 20.4% of *Local Relevance* use was by EPSCoR institutions, which represent 26% of the contributing academic institutions.

- 14.2% of *Local Relevance* use was by minority-serving institutions, which represent 21% of the contributing academic institutions.

We then considered whether the reweighting of the column "*Multi-Institutional Collaboration*" might have had a disproportionate effect, since a low value in this column had a stronger negative impact on the computed coverage in 2020 than with the new weighting.

- The average *Multi-Institutional Collaboration* values for R1s is 10% higher than for non-R1s, which should have worked to narrow the gap with the reweighting, and so may have moderated what would have been a larger gap due to the disproportionate use of *Local Relevance* (as discussed above).

- The average *Multi-Institutional Collaboration* values for EPSCoR institutions was more than 20% higher than for non-EPSCoR institutions (2.07 versus 1.71), and the average *Multi-Institutional Collaboration* values for minority-serving institutions was 10% higher than for non-minority-serving institutions (2.34 versus 2.13). These differences were likely a factor in the disparate impact of reweighting.