2020 RCD CM Community Data report

Prepared by Patrick Schmitz (Semper Cogito Consulting), with input and review by the Research Computing and Data Capabilities Model Working Group members: Claire Mizumoto, Dana Brunson, Doug Jennewein, Galen Collier, Joel Cutcher-Gershenfeld, John Hicks, and Thomas Cheatham

Table of Contents:

Executive Summary 1
1. Introduction 2
2. Significant Themes for Capabilities Coverage 12
3. Significant Themes for Priorities 18
4. Conclusions and Looking Ahead 22

Appendix A: Graphs by Demographics
Appendix B: Detailed Graphs by Demographics (only available to contributing institutions)
Appendix C: Priorities by Facing and Area
Appendix D: Complete Priorities Detail (only available to contributing institutions)
Appendix E: Top Priorities by Demographics

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) 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. These institutions represent 28 states, a mix of public and private, R1, R2, and other Carnegie Classifications, and also include a number of key institutional demographic sub-communities. 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*. 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 as well as a 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 support across the community and 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 that we have a baseline for understanding RCD support broadly and in sub-communities. In several cases the data made clear that differences among certain groups are even more profound than many may have expected and this allows RCD leadership and others to refine their understanding of which particular areas of RCD support merit attention.

Among the significant themes that emerged in our analysis of the data are:

- 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 this 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.
- Institutions in EPSCoR-eligible states have significant gaps in capabilities coverage relative to institutions in other states, including dramatic gaps in certain areas of *Data Facing* support.

In addition to the capabilities assessment data, the aggregated priorities data provides insight into the areas in which institutions plan to place emphasis, devote resources, etc. This will provide additional information for RCD leadership (among others) as they develop strategic plans. Priorities were spread widely across the capability areas, however several themes emerged in our analysis of the priorities data:

- The community as a whole marked many priorities in the *Researcher Facing* and *Data Facing* areas, and a fair number are in the *Strategy and Policy Facing* area, with strong emphasis on **staffing**, **research lifecycle management**, and various aspects of **research data management**.
- Private institutions have more priorities in the *System Facing* area and less in the other facings, as compared to Public institutions.
- Minority-serving institutions strongly emphasize the pattern of the broad community, with a significantly higher emphasis in the *Data Facing* and *Researcher Facing* areas, and very little emphasis in the *System Facing* topics.

The 2020 Community Dataset provides an initial snapshot of RCD support. Over time, longitudinal data will provide additional insight into trends, and a means of evaluating the impact of programs designed to increase RCD support, for institutions, for collaborations and sub-communities, and for the community as a whole.

1. Introduction

This report describes the first Research Computing and Data Capabilities Model Community Dataset, aggregating the assessments of 41 Higher Education Institutions. These assessments were completed using the 1.0 version of the RCD CM, over a period of several months in the Spring and Summer of 2020. This data provides 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 of use to:

- The Higher Education research community (including campus leadership, funding agencies, and others) who are interested in research support;
- 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 Community Dataset and would like additional context on the individualized benchmarking reports that they can request for their institutions.

The report includes background on the Capabilities Model; the structure of the Community Dataset; a description of and some reflections on our process to gather and analyze the data; visualization and analysis of the significant patterns and themes in the capabilities coverage data; and a description of the priorities identified by institutions. We close with conclusions and future work.

We present high level data visualizations in Appendices A, C, and E, available in the public version of this report. More detailed visualizations of the capabilities coverage and priority data is provided in Appendices B and D, 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

Research is increasingly dependent upon Cyberinfrastructure (CI), from instruments and sensors to Research Computing and Data[1] (RCD) infrastructure and services. Previously limited to physical sciences like chemistry and physics, RCD has moved far beyond the desktop for all research domains. High Performance Computing (HPC) is still an important element, however RCD has expanded well beyond HPC into secure enclaves for data compliance; big data management, analytics, and movement; AI/machine learning; and more recently 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 lack of a shared vocabulary to describe the various aspects of RCD support hinders efforts to discuss and plan coordinated efforts to advance support of, and for, researchers. These challenges are especially acute for smaller and emerging RCD support organizations, which often lack experience supporting RCD and have limited resources to develop an analysis framework for strategic planning.

To address these gaps, a collaborative team within the RCD ecosystem developed a Research Computing and Data Capabilities Model[2] that allows an organization to self-evaluate across a range of RCD services and capabilities for supporting research. The Model is designed to be useful to a diverse set of stakeholders including campus RCD professionals; PIs and research team members; and campus leadership. The Model provides structured input to guide strategic planning leveraging a defined and shared community vocabulary, and it enables benchmarking relative to peer institutions and/or to various segments of the community.

The initial version of the RCD Capabilities Model was developed as a collaboration among the Campus Research Computing Consortium (CaRCC)[3], Internet2[4], and EDUCAUSE[5], with support from the National Science Foundation (NSF OAC-1620695) and from many volunteers who provided input and review from a diverse set of universities (large and small, public and private) and related organizations. The 1.0 version became publically available in January 2020, and has been downloaded by over 120 institutions across 44 states and 2 Canadian Provinces, including both public and private institutions, a range of Carnegie classifications, and many EPSCoR-eligible and minority-serving institutions.

A very high proportion[6] of institutions that requested a copy of the RCD CM Assessment Tool indicated “Benchmarking of current service offerings” as an intended use and a total of 41 institutions completed the assessment and contributed their results to the 2020 Community Dataset. This report describes this initial Community Dataset which is not only a baseline for benchmarking but also provides important insights into the state of support for research computing and data across the community and within specific sectors and regions.

---

[1] “Research Computing and Data” (abbreviated as RCD) includes technology, services, and people supporting the needs of researchers and research, and is intended as a broad, inclusive term covering computing, data, networking, and software. The National Science Foundation (NSF) uses the term “cyberinfrastructure,” and others use “Research IT.”

[2] Patrick Schmitz, Claire Mizumoto, John Hicks, Dana Brunson, Gail Krovitz, James Bottum, Joel Cutcher-Gershenfeld, Karen Wetzel, Thomas Cheatham. 2020. A Research Computing and Data Capabilities Model for Strategic Decision-Making. *In Proceedings of Practice & Experience in Advanced Research Computing (PEARC20). ACM, New York, NY, USA,* https://dl.acm.org/doi/10.1145/3311790.3396643

[3] https://carcc.org/

[4] https://www.internet2.edu/

[5] https://www.educause.edu/

[6] 86% as of November 2020

1.2. Structure of the RCD CM Community Dataset

The Community Dataset structure mimics that of the RCD Capabilities Model to facilitate benchmarking analyses by institutions that completed the RCD CM Assessment Tool. The Model recognizes different roles that staff and faculty fill in supporting Research Computing and Data with names that reflect who or what each role is facing (i.e., focused on), noting that a given individual may fill roles in multiple facings.

1. **Researcher Facing Roles**. Includes research computing and data staffing, outreach, and advanced support, as well as support in the management of the research lifecycle. Example roles include: Research IT User Support, Research Facilitator, CI engineer.
2. **Data Facing Roles**. Includes data creation; data discovery and collection; data analysis and visualization; research data curation, storage, backup, preservation, and transfer; and research data policy compliance. Example roles include: Research Data Management specialist, Data Librarian, Data Scientist
3. **Software Facing Roles**. Includes software package management, research software development, research software optimization or troubleshooting, workflow engineering, containers and cloud computing, securing access to software, and software associated with physical specimens. Example roles include: Research Software Engineer, Research Computing support.
4. **Systems Facing Roles**. Includes infrastructure systems, systems operations, and systems security and compliance. Example roles include: HPC systems engineer, Storage Engineer, Network specialist.
5. **Strategy- and Policy Facing Roles**. Includes institutional alignment, culture for research support, funding, and partnerships and engagement with external communities. Example roles include: Research IT leadership.

The initial version of the Assessment Tool is implemented as a spreadsheet, developed in Google Sheets to facilitate collaborative work among campus teams conducting an assessment. The tool is presented as a series of sheets, each of which represents one of the facings described above. On each of these sheets there is a list of questions that represent key aspects or factors associated with supporting Research Computing and Data; an assessment team will answer the question from three perspectives or 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 thematic groupings within each facing, and are then aggregated into a coverage value for each facing. In addition to the facings sheets, the Assessment Tool also presents a summary sheet that rolls up the assessment results into a single page for use in presentation to leadership. 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).

The Community Dataset aggregates coverage values for all the contributing institutions from the question level up through the summary values for the facing. When institutions request a copy of the RCD CM Assessment Tool, they provide some basic institutional demographic information; this allows us to understand how representative the data is, and also to filter the data according to broad subgroups including public and private institutions, Carnegie classifications, EPSCoR eligibility and minority-serving status, etc.

Figures 1 to 5 below illustrate the demographic characteristics of the 41 institutions that completed the assessment and contributed their data. Some points to note:

- 28 states are represented in the data. While we look forward to more complete coverage of US states, we note that these 28 states include roughly 80% of the R1 and R2 Universities in the US (as reported by Carnegie[7]), and so should be fairly representative.

[7] The Carnegie Classification of Institutions of Higher Education, https://carnegieclassifications.iu.edu/. Accessed on 11/30/2020.

- Roughly ¾ of our reporting institutions are public, and ¼ 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 comparable to the broader US mix.
- Just under 20% 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[8]). However, we note that our Dataset includes no Historically or Predominantly Black Colleges and Universities (HBCU/PBIs) or American Indian-serving (Tribal Colleges and Universities/TCU) institutions (a gap that we hope to remedy in future years through targeted outreach and support).
- Somewhat over 24% of our reporting institutions are in EPSCoR eligible states. This is slightly higher than the ~22% (58 out of 266) of R1 and R2 institutions Carnegie reports in these states, although only 9 of the 25 EPSCoR states are represented.

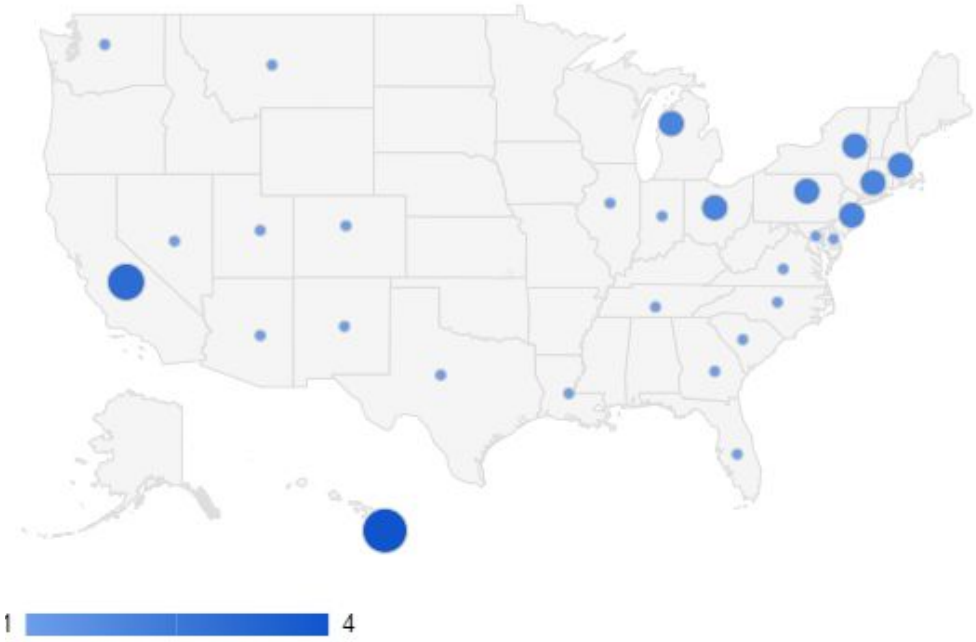


Figure 1: Contributing Institutions by Carnegie Classification

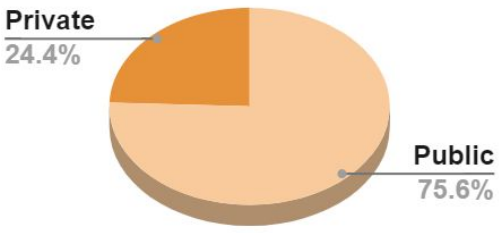


Figure 2: Contribution Institutions by type of control

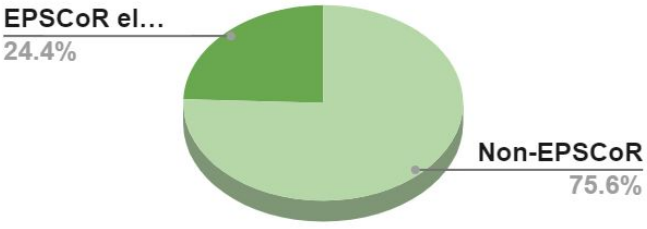


Figure 3: EPSCoR Eligibility of Contributing Institutions

[8] National Center for Education Statistics, DataLab Tables Library, https://nces.ed.gov/Datalab/TablesLibrary/TableDetails/3995. Accessed on 11/30/2020.

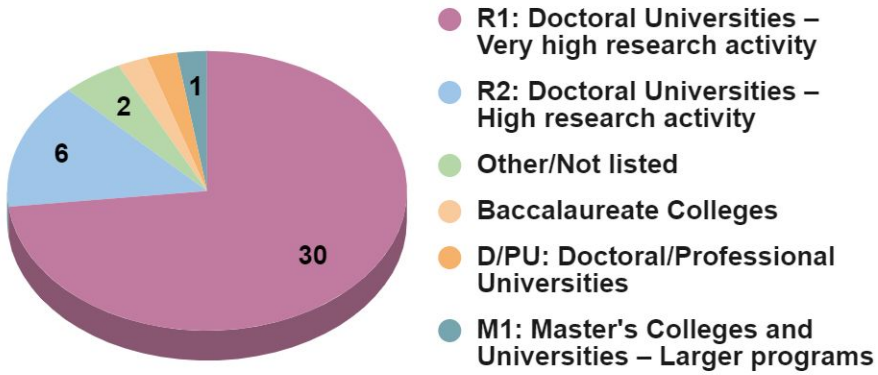


Figure 4: Contributing Institutions by Carnegie Classification

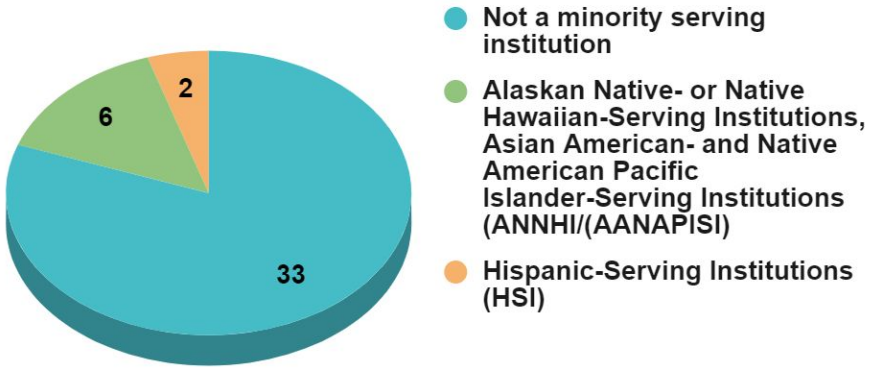


Figure 5: Contributing Institutions by Minority Serving Status

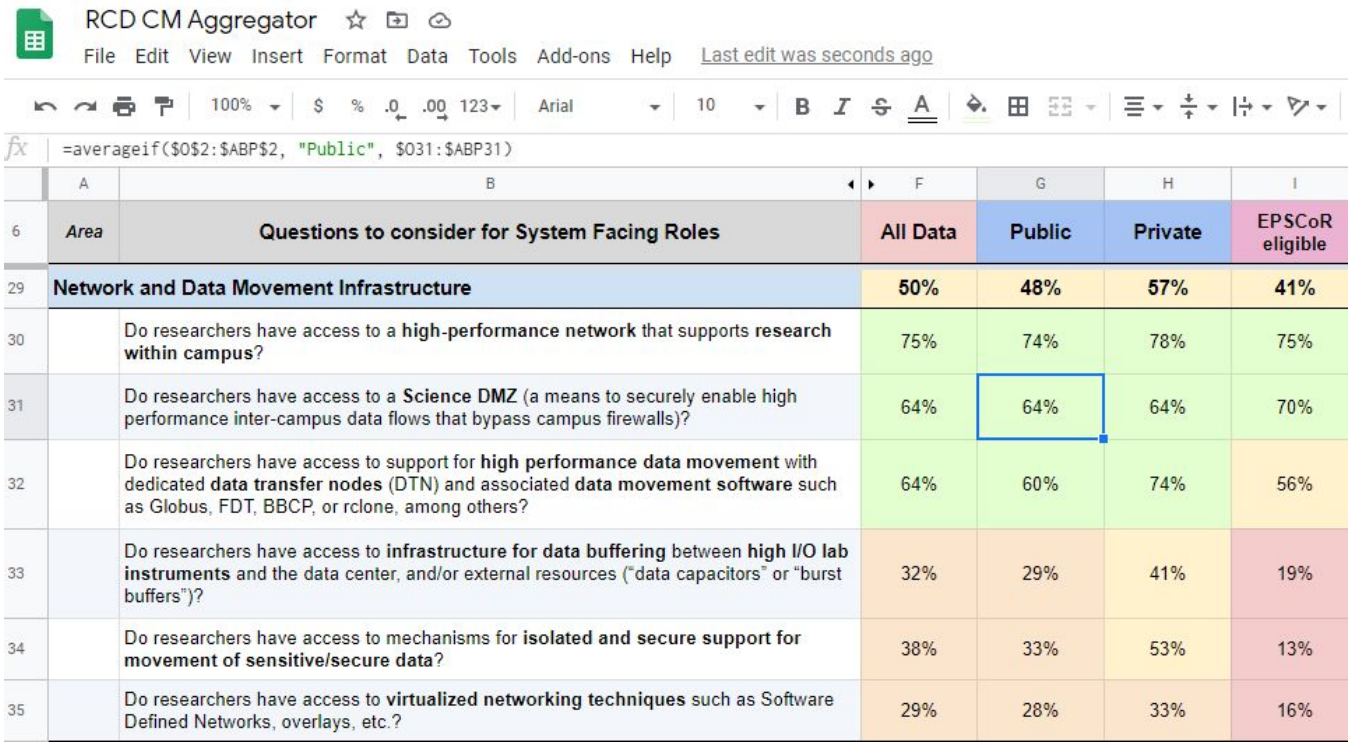
1.3. Methodology for gathering and analyzing the data

As noted above, the initial version of the Assessment Tool is implemented as a spreadsheet, which does not easily lend itself to data aggregation and analysis[9]. We assembled the 2020 Dataset using the same underlying platform (Google sheets), using this methodology:

1. We created a snapshot copy of each contributed and completed assessment to ensure that the primary data was stable.
2. We defined an aggregation sheet with the same primary structure as the RCD CM Assessment Tool with a tab for each facing and a summary tab. This included the original questions and groupings as the Assessment Tool so that the aggregated data could be viewed in the same basic format as the assessment itself. We used the same conditional formatting to produce the heat-map visualization of coverage (green for high coverage levels through yellow and red for low coverage values/gaps).
3. A data column was added for each institution that:
    a. Included the basic institutional demographic information from the assessment request tool, and
    b. Leveraged the IMPORTRANGE[10] function to capture the coverage values for each row (corresponding to the questions) in each facing.

[9] For this reason (and others), we have proposed a version 2 of the RCD CM Assessment Tool that would be survey based and that would gather data into a database with a data exploration portal.
[10] See: https://support.google.com/docs/answer/3093340

4. We computed average overall coverage for each row and used conditional aggregation functions (e.g., AVERAGEIF[11]) to compute the average coverage for each row by major demographic filters (e.g., all Publics, R1s, etc.). This is illustrated in Figure 6 below.
5. We modified the summary tab only slightly to gather the associated values from the respective facings tabs.
6. Although the above was sufficient for basic visual analysis, we found that we needed to create a data tab that gathered all the institutional data (and associated metadata) into a single range, to more easily manage complex queries that supported computation (e.g., of standard deviations).



	A	B	F	G	H	I
6	*Area*	**Questions to consider for System Facing Roles**	**All Data**	**Public**	**Private**	**EPSCoR eligible**
29	**Network and Data Movement Infrastructure**		**50%**	**48%**	**57%**	**41%**
30		Do researchers have access to a **high-performance network** that supports **research within campus**?	75%	74%	78%	75%
31		Do researchers have access to a **Science DMZ** (a means to securely enable high performance inter-campus data flows that bypass campus firewalls)?	64%	64%	64%	70%
32		Do researchers have access to support for **high performance data movement** with dedicated **data transfer nodes** (DTN) and associated **data movement software** such as Globus, FDT, BBCP, or rclone, among others?	64%	60%	74%	56%
33		Do researchers have access to **infrastructure for data buffering** between **high I/O lab instruments** and the data center, and/or external resources ("data capacitors" or "burst buffers")?	32%	29%	41%	19%
34		Do researchers have access to mechanisms for **isolated and secure support for movement of sensitive/secure data**?	38%	33%	53%	13%
35		Do researchers have access to **virtualized networking techniques** such as Software Defined Networks, overlays, etc.?	29%	28%	33%	16%

Figure 6: Screenshot of the Aggregation tool for a segment of the System Facing tab

We found the graphing tools in Google Sheets too simplistic, and copied the data into Microsoft Excel for additional filtering and graphic creation. We debated creating a database and using more sophisticated tools for analysis, but given our time and resource constraints decided to stay with the simpler (if rather inelegant) tools.

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. At the same time, some institutions reported some consternation at the coverage value percentages and the possible comparison to *grades* in a course. The capabilities coverage values are on a scale that is intended to be highly aspirational and most institutions will likely be in a mid range of coverage percentage for many areas. In contrast, course grades are intended to measure against a reasonable expectation of full comprehension and so are not at all comparable. Nevertheless, some felt that the comparison would be inevitable and would result in unnecessarily harsh evaluation of the numeric assessment results. As the data shows, the bulk of coverage values range from 35% to 70%; while these make sense on the capabilities coverage scale *as it was intended to be used* these would correspond to grades ranging from 'F' to 'C' and looked at through this lens, the assessment results would be discouraging, if not damning.

[11] https://support.google.com/docs/answer/3256529

We developed a tool to produce a benchmarking report that places an institution's results in context to the community results. Given the feedback on simple grade associations RCD CM working group members suggested that a useful tool would allow institutions to see their assessment data presented as percentile values relative to the community in the hope that this would approximate "grading on a curve," and we implemented such a tool. However, as we tested the output it became clear that the tool provided little value beyond a comparison to the community values and it just amplified the distance from the mean. While the mean values do range well below what would be an 'A' or 'B' grade, the range of values extends far enough that some institutions have "ruined the curve" (as students might describe it).

We then tried computing a *curved grade* based upon a bell-curve distribution of results, but we felt that the resulting "grades" were still a significant distraction from the goals of a benchmarking exercise. In our third iteration we changed the model to simply indicate which of four tiers an institution's coverage sits, allowing them to see their comparative capabilities coverage in a less semantically loaded manner. The Tiers are just four equal divisions of the institutions (computed for each demographic slice). We considered a model in which tier 1 was lowest, in part to allow for upward adjustment of the scale upwards (e.g., adding a new top Tier 5) as institutions develop broader coverage over time. However, this would alter the definition of the tiers and undermine benchmarking evaluation over time (e.g., to track the progress of one's institution in response to strategic investment in priority areas). We settled on a model in which the Tiers are labelled from I to IV (highest to lowest) and are color coded from blue (Tier I) to red (Tier IV). Figure 7 illustrates the resulting tool (edited to highlight the indication of an institution's RCD coverage Tier relative to several sub-communities).

The RCD CM Working Group is engaged with the group of EPSCoR-eligible institutions that contributed to the 2020 Dataset and is planning a deeper analysis of the data from these institutions; as a result, we do not include benchmarking information for this demographic slice in the reports for other institutions.

	Ranking for:			All Data (41)	Publi (31)	Tier among Publics	Tier among Privates	Tier among R1s	Tier among R2s
	Facing Area *(click the "+" to the left of each to expand)*	Coverage	Tier	Average Coverage	Averag Coverag				
4	**Researcher Facing Capabilities**	52%	III	53%	52%	II	IV	III	III
5	*Research Computing and Data Staffing*	61%	III	60%	58%	II	IV	III	III
6	*Research Computing and Data Outreach (Initial Contact)*	39%	III	46%	44%	III	IV	III	III
7	*Research Computing and Data Advanced Support*	60%	II	53%	52%	II	III	III	I
8	*Research Computing Management of the Research Lifecycle*	55%	III	59%	58%	III	III	III	I
10	**Data Facing Capabilities**	44%	III	49%	47%	III	IV	III	II
19	**Software Facing Capabilities**	49%	II	46%	45%	II	III	II	III
28	**System Facing Capabilities**	60%	II	56%	57%	II	III	II	III
44	**Strategy and Policy Facing Capabilities**	35%	IV	56%	53%	IV	IV	IV	III
52	**Total Organizational RCD Coverage**	48%	III	52%	51%	III	IV	III	III

Figure 7: Screenshot of Benchmarking report tool

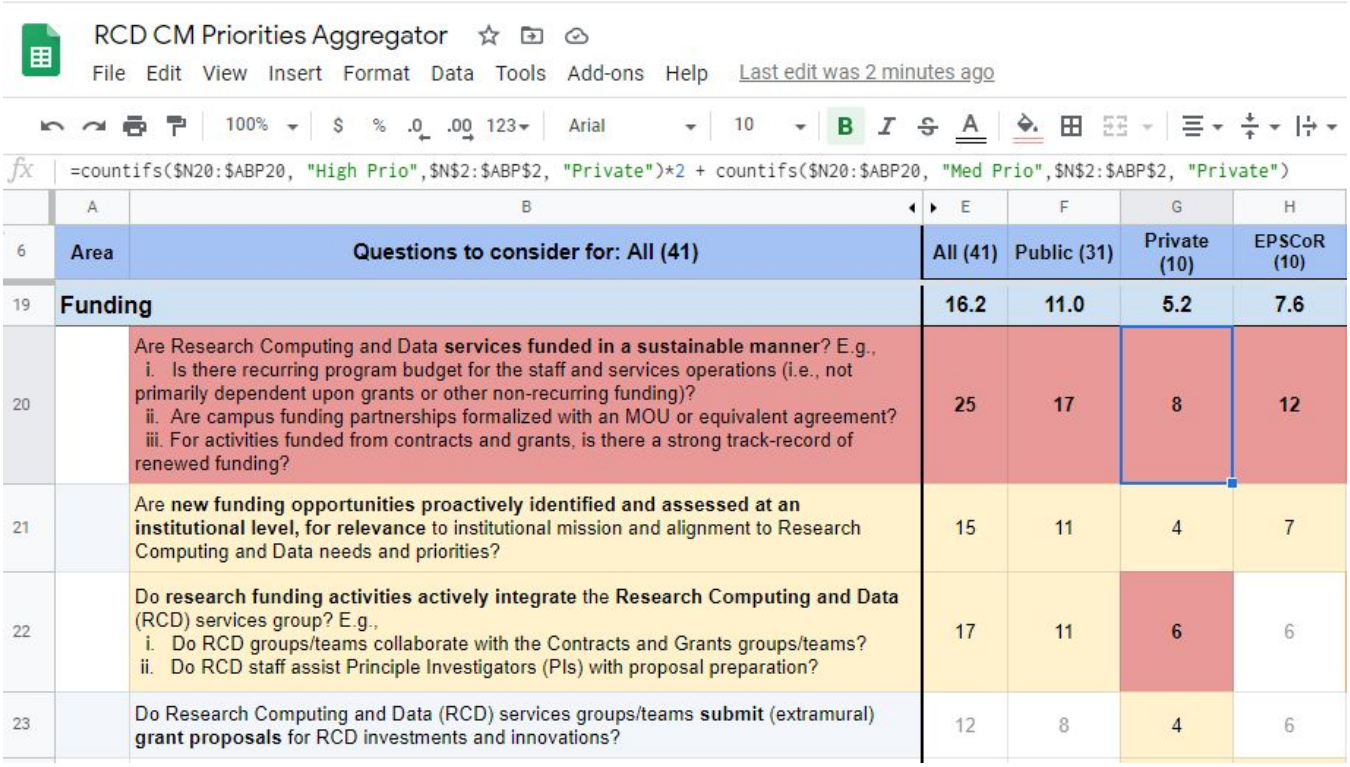
1.3.2. Aggregating Institutional Priorities

Aggregation of the priorities was somewhat less straightforward for a number of reasons:

- Not every institution marked items for priority, yielding some sparsity (especially for filtered sets).
- There are no constraints on minimum or maximum numbers of priorities (although we did encourage teams to recognize that if everything is a priority, nothing is a priority). The summary page includes the first 5 items marked as a priority on each facing tab, but we wanted to recognize any and all items marked as strategic priorities.
- Inasmuch as this feature was added in response to requests from the community, we did not incorporate it into the capabilities coverage calculations. As a result, we had no mathematical formalization that we could leverage to develop summary statistics.
- In the initial exploration of the resulting data we found that summarizing priorities by broad themes, and/or by facings, left out many interesting features in the data.

We initially developed a tool that had similarities to the capabilities coverage aggregator so that we could present the data in a manner analogous to the Assessment Tool:

1. There are tabs for each facing and each question has a row; rows roll up to themes.
2. There is a summary tab that aggregates the data by facing and broad theme.
3. We developed a mathematical model in which a “Medium Priority” was assigned 1 Priority Point, and a “High Priority” was assigned 2 Priority Points.
4. We summed all points for each row/question and developed filters based upon institutional demographics.
5. We summarized the broad themes by averaging the Priority Point sums for the row/questions in each theme and averaged all the rows for each facing.



	A	B	E	F	G	H
6	Area	Questions to consider for: All (41)	All (41)	Public (31)	Private (10)	EPSCoR (10)
19	**Funding**		16.2	11.0	5.2	7.6
20		Are Research Computing and Data **services funded in a sustainable manner**? E.g., i. Is there recurring program budget for the staff and services operations (i.e., not primarily dependent upon grants or other non-recurring funding)? ii. Are campus funding partnerships formalized with an MOU or equivalent agreement? iii. For activities funded from contracts and grants, is there a strong track-record of renewed funding?	25	17	8	12
21		Are **new funding opportunities proactively identified and assessed at an institutional level, for relevance** to institutional mission and alignment to Research Computing and Data needs and priorities?	15	11	4	7
22		Do **research funding activities actively integrate** the **Research Computing and Data** (RCD) services group? E.g., i. Do RCD groups/teams collaborate with the Contracts and Grants groups/teams? ii. Do RCD staff assist Principle Investigators (PIs) with proposal preparation?	17	11	6	6
23		Do Research Computing and Data (RCD) services groups/teams **submit** (extramural) **grant proposals** for RCD investments and innovations?	12	8	4	6

Figure 8: Screenshot of the Priorities Aggregation tool for a segment of the Strategy and Policy Facing

The resulting tool is illustrated in Figure 8. As with the coverage values, we wanted to quickly identify significant features in the dataset (i.e., areas that many institutions in a given demographic slice had indicated as a priority). We developed a conditional formatting scheme based upon an empirical analysis of the distribution of sums and averages; the formatting rules are calculated based upon the average value of Priority Points for that facing (i.e., for the demographic slice in that facing):

- Cells with values below the average value for that facing are left uncolored and text is shaded down
- Cells with values 1.0 and 1.2 times the average value for that facing are colored yellow
- Cells with values 1.2 and 1.4 times the average value for that facing are colored orange
- Cells with values at or above 1.4 times the average value for that facing are colored red

The summary tab uses a variant on the same conditional formatting rules, but is based upon the overall average of Priority Points. As Figure 8 shows, however, the average an area can effectively "hide" significant priorities on individual rows (the 25 Priority points on the funding question make this a very significant question, but the average value for the Funding theme, and for the Strategy and Policy Facing tab, are so diluted by the other areas that the summary view only shows these summary values as low to moderate features).

Given this result, we developed a second tool that simply filters the data to select the top ten individual areas of priority across the entire RCD Capabilities Model (with filters for different demographic slices). While the first tool does give some broad indication of where institutions planned to prioritize attention and resources, the second tool does a better job of pulling out specifics. Nevertheless, marked priorities are spread widely and so the top ten lists should not be understood to be the main or by any means exclusive areas of priority for the community.

The results are discussed in Section 3 below and the resulting data for the Priorities Aggregation Tool is presented in Appendices C, D, and E.

1.4. Observations and reflections on the analysis

A number of things emerged as we worked through the analysis, in addition to the most obvious issue: we need a proper data analytics platform to support this work going forward (especially so if we hope to consider longitudinal analysis).

Mean vs. Median in the data

In addition to the broad variance in the data from different institutions, there is a fairly large difference between the ***mean*** and the ***median*** for many individual capabilities (rows), and also for the summary values for broader themes of capabilities. This seems to indicate that a few outliers are skewing some of the average values and we debated using (or at least exploring the use of) a trimmed mean. However, the current data analysis environment does not easily afford such exploration (this is a good use-case for a more powerful analysis platform such as has been proposed for a future implementation of the Assessment Tool).

The argument for a trimmed mean posits that there are relatively few outliers and that a modest trim value (e.g., 5% - 10%) would bring the mean and median into closer alignment. However, we would need to explore whether there were *outlier institutions* broadly, or simply outlier *values* for individual capabilities (rows). This leads to the question of whether it is reasonable to only trim the outlier values looking at each capability area individually, or whether we should attempt to identify outlier institutions to be more consistent in the data that is filtered.

In part to understand the variability of the data more broadly, we developed a scatter plot of the facings summary values for all 41 institutions (this is discussed in section 2.1, below). Looking at this scatter plot there is really only one institution that is both fairly consistent among the facings, and an outlier among the others (with a very high assessment across all facings). However, it is not clear that their assessment is any less valid than the others (even for this institution, one facing is in the mid-range of coverage). For each individual (facing-level) value that is at an

edge of the distribution, the corresponding values in the other facings for that institution are generally well within the distribution range. Given all this, it is not clear that a trimmed mean will produce a more representative value.

In particular for institutions reviewing the data as part of a benchmarking exercise, the full data set should be likely be retained (i.e., trimming no values) to represent the range across institutions.

Looking ahead, a more full-featured data exploration portal should perhaps support selection of either mean or median values in the visualizations, etc. Here is one example of the utility this would provide: There are some notable examples in which the average coverage is low, but the median coverage value is **zero** (e.g., for the questions: "*Do researchers have access to usability testing for research software developed on campus?*" and "*Is there a practice in place for whole system testing (e.g./chaos monkey) on resources that support research?*") indicating a broad pattern of gaps in these specific capabilities. The ability to readily identify patterns like this would be helpful to anyone exploring the Dataset.

Notes about Carnegie Classifications

We had sufficient demographic range across institutions to present three demographic slices based upon Carnegie Classification: R1s, R2s, and "Other than R1 & R2." However, the "Other than R1/R2" group is a very mixed bag, including a mix of *Master's Colleges and Universities*, *Baccalaureate Colleges*, and the *Other* category in the Carnegie Classification which can include research centers and laboratories. As such, the patterns for this group may not be all that meaningful. We are inclined to wonder which institutions the leaders of a large national center or lab would consider to be peers for the purpose of benchmarking. Would they want to benchmark relative only to other such centers/labs, or to R-1 universities, or some other group? As we gain greater experience (and benefit from more feedback) we may reconsider how we support benchmarking.

We have several institutions in Canada that are using the RCD CM. The Carnegie Classification does not strictly cover Canadian institutions, however they have been self-assessing their category and we assume this is accurate enough to use as we group institutional data. We have generally tried to align our model for institutional demographics to the model used by the EDUCAUSE Core Data Service (CDS)[12]. Going forward, it is our understanding that the CDS model may evolve to better accommodate institutions outside the US, while still retaining the key aspects of the Carnegie Classification system, and we hope to leverage the CDS model as it evolves.

Custom Demographic slices

We have been approached by several groups of institutions that would like to analyze the data from their group to understand their collective state, to identify patterns and opportunities for collaboration, and of course as a benchmarking exercise. Our current resources are stretched to fulfill these requests, but they provide additional requirements input for a data exploration portal in a future version.

1.4.1. Observations on Aspects of the RCD Capabilities Model

The analysis of contributed assessment data revealed issues or in some cases underscored issues that had been under discussion in the community. Two issues in particular are worth mentioning:

Unweighted vs. Domain-weighted summary coverage values

In the full RCD Capabilities Model Assessment Tool, we included support for assessing the breadth of support across different academic domains. The summary page of the Assessment Tool included the summary capabilities coverage values, as well as summary values that were *weighted* by the domain coverage assessment values. In our review of the contributed assessment data it was not clear that institutions were at all consistent in their

[12] EDUCAUSE Core Data Service. Available (as of 12/1/2020) at: https://www.educause.edu/research-and-publications/research/core-data-service.

interpretation and use of the domain coverage portion of the assessment, and so we used the unweighted capabilities coverage values in our data analysis.

Issues with *Local Weight/Relevance*

The Assessment Tool includes a column in which institutions can indicate that a given question or aspect has little or no meaning or application at their institution[13]. When a given question is marked as "***Not relevant or applicable***" the computed coverage value is 100% (even though an assessment of coverage would typically indicate "***No existing service or support***"). The intended effect was that not supporting that aspect of capability would not (unfairly or unreasonably) reduce the summary coverage for a given theme or facing. However, not all institutions used this feature consistently, and so for some rarely supported capabilities, some institutions had coverage values of 0% while others had 100%. This introduces several problems:

1. It is difficult to interpret an average of these coverage values given a mix of actual coverage that varies from 0% to 100% and some weighted coverage values that should be 0% but are converted to 100%.
2. The weighted coverage values skew the summary values upward for the broader themes, and to a lesser extent (given the number of rows involved) for the facing summary values.
3. The option to indicate that a given capability was "***Somewhat relevant or applicable***" (with a partial weighting of the capabilities coverage value) further complicates the interpretation of aggregated values and seems to have been poorly understood and/or rarely used by the community.

In reviewing all this, we have concluded that a better model is to refine the Assessment Tool to 1) simplify the question of local applicability to a simple boolean, and 2) refine the computational model to simply *exclude* rows that are marked as not applicable from all summary calculations.

2. Significant Themes for Capabilities Coverage

2.1. Community-wide patterns

One of the most striking aspects of the Community Dataset is the significant variation in the data. 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, we had to cap the minimum values in visualizations graphs to 0 because error bars (at one standard deviation) extended below the associated 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 Dataset. Each vertical stripe represents a given institution (in no particular order) with 5 colored dots indicating the summary coverage for the 5 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. I.e., different institutions have strengths and weaknesses in different areas.

The last point may be of interest for sub-communities, regional groups, and other potential collaborators, as it points to potential opportunities for collaboration where partners are likely to have complementary areas of strength, with the potential to share leading practices in different areas.

[13] This was intended for use only in rare cases, e.g., when an institution conducts no research (and has no plans or aspirations to conduct research) that require a given aspect of support.

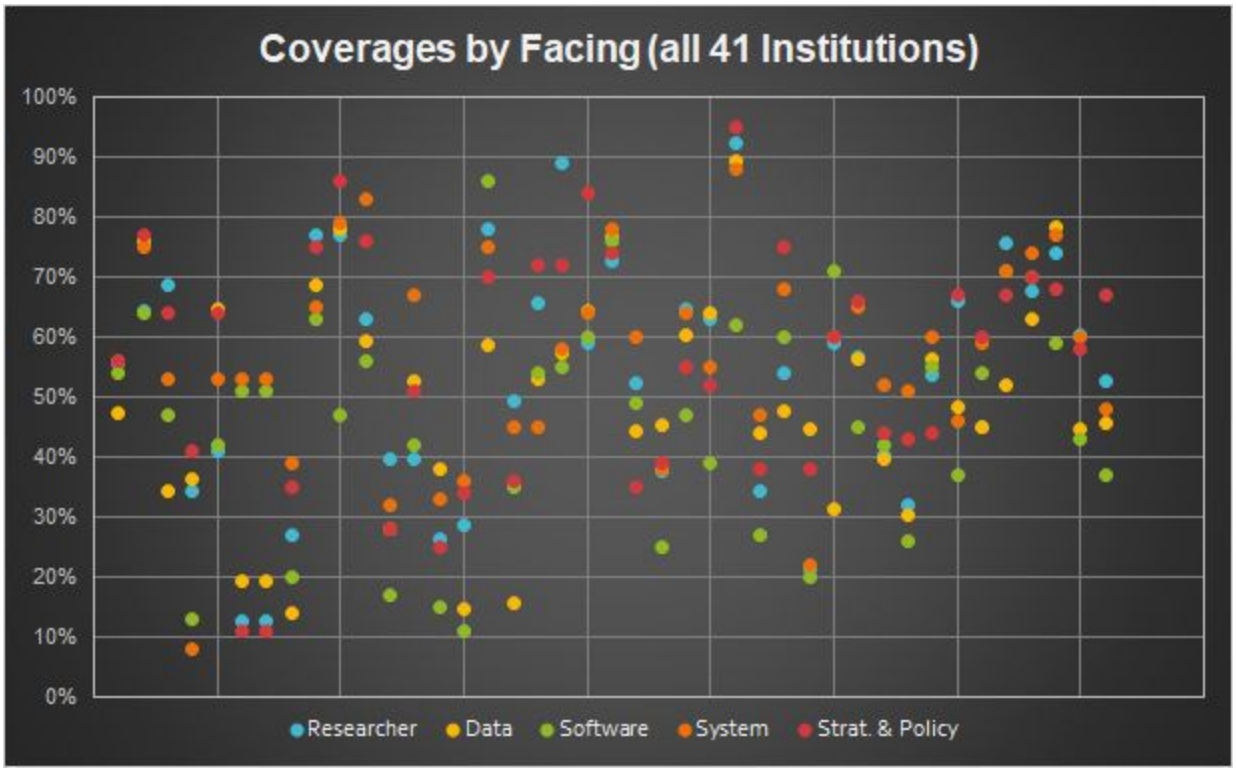


Figure 9: Scatter graph of capabilities coverage by facing for all 41 institutions

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 the institutions. While the variance is slightly smaller for the System Facing, all the facings show considerable variation across the contributing institutions.

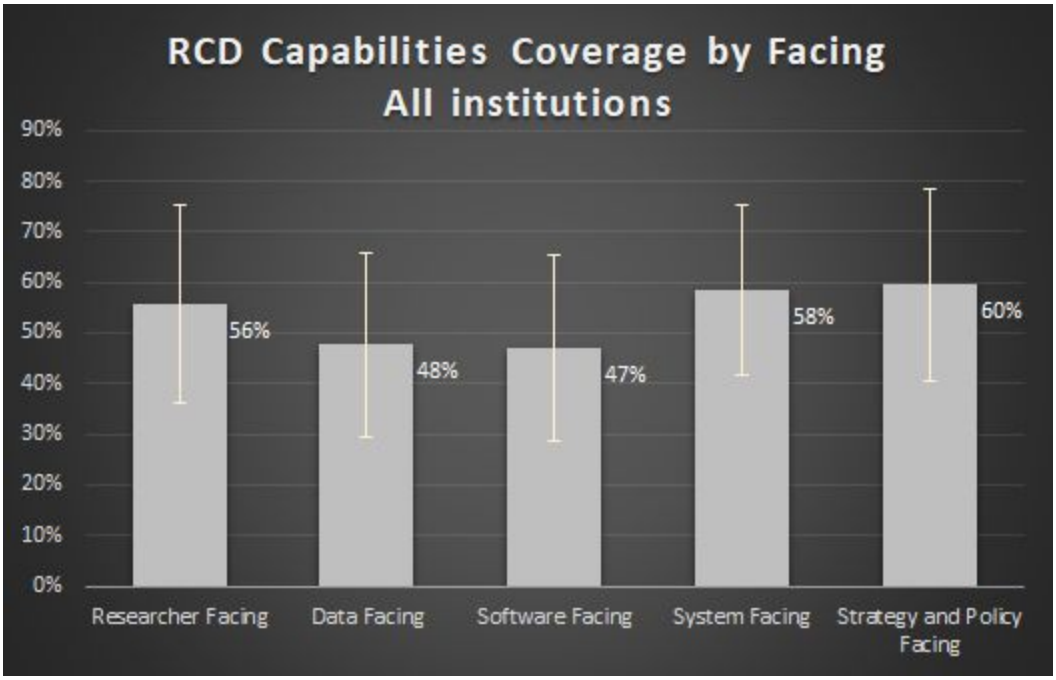


Figure 10 - Capabilities Coverage for all institutions, for each Facing

2.2. Demographic commonalities and differences

Figure 11 presents the summary capabilities coverage for different demographic slices.

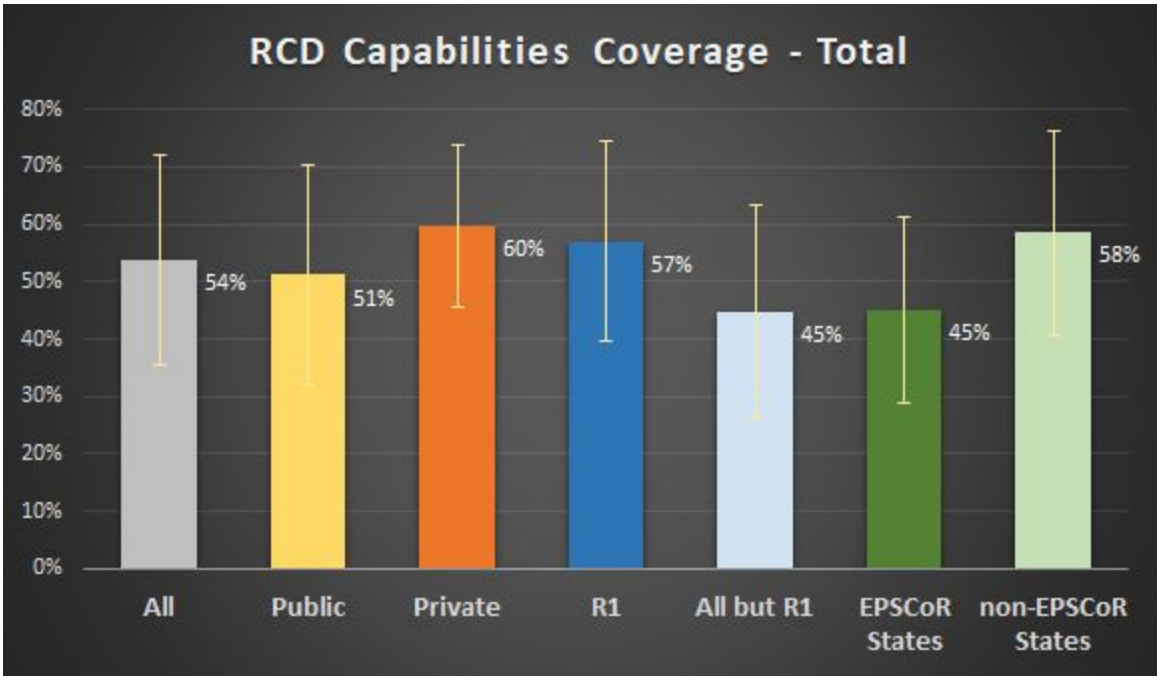


Figure 11 - Total RCD Capabilities coverage by key institutional demographics

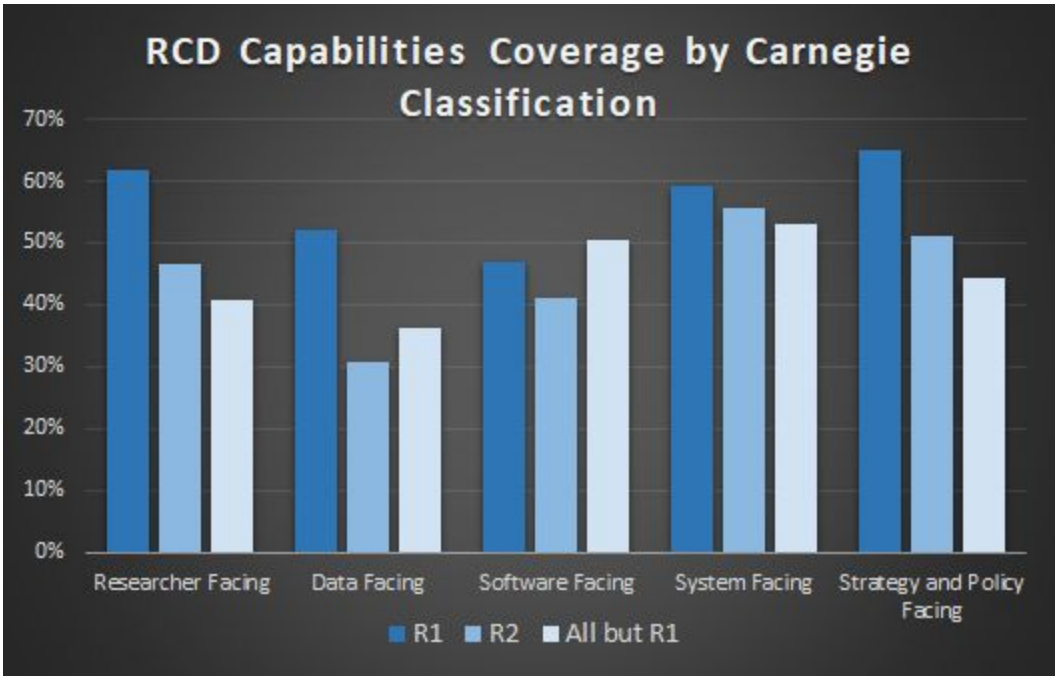


Figure 12 - Median Capabilities coverage across facings by Carnegie Classification

Note the significant difference between public and private institutions, between R1 institutions vs. other institutions, and between institutions in EPSCoR-eligible states vs. those in other states. While each of these might well conform to expectations, the data clearly shows that popular conceptions are borne out in experience and that

these differences are often fairly substantial (although only the EPSCoR distinction is statistically significant, in large part because of the relatively large variance in the data)[14].

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 actually fairly comparable in the System Facing and Software capabilities, perhaps reflecting the longer traditions and understanding of requirements and good practices for systems definition, administration, and maintenance, and of 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.

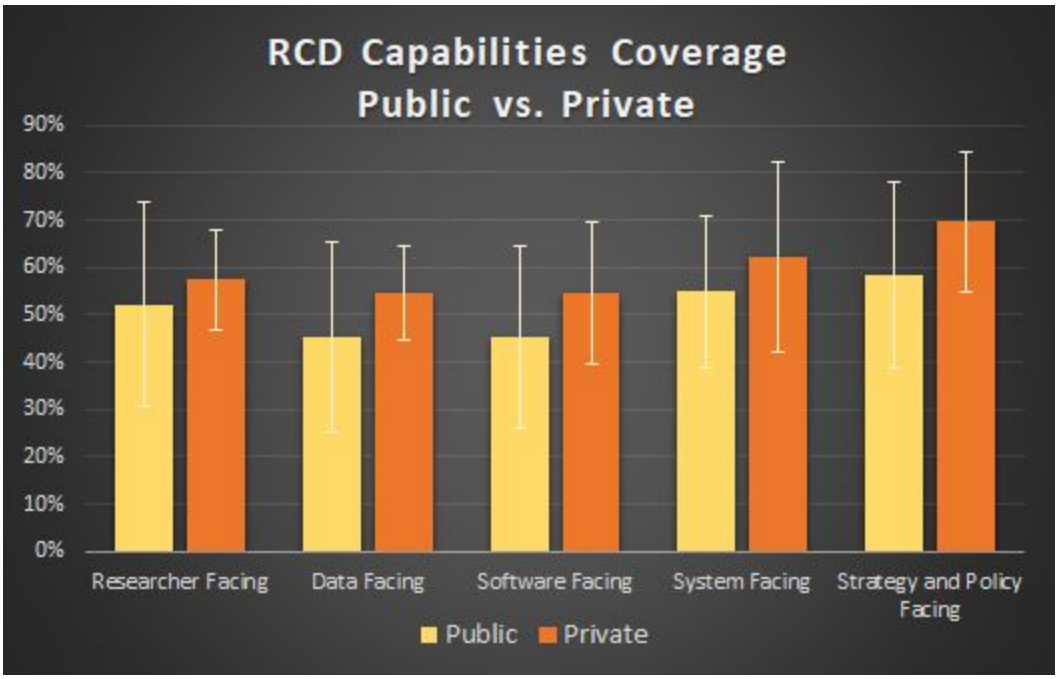


Figure 13 - Capabilities coverage across facings, Public vs. Private Institutions

Figures 13, 14, and 16 present similar comparisons by facing, for Public vs. Private institutions, for institutions in EPSCoR-eligble[15] states vs. other states, and by Minority serving status. The data show that the private institutions consistently have higher capabilities coverage than the public institutions (again, unsurprising to many observers but confirmed here by actual assessment data). It is interesting to note that the *variation* among public institutions is considerably greater than that of the privates, 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 B).

- In the Data Facing themes for ***Data Analysis*** and ***Data Visualization***, private institutions average 25% to 30% higher coverage than public institutions
- Private institutions average 20% higher coverage of capabilities in the Data Facing theme of ***Data Security/Sensitive Data Support***, and the median value for private institutions is 50% higher than that for public institutions.
- Private institutions average 20% to 25% higher coverage of capabilities in the Strategy and Policy Facing themes of ***Institutional Culture for Research Support***; ***Funding***; and ***Partnerships / Engagement with External Communities***. The median values for private institutions in these same themes are from 25% to 44% higher than for public institutions. Although the two groups show much closer values for the

[14] We have not attempted to calculate statistical significance for differences among the various slices of the data. If the reader is interested in this, the full data set is available.

[15] An EPSCoR-eligible 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/

Institutional Alignment theme, it is clear that there are some major gaps in these key areas (and perhaps opportunities for public institutions to learn leading practices from their private peers).
- Counter-examples emerge in the System Facing themes where public institutions have comparable or higher average coverage across a number of themes.

A similar pattern emerges for the EPSCoR vs. non-EPSCoR institutions, but with even starker differences. There is again somewhat closer parity in System Facing capabilities, but the EPSCoR-eligible institutions experience wide gaps across Researcher Facing, Data Facing, and Strategy and Policy Facing areas. The more detailed views (in Appendix B) show significant gaps and a few areas of parity:

- Gaps in the areas of RCD Staffing and RCD Outreach.
- All areas of Data Facing capability, and especially in ***Data Security/Sensitive Data Support*** where EPSCoR-eligible institutions have a median value roughly **one-fourth** that of other institutions (see Figure 15).
- ***Software Associated with Physical Specimens*** (about half the average support and a median support level of 0 (zero)).
- While System Facing capabilities are lower for EPSCoR-eligible institutions, median values are closer to parity between these groups in many areas, except for the themes of ***Storage Infrastructure***, ***Network and Data Movement Infrastructure***, and ***Security practices for open environments*** where there are significant gaps.
- EPSCoR-eligible institutions show much lower assessed coverage values in the areas of ***Institutional Culture for Research Support*** and ***Diversity, Equity, and Inclusion***.

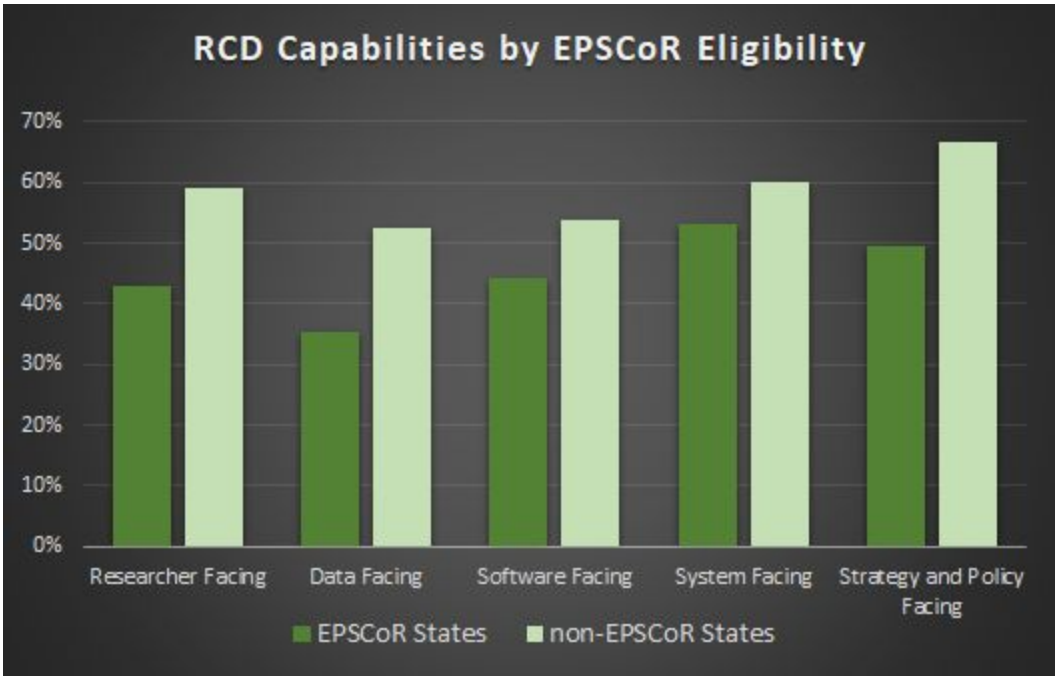


Figure 14 - Median Capabilities coverage across facings by EPSCoR eligibility

Minority-serving institutions (Figure 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. Notably, the median values for themes in the System Facing and Software Facing areas are much closer to parity than the averages are, indicating that some Minority serving institutions are facing even greater gaps than many of their peers (this is echoed by the standard deviations for these Facings, which are a significant proportion of the average values).

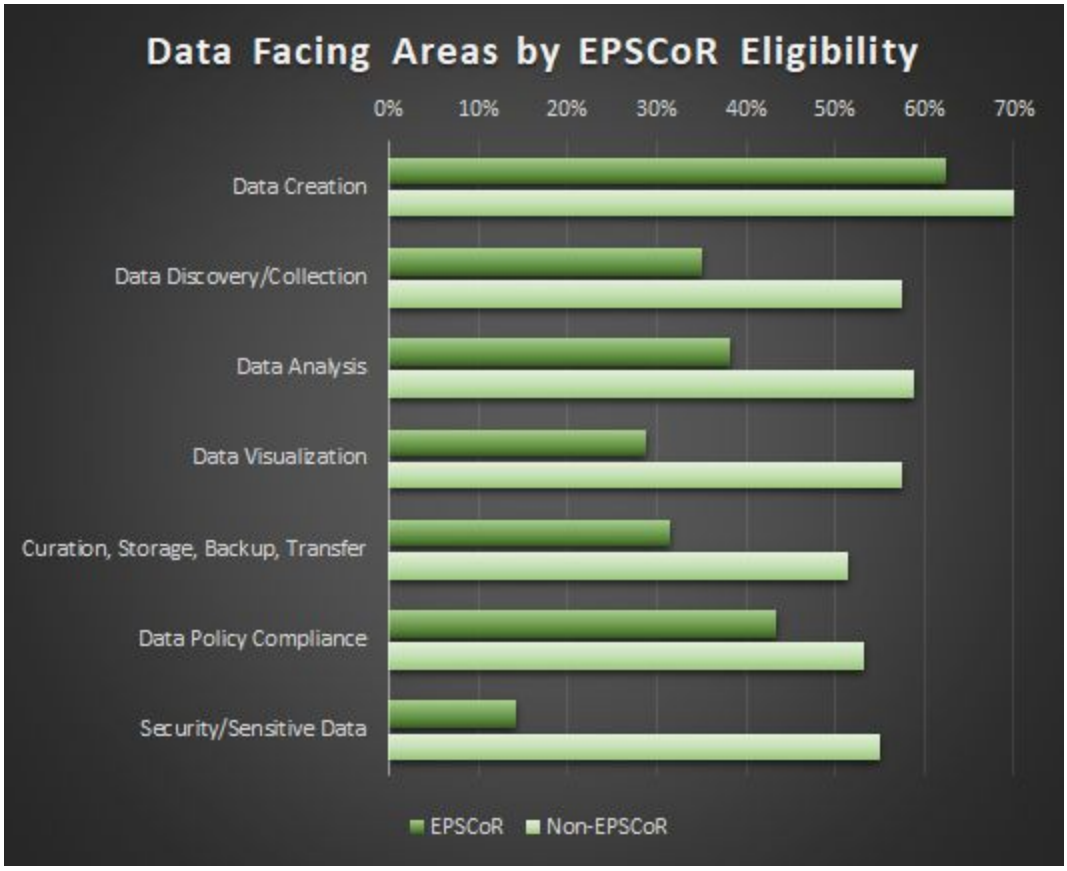


Figure 15 - Median Data Facing Capabilities coverage by EPSCoR eligibility

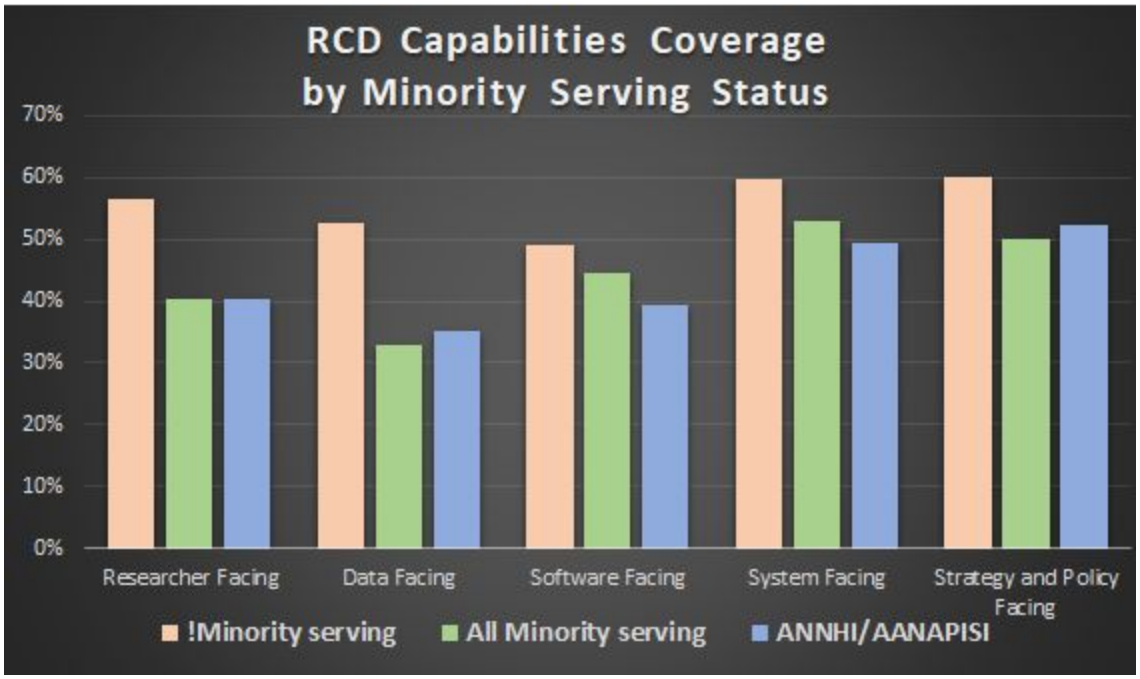


Figure 16 - Median Capabilities coverage across facings by Minority Serving status.

3. Significant Themes for Priorities

We consider both of the approaches to analyzing the institutional priorities (summarizing by theme and facing as for the capabilities coverage, and ranking by individual question), as they each tell a slightly different part of the story. The summary tool with a heat map visualization yields some indication of where the bulk of priorities gather in a broad sense, where the top ranked individual priorities make clear where particular issues are likely to be the focus of resources. In both cases, the numeric data is a sum of all marked priorities from the 41 institutions, where a High Priority counted as 2 Priority Points, and a Medium Priority as 1. As noted above, this data is somewhat sparse, however there is enough to see some patterns emerge.

3.1. Community-wide patterns in the summary data

A graphical representation of the priority points at a summary level is presented in Appendix C, and gives some idea of the areas of emphasis, although (as discussed above) the summary values mask some important individual priorities. This qualification notwithstanding, it is worth noting that for the community as a whole, many of the priorities are in the Researcher Facing and Data Facing areas, and a fair number are in the Strategy and Policy Facing area, with very strong emphasis in:

- *Research Computing and Data Staffing*
- *Research Computing Management of the Research Lifecycle*
- *Data Creation*

...and relatively strong emphasis in these areas as well:

- *Data Analysis*
- *Research Data Curation, Storage, Backup, and Transfer*
- *Research Data Policy Compliance*
- *Software Portability, Containers, and Cloud Computing*
- *Best security practices for open environments*
- *Institutional Alignment*

However, the distribution of priorities varies quite a bit when considering the demographic communities, as shown in Figure 17, below. Some notable differences include:

- Private institutions have more priorities in the System Facing area and less in the other facings, as compared to Public institutions.
- A similar pattern is seen for institutions that are not in EPSCoR-eligible states, as compared to EPSCoR-eligible institutions[16].
- Minority-serving institutions strongly emphasize the pattern of the broad community, with a significantly higher emphasis in the Data Facing and Researcher Facing areas and very little emphasis in the System Facing topics.
- R2 institutions show a strong counter-pattern to the broader community, with a much higher proportion of priorities in the System Facing area and much lower proportions in the other areas (especially Software Facing).
- The group of *Other than R1 and R2* institutions has a distinctly higher proportion of priorities in the Software Facing than the broader community (however this group is a rather mixed bag and so it is not clear how significant this is).

[16] While most of the Private institutions are not in EPSCoR-eligible states, the Private institutions constitute only ⅓ of the total institutions in EPSCoR-ineligible states, and so do not dominate that group (i.e., the pattern comparing institutions by EPSCoR eligibility is distinct from the Public/Private comparison above).

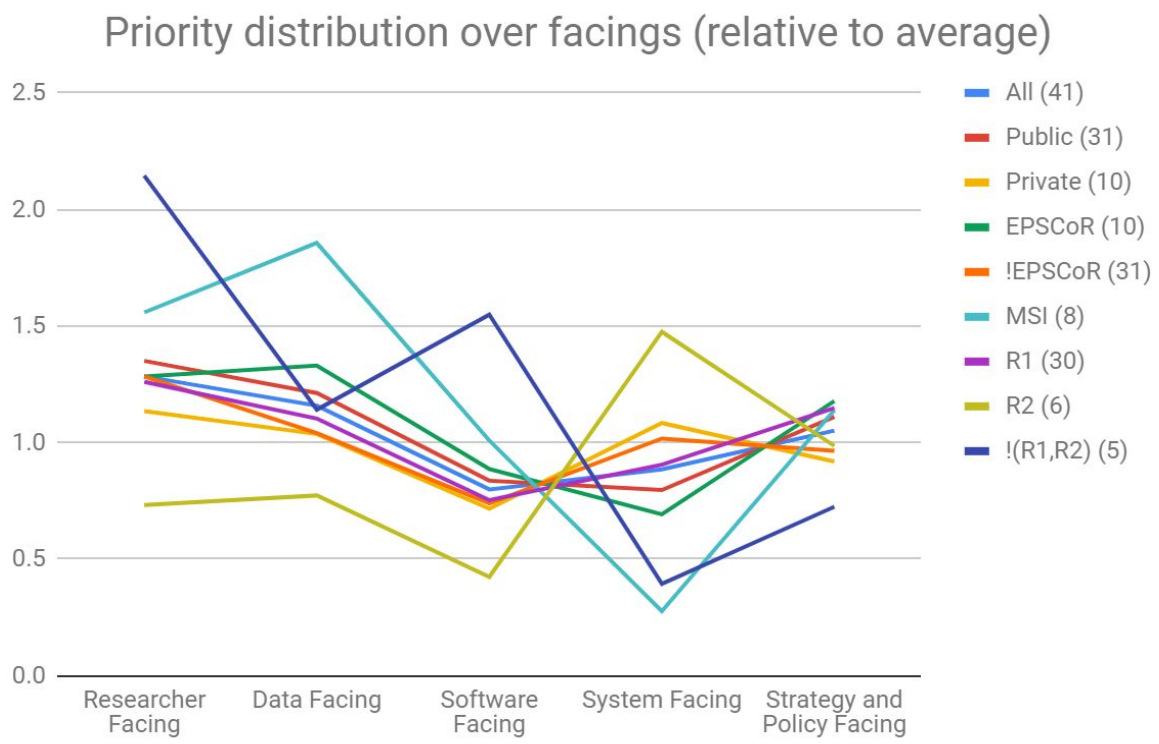


Figure 17 - Priority distribution over facings, by demographic group

#[17]	Capability[18]	Priority Points
1	**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?**	**25**
2	**Are Research Computing and Data services funded in a sustainable manner?**	**25**
3	**Do researchers have access to consulting and expertise to help them identify appropriate data repositories?**	**24**
4	**Are researchers supported across the full research lifecycle?**	**23**
5	**Does your Research Computing and Data (RCD) team/group have a strategic plan?**	**23**
6	**To what extent is there a clear vision, effective guidance, and strategy for the allocation and prioritization of support resources/personnel?**	**22**
7	**Do researchers have access to advice on research compute and data compliance, security, management, and governance?**	**22**
8	**Do researchers have access to tools/software that supports data backup, storage, and integrity checking?**	**22**
9	**Does your institution have research data governance processes in place to establish data policies for research data?**	**22**
10	**Do researcher-facing staff have the skills and capacity to broadly support researchers across levels (graduate students to PIs) and across domains with information about the use and effectiveness of new technologies?**	**21**
11	**Do researchers have access to consulting and expertise on data wrangling/manipulation and data analysis?**	**21**

Table 1 - Top priorities for the entire community

[17] Numbers only for reference. Note that many have the same priority point count.
[18] These are somewhat abbreviated - the full capability titles are listed in the Assessment Tool.

3.2. Top priorities for the community and by demographic grouping

We sorted all the capabilities by total *priority points* and filtered to the top 10 values (which yields one or two additional capabilities where there were ties for the 10th place). Across the full community, the top priorities are listed in Table 1. The list includes an interesting mix of consulting and engagement support for researchers, funding and strategic planning, and several data governance and compliance items.When we filter for particular sub-communities, there are many overlaps but also some interesting differences. The lists of top priorities for each demographic group are presented in Appendix E. Some aspects of these worth noting are described below, for each demographic group.

Public Institutions:

7 of the top 8 priorities for public institutions are the same as those for the broader community, with only very slight differences in the order (and inasmuch as the top priority point counts vary by only a few points, the order is not all that significant). However, there are several additional priorities that appear for the public institutions. Two are closely related to outreach and communications:

- ***Researcher awareness of RCD services across a spectrum of resources****.*
- ***Strategic and policy practices that support Researcher awareness of RCD services (dedicated role, documentation, recognition of importance by leadership)***.

In our workshops and other discussions, many institutions have reported the challenge of outreach and communications to make researchers aware of their services. While the first priority above emphasizes the outcome, the second considers institutional support for making researchers aware of RCD services.
Two additional priorities (not in the broader community list) focus on support for software that researchers use:

- ***Application support for common software packages****.*
- ***Support for software to perform data wrangling, manipulation, etc***.

The second one is closely related to the community priority ***Do researchers have access to consulting and expertise on data wrangling/manipulation and data analysis?*****,** but emphasizes tools in addition to consulting and expertise.

Private Institutions:

Only 5 of the top priorities for private institutions are shared with those for the broader community, including the top two, and numbers 6, 9, and 10. The 5 additional priorities (all of which are at the same level) include:

- ***Compute and data environments for sensitive/secure data***
- ***Support for Software compilation, software good practices***
- ***Institutional support for cloud services for research***
- ***Staff skills for containers, orchestration***
- ***Support for interactive computing (VDI, Jupyter, etc.)***

R1 Institutions:

The top priorities among R1 institutions are largely the same as for the community as a whole, sharing all the but #7 and #9 (although some of these are ranked quite a bit lower, e.g., #4 and #6). The 2 additional priorities are common with the top priorities for public institutions. Given the preponderance of public R1 institutions represented in the Dataset, the overlap is unsurprising.

R2 Institutions:

The top priorities among R2 institutions vary widely from those of the community as a whole, sharing only numbers 2, 4, 5 and 7 (and even these are ranked somewhat differently). Of the additional priorities, 2 are common with the top priorities for private institutions:

- ***Institutional support for cloud services for research***
- ***Support for interactive computing (VDI, Jupyter, etc.)***

Another 5 priorities are distinctive to the R2 institutions, and are listed here in priority order. It is notable that HPC is ranked so high (although below support for cloud services!), and that several are associated with research data management:

- ***Do researchers have access to a place to store final research data to address institutional policy and/or funding agency requirements?***
- ***Do researchers have access to high performance (batch) computing (HPC)?***
- ***Do researchers have access to policies and technologies that facilitate management and wide access to data?***
- ***Is your Research Computing and Data (RCD) strategic plan aligned to campus plans?***
- ***Do research funding activities actively integrate the Research Computing and Data (RCD) services group?***

EPSCoR-eligible Institutions:

EPSCoR-eligible institutions shared 7 of their top 10 priorities with the broader community (numbers 2-6, 9 and 11), although in a different order, with the top three priorities focusing on strategy and funding:

1. ***Does your Research Computing and Data (RCD) team/group have a strategic plan?***
2. ***Are Research Computing and Data services funded in a sustainable manner?***
3. ***To what extent is there a clear vision, effective guidance, and strategy for the allocation and prioritization of support resources/personnel?***

Another priority not in the community-wide list is shared with public institutions: **Researcher awareness of RCD services across a spectrum of resources**, which is a variant on the common theme of outreach and communications for awareness, discussed above. The other 2 on the list are ***Do researchers have access to guidance or training for cloud computing?*** which (in contrast to support for consulting and guidance about cloud services) seems to indicate an interest among their associated researchers in learning the skills themselves, and ***Do researchers have access to dedicated resources (e.g., staff) who can perform data wrangling/manipulation and data analysis?*** which is interesting in that where other demographic groups have prioritized expertise and/or tools, the EPSCoR-eligible institutions seem interested in developing staff resources who can *perform* data wrangling for researchers.

Minority Serving Institutions:

Minority serving institutions seem to have a fair amount in common with EPSCoR-eligible institutions (not surprising given the overlap between these two groups). The two demographic groups share 9 of the same 10 priorities with the only difference being a subtle one: prioritizing ***software*** rather than ***dedicated staff*** in support of *data wrangling/manipulation and data analysis*.

4. Conclusions and Looking Ahead

We have presented an analysis of the first Community Dataset that aggregates 41 institutional assessments using the Research Computing and Data Capabilities Model. The Model itself was developed through a collaboration of Internet2, EDUCAUSE, and CaRCC, and reflects the contributions of many subject matter experts across a range of roles and representing a diverse set of universities and organizations. The 2020 Community Dataset similarly represents a diverse set of institutions and provides significant insights into the state of support programs for RCD.

The assessment itself allows institutions to answer key questions as part of their strategic planning efforts:

- **How well is my institution supporting computationally- and data-intensive research, and how can we get a comprehensive view of our support?**
- **What is my institution not thinking about or missing that the community has identified as significant?**
- **How can my institution (and my group) identify potential areas for improvement?**

4.1. Overall value of the dataset

The 2020 RCD CM Community 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 interest within the community for a community dataset that aggregates the assessments of many institutions has been very high and we expect that institutions will find considerable value in the data presented herein. For institutions that contributed assessment data to the 2020 Community Dataset, the extra detail they have access to should provide important input as they prepare strategic plans.

In many cases, the patterns we describe in the data confirm common perceptions about support across the community and 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 that we have a baseline for understanding not just the current broader state of RCD, but also how support varies across the community and how it evolves. In several cases the data made clear that differences between certain groups within the community are even more profound than many may have expected, and allows RCD leadership and others to refine their understanding of which particular areas of RCD support have wider gaps among sub-communities.

We described in Section 2 how the variation of institutional strengths (and weaknesses) across different areas of capability may present opportunities for collaboration. The RCD CM model itself provides a structured vocabulary that allows institutions to more clearly communicate about RCD support, and the Community Dataset now provides granular insights to groups of institutions that are seeking to collaborate on shared solutions and strategies to advance RCD support.

In discussion of these early results with some community members, institutional leaders confirmed that the RCD CM will make it easier to compare programs and practices, to understand where collaborations can be of benefit, and to identify opportunities for joint projects, funding proposals, etc.

In addition to the capabilities assessment data, the aggregated priorities data provides insight into the areas in which institutions plan to place emphasis, devote resources, etc. This will provide additional information for RCD leadership (among others) as they develop strategic plans.

Figure 18 below shows how the institutions that have downloaded the RCD CM plan to use the model.

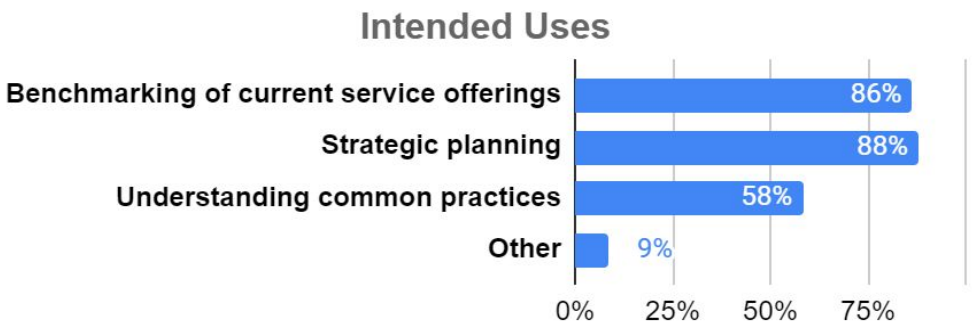


Figure 18 - Intended use of the RCD CM Assessment Tool for all institutions requesting a copy, as of December 2020

We have developed a simple survey to capture more information about institutions' investment (in people's time) in completing the assessment, and the value they perceive in the results and the Community Dataset (institutions will be asked to complete this survey as they request their copy of the more detailed report). We plan to follow up with qualitative interviews with institutions to understand how they are using the RCD CM, the benchmarking reports, and this report as part of their strategic planning process. We plan to share the results of these efforts as well.

4.2. Value over time

In workshops facilitated by the RCD CM Working Group, a number of institutions expressed a desire to repeat the RCD CM assessment over time, to gain insight into the evolution of their support. Especially when they use the priority marking mechanism as input into strategic planning, they would like to have a simple view of where their efforts resulted in improved capabilities coverage.

For groups within the community that are working to improve RCD support for their associated institutions (e.g., EPSCoR-eligible institutions, HBCUs, TCUs, other minority serving institutions, etc.), a shared repository of longitudinal assessment data for their community provides a baseline for planning at any given point, as well as a means of evaluating the impact over time of programs devoted to advancing RCD support in their communities.

At a broader community level we are seeing increased interest in understanding the broad state of RCD support and in working collaboratively to advance such support (e.g., the RCD Decadal Survey currently in planning at CaRCC). A shared repository of longitudinal assessment data will provide essential baseline data for such an effort, as well as a framework for describing the broad scope of RCD support (not to mention the means of evaluating the impact of the decadal survey itself).

The Community Dataset provides clear value from the level of a given institution up to the community as a whole. The 41 institutions that completed an assessment represented about one-third of those who requested a copy of the Assessment Tool, in just the first 6 months since it was released. There is strong interest in the community for this sort of tool and this sort of data, and as the number of contributing institutions grows, so too will the value of the aggregated data.

4.3. Assessment of the tools

In addition to our plans for developing a more robust platform, we have been gathering input on the model itself to understand where it may need refinement to ensure a good experience for institutional assessment teams. The RCD CM Working Group has conducted interviews with several institutions who completed an assessment in 2020, and plans to expand this with additional interviews and discussions. The current set of questions was developed with the input from many institutions over a series of workshops in 2018 and 2019, and we hope that they can remain relatively stable to facilitate longitudinal analysis. However, we will consider whether we can simplify the

means of answering these questions and how we can better support institutions that will be repeating an assessment and may wish to draw upon the previous assessment to streamline the process.

We are particularly interested to compare experiences across institutions and provide additional resources to those interested in using the RCD CM. This will inform our plans for refining the tools, and will support additional work to support institutions using the tool, including:

1. Developing a set of narratives about how other institutions went about gathering the data (with examples from a range of different types of institutions).
2. Support for small institutions to facilitate the assessment process (how to gather a team, how to streamline the process when resources are particularly limited).
3. Support for taking next steps: approaches to developing a RCD strategic plan, drawing upon the RCD CM as input.

4.4. Refining the Assessment Tool and the data analysis platform

The process of completing our analysis of the 2020 RCD Community Dataset (described in section 1.3, above) has underscored the need for a more robust data analysis platform. The RCD CM Working Group has documented plans to re-implement the RCD CM Assessment Tool on a more robust and functional survey platform (although we still need funding for this). Version 1.0 is a Google Sheet (for more about the current implementation, and the structure of the tool, see *Schmitz et al. 2020*, cited in section 1.1 above). Although functional, mining the data from this platform is challenging and cannot easily be automated. Moreover, the Google Sheets platform does not provide a particularly great user experience, lacks affordances for accessibility and localization, and does not provide our desired level of stability, privacy, and security controls. Some of the additional functionality (relative to the Community Dataset analysis and exploration) that is under discussion for the new Data Portal functionality includes:

- Support for interactive exploration of the data, including filtering by the broad demographic categories used in this report. This may replace the expanded report or at least reduce the number of graph appendices.
- Support to tag institutions with consortia and regional communities and then to filter on these tags. This will allow these groups to gather data from their members into reports (for shared planning, etc.). Examples include the Pacific Research Platform (PRP), regional groups like the Pac12 and Big10, the Eastern Regional Network (ERN), RMACC, etc.
- Support for geographic filters by regions, or even states or provinces.
- Support to combine filters to explore, for example, R2 institutions in a geographic region (the ability to narrow an analysis in this way will be subject to a minimum number of institutions in such a group, to prevent identification of individual institutions in the Dataset).
- Support to visualize data for a given time span (e.g., a year in the past), and over multiple years (to explore trends and evolution).

We believe that this new functionality will provide additional value to institutions and to the community, and will further motivate contributions of institutional assessment data.

Acknowledgements

This work has been supported in part by an RCN grant from the National Science Foundation (OAC-1620695, PI: Alex Feltus, “RCN: Advancing Research and Education through a national network of campus research computing infrastructures – The CaRC Consortium”). 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, for which the working group is very grateful.

Appendix A: Graphs by Demographics

We include the graphs of median values for each demographic slice, with error bars indicating a single standard deviation. This is something of a compromise between using the median which is less susceptible to outliers in the data, and the mean from which the standard deviation is properly measured. As noted in the report, there are cases where the median values and the averages differ in interesting ways, indicating that there are a number of outlier values. We have deferred the presentation of average-based vs. median-based visualizations to the future when we hope to have an interactive data exploration portal that would allow users to switch between the two and more easily see the distinctions.

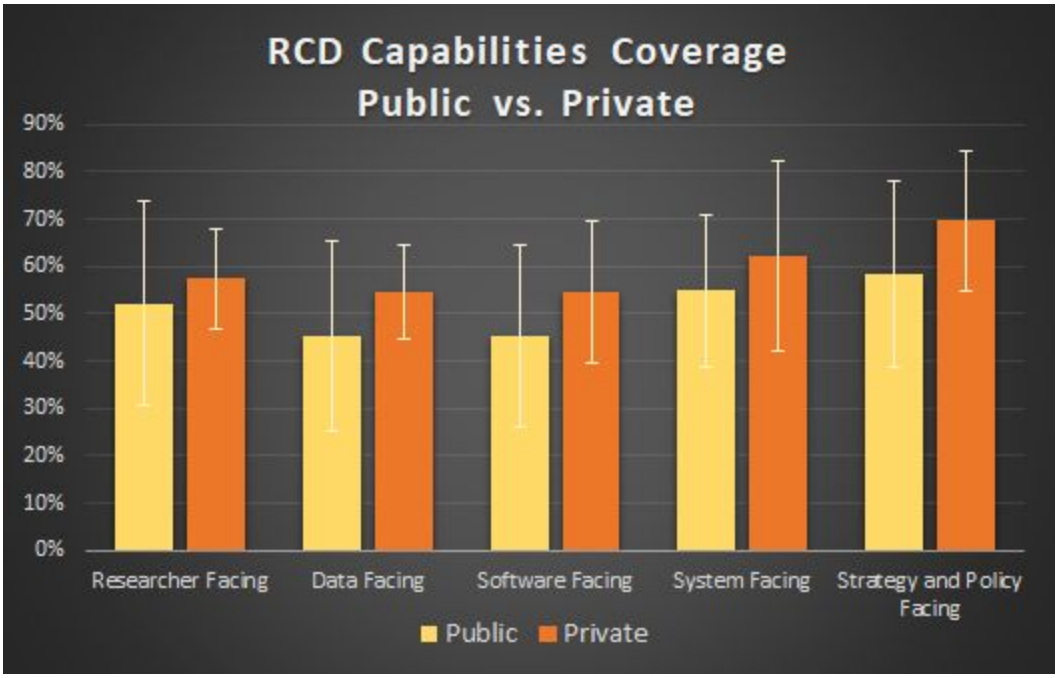


Figure A-1: Capabilities coverage across facings, Public vs. Private Institutions

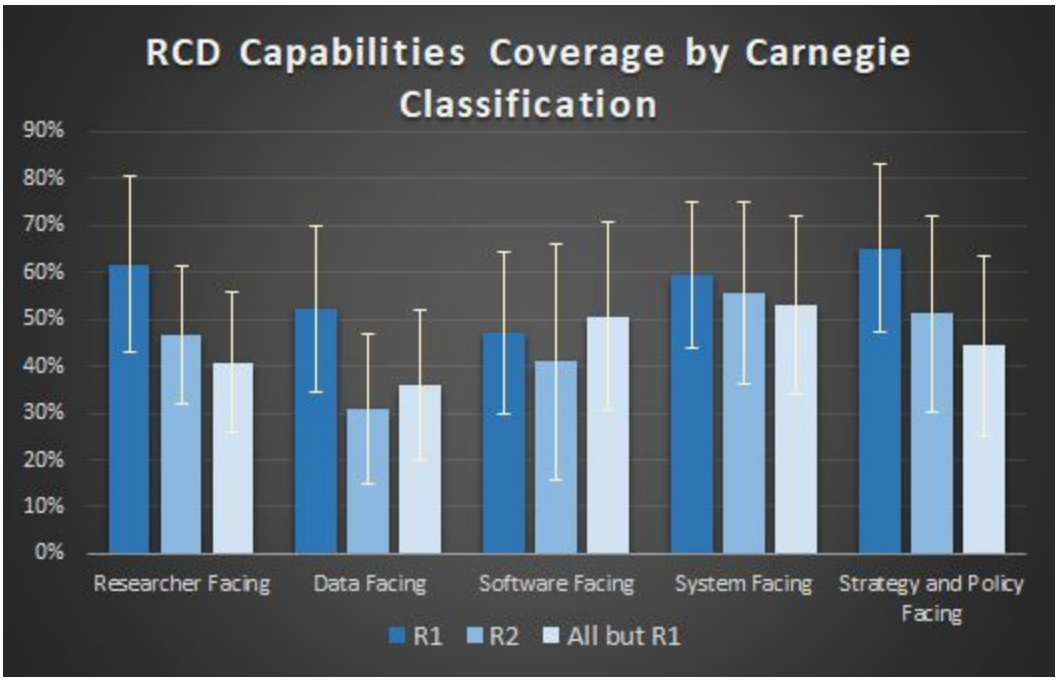


Figure A-2: Capabilities coverage across facings by Carnegie Classification

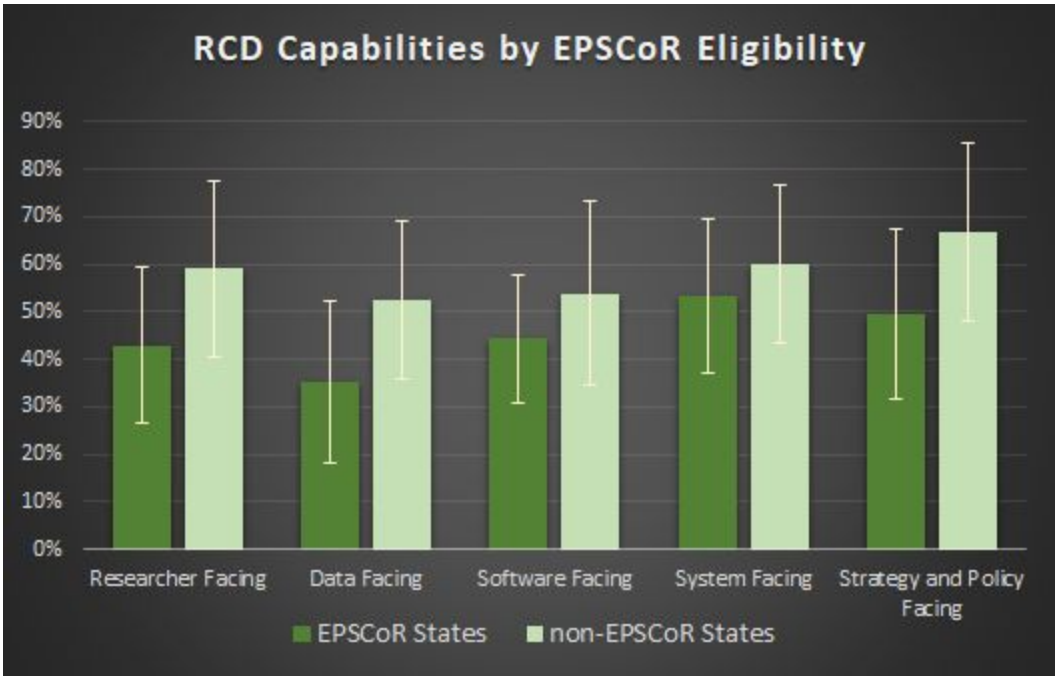


Figure A-3: Capabilities coverage across facings by EPSCoR eligibility

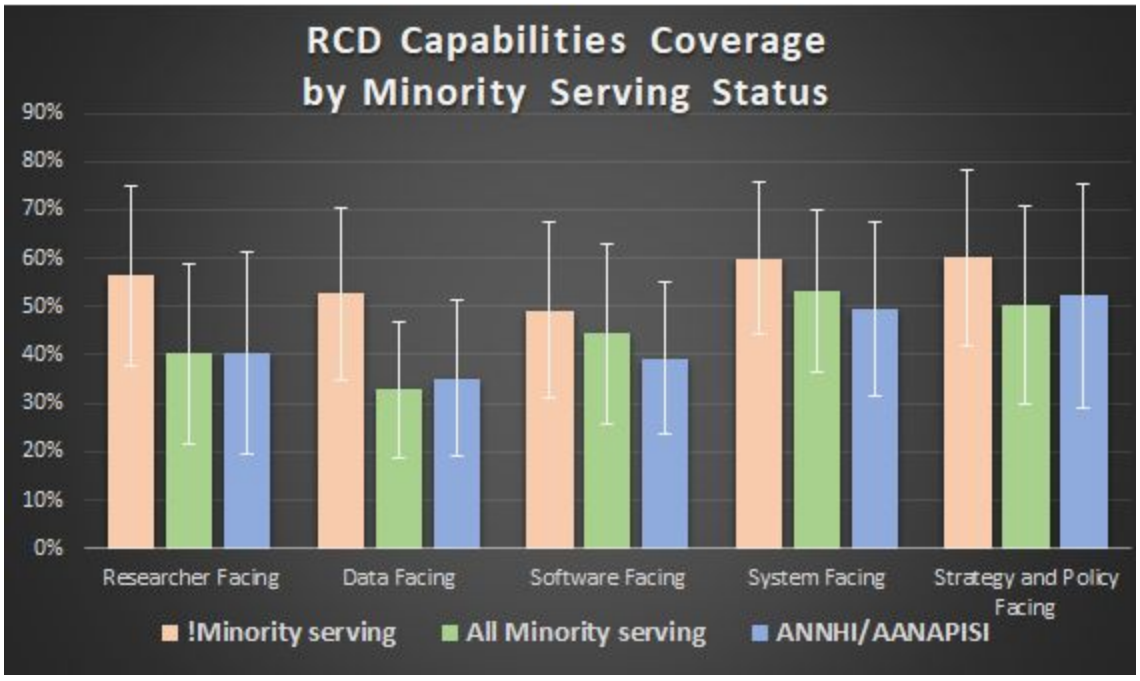


Figure A-4: Capabilities coverage across facings by Minority serving status

Appendix B: Detailed Graphs by Demographics

Appendix B is only available to contributing institutions; it is not included in the public report.

Summary page RCD Capabilities Model Aggregated Priorities

Facing Area (click the "+" to the left of each to expand)	All (41) Priority Points	Public (31) Priority Points	Private (10) Priority Points	EPSCoR (10) Priority Points	!EPSCoR (31) Priority Points	MSI (8) Priority Points	R1 (30) Priority Points	R2 (6) Priority Points	!(R1,R2) (5) Priority Points
Researcher Facing Capabilities Priority Points (average across topics)	18.12	13.18	4.94	7.35	10.76	3.65	13.12	1.18	3.82
Research Computing and Data Staffing	20.40	13.60	6.80	7.80	12.60	3.80	14.80	1.00	4.60
Research Computing and Data Outreach (Initial Contact)	15.83	11.67	4.17	7.00	8.83	3.17	12.00	0.50	3.33
Research Computing and Data Advanced Support	16.50	13.25	3.25	6.50	10.00	3.25	12.00	1.00	3.50
Research Computing Management of the Research Lifecycle	22.50	16.50	6.00	9.00	13.50	5.50	14.50	4.00	4.00
Data Facing Capabilities Priority Points (average across topics)	16.36	11.91	4.45	7.65	8.71	4.38	11.55	1.20	2.04
Data Creation	20.00	13.00	7.00	9.00	11.00	5.00	13.00	3.00	4.00
Data Discovery and Collection	11.75	8.50	3.25	5.50	6.25	3.25	7.75	1.00	3.00
Data Analysis	17.50	13.50	4.00	8.75	8.75	5.75	13.25	0.75	3.50
Data Visualization	15.00	11.75	3.25	7.25	7.75	4.25	11.00	0.50	1.00
Research Data Curation, Storage, Backup, and Transfer	17.14	12.86	4.29	7.86	9.29	4.00	12.43	1.00	1.71
Research Data Policy Compliance	18.17	12.00	6.17	8.00	10.17	4.33	12.17	2.00	1.67
Data Security/Sensitive Data Support	16.00	11.00	5.00	7.67	8.33	4.67	10.67	1.67	1.00
Software Facing Capabilities Priority Points (average across topics)	11.28	8.16	3.12	5.08	6.20	2.36	7.84	0.68	2.76
Software Package Management (installation, documentation, validation, a	13.00	8.67	4.33	6.00	7.00	3.33	7.00	1.67	4.33
Research Software Development	10.00	7.20	2.80	5.00	5.00	2.40	7.40	0.40	2.20
Research Software Optimization or Troubleshooting	10.00	7.50	2.50	3.50	6.50	1.00	7.50	0.75	1.75
Workflow Engineering	13.67	9.00	4.67	6.00	7.67	2.67	9.00	1.33	3.33
Software Portability, Containers, and Cloud Computing	17.33	12.00	5.33	8.33	9.00	4.33	13.00	0.67	3.67
Securing Access to Software	9.50	8.25	1.25	4.25	5.25	2.00	7.25	0.25	2.00
Software Associated with Physical Specimens (e.g., samples, research / n	7.33	5.33	2.00	3.33	4.00	1.33	4.33	0.00	3.00

Summary page RCD Capabilities Model Aggregated Priorities

	All (41)	Public (31)	Private (10)	EPSCoR (10)	!EPSCoR (31)	MSI (8)	R1 (30)	R2 (6)	!(R1,R2) (5)
System Facing Capabilities Priority Points (average across topics)	12.49	7.77	4.72	3.96	8.53	0.65	9.42	2.37	0.70
Infrastructure Systems									
Infrastructure Support	15.40	8.60	6.80	4.00	11.40	0.20	10.40	2.80	2.20
Compute Infrastructure	14.17	8.50	5.67	4.00	10.17	0.50	10.83	2.83	0.50
Storage Infrastructure	14.57	9.29	5.29	4.71	9.86	0.57	11.14	2.86	0.57
Network and Data Movement Infrastructure	11.50	7.33	4.17	3.83	7.67	0.17	9.00	2.33	0.17
Specialized Infrastructure	8.80	5.40	3.40	2.60	6.20	0.20	6.40	2.20	0.20
Infrastructure Software	10.67	6.83	3.83	3.00	7.67	0.17	8.50	2.00	0.17
Systems Operations									
Monitoring and Measurement	9.71	5.71	4.00	3.57	6.14	0.43	7.43	2.00	0.29
Change Mngmnt, version control, administration, and ticketing	11.60	8.00	3.60	3.80	7.80	0.80	9.20	2.00	0.40
Documentation	12.50	8.50	4.00	3.00	9.50	0.00	10.50	2.00	0.00
Planning	13.80	8.40	5.40	4.60	9.20	1.40	10.20	2.20	1.40
Systems Security and Compliance									
Best security practices for open environments	17.00	11.00	6.00	7.33	9.67	4.00	11.67	2.67	2.67
Strategy and Policy Facing Capabilities Priority Points (average across topics)	14.83	10.83	4.00	6.75	8.08	2.67	11.96	1.58	1.29
Institutional Alignment (How policies and priorities are set)	17.14	13.29	3.86	8.29	8.86	3.43	14.00	1.86	1.29
Institutional Culture for Research Support	15.67	11.67	4.00	6.67	9.00	2.67	13.33	1.67	0.67
Funding	16.20	11.00	5.20	7.60	8.60	3.20	12.40	2.20	1.60
Partnerships / Engagement with External Communities	12.33	8.67	3.67	5.67	6.67	2.33	10.00	1.00	1.33
Professional Development of Research Computing and Data Staff	12.33	9.00	3.33	5.00	7.33	1.33	10.00	1.00	1.33
Diversity, Equity, and Inclusion	11.33	8.00	3.33	4.67	6.67	1.67	9.00	1.00	1.33
Average of RCD Priority Point Values	14.6	10.4	4.2	6.2	8.5	2.7	10.8	1.4	2.1

Note that each "High Prio" counts as 2, and each "Med Prio" counts as 1, so these are not true counts of instances.

Appendix D: Complete Priorities Detail

Appendix D is only available to contributing institutions; it is not included in the public report.

Appendix E: Top Priorities by Demographics

In the tables below, the last column indicates where priorities duplicate (share) a priority with the list (i.e., Table 1 in the main report section) for all institutions (e.g., “Full Community #1”), is distinct among groups (“Distinct”), or is distinct from the list for all institutions but overlaps with one of the other demographic groups (e.g., “Distinct, common with Public”).

E.1: Top Priorities for Public Institutions

Capability[1]	Priority Points	Duplicate?
Do researchers have access to consulting and expertise to help them identify appropriate data repositories (on campus, in domains, and more generally) to place their data?	**18**	Full Community #3
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?	**17**	Full Community #1
Are Research Computing and Data services funded in a sustainable manner?	**17**	Full Community #2
Are researchers supported across the full research lifecycle?	**17**	Full Community #4
Does your Research Computing and Data (RCD) team/group have a strategic plan?	**17**	Full Community #5
Do researchers have access to advice on research compute and data compliance, security, management, and governance?	**16**	Full Community #7
Do researchers have access to tools/software that supports data backup, storage, and integrity checking?	**16**	Full Community #8
Do researchers have access to consulting and expertise on data wrangling/manipulation and data analysis?	**15**	Full Community #11
Are researchers made aware of research computing and data related resources?	**15**	Distinct
Do researchers have access to application support (training, help) for standard software packages, middleware, libraries, and modules?	**15**	Distinct
Do researchers have access to software that supports data wrangling/manipulation and data analysis?	**15**	Distinct
Are researchers effectively informed and made aware of Research Computing and Data (RCD) resources and services?	**15**	Distinct

Table E-1: Top Priorities for Public Institutions, and showing overlap with the full community priorities

E.2: Top Priorities for Private Institutions

Capability	Priority Points	Duplicate/ share?
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?	**8**	Full Community #1
Are Research Computing and Data services funded in a sustainable manner?	**8**	Full Community #2
To what extent is there a clear vision, effective guidance, and strategy for the allocation and prioritization of support resources/personnel?	**8**	Full Community #6
Does your institution have research data governance processes in place to establish data policies for research data?	**8**	Full Community #9
Do researcher-facing staff have the skills and capacity to broadly support researchers across levels (graduate students to PIs) and across domains with information about the use and	**8**	Full Community #10

[1] The capability titles throughout these tables are somewhat abbreviated - the full titles are in the assessment tool.

effectiveness of new technologies?		
Do researchers have access to compute and data environments to manage and use moderately sensitive data (e.g. NIH dbGaP data controls)?	**8**	Distinct
Do researchers have access to support, facilitation or training on how to compile, install, and deploy research software?	**8**	Distinct
Are there institutional resources for leveraging commercial cloud services for research computing and researchers?	**8**	Distinct
Do systems staff have the skills and capacity to support container deployment and orchestration (via APIs, kubernetes, docker, singularity)?	**8**	Distinct
Do researchers have access to interactive computing services? E.g., support for VDI, Gateways, JupyterHub.	**8**	Distinct

Table E-2: Top Priorities for Private Institutions, and showing overlap with the full community priorities

E.3: Top Priorities for R1 Institutions

Capability	Priority Points	Duplicate/ share?
Are Research Computing and Data services funded in a sustainable manner?	**20**	Full Community #2
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?	**18**	Full Community #1
Do researchers have access to consulting and expertise to help them identify appropriate data repositories (on campus, in domains, and more generally) to place their data?	**18**	Full Community #3
Does your Research Computing and Data (RCD) team/group have a strategic plan?	**18**	Full Community #5
Do researchers have access to consulting and expertise on data wrangling/manipulation and data analysis?	**17**	Full Community #11
Do researchers have access to tools/software that supports data backup, storage, and integrity checking?	**16**	Full Community #8
Are researchers made aware of research computing and data related resources?	**16**	Distinct, common with public
Are researchers effectively informed and made aware of Research Computing and Data (RCD) resources and services?	**16**	Distinct, common with public
Are researchers supported across the full research lifecycle?	**15**	Full Community #4
To what extent is there a clear vision, effective guidance, and strategy for the allocation and prioritization of support resources/personnel?	**15**	Full Community #6
Do researcher-facing staff have the skills and capacity to broadly support researchers across levels (graduate students to PIs) and across domains with information about the use and effectiveness of new technologies?	**15**	Full Community #10

Table E-3: Top Priorities for R1 Institutions,
and showing overlap with the full community priorities and other demographic groups

E.4: Top Priorities for R2 Institutions

Capability	Priority Points	Duplicate/ share?
Do researchers have access to a place to store final research data to address institutional policy and/or funding agency requirements?	**6**	*Distinct*
Are there institutional resources for leveraging commercial cloud services for research computing and researchers?	**5**	*Distinct, common with private*
Are researchers supported across the full research lifecycle?	**4**	Full Community #4
Do researchers have access to advice on research compute and data compliance, security, management, and governance?	**4**	Full Community #7
Do researchers have access to interactive computing services? E.g., support for VDI, Gateways, JupyterHub.	**4**	*Distinct, common with private*
Do researchers have access to high performance (batch) computing (HPC)?	**4**	*Distinct*
Do researchers have access to policies and technologies that facilitate management and wide access to data?	**4**	*Distinct*
Are Research Computing and Data services funded in a sustainable manner?	**3**	Full Community #2
Does your Research Computing and Data (RCD) team/group have a strategic plan?	**3**	Full Community #5
Is your Research Computing and Data (RCD) strategic plan aligned to campus plans?	**3**	*Distinct*
Do research funding activities actively integrate the Research Computing and Data (RCD) services group?	**3**	*Distinct*

Table E-4: Top Priorities for R2 Institutions, and showing overlap with the full community priorities and other demographic groups

E.5: Top Priorities for EPSCoR-eligible Institutions

Capability	Priority Points	Duplicate/ share?
Does your Research Computing and Data (RCD) team/group have a strategic plan? i. Is this strategic plan updated on a regular basis (e.g., annually, semi-annually)?	**14**	Full Community #5
Are Research Computing and Data services funded in a sustainable manner?	**12**	Full Community #2
Do researchers have access to consulting and expertise to help them identify appropriate data repositories (on campus, in domains, and more generally) to place their data?	**11**	Full Community #3
To what extent is there a clear vision, effective guidance, and strategy for the allocation and prioritization of support resources/personnel?	**11**	Full Community #6
Are researchers made aware of research computing and data related resources?	**11**	*Distinct, common with public*
Do researchers have access to guidance or training for cloud computing?	**11**	*Distinct*
Are researchers supported across the full research lifecycle?	**10**	Full Community #4
Does your institution have research data governance processes in place to establish data policies for research data?	**10**	Full Community #9

Do researchers have access to consulting and expertise on data wrangling/manipulation and data analysis?	**10**	Full Community #11
Do researchers have access to dedicated resources (e.g., staff) who can perform data wrangling/manipulation and data analysis?	**10**	*Distinct*

Table E-5: Top Priorities for EPSCoR-eligible Institutions, and showing overlap with the full community priorities and other demographic groups

E.6: Top Priorities for Minority Serving Institutions

Capability	Priority Points	Duplicate/ share?
Do researchers have access to software that supports data wrangling/manipulation and data analysis?	**7**	*Distinct, common with public*
Are Research Computing and Data services funded in a sustainable manner?	**6**	Full Community #2
Are researchers supported across the full research lifecycle?	**6**	Full Community #4
Does your Research Computing and Data (RCD) team/group have a strategic plan?	**6**	Full Community #5
To what extent is there a clear vision, effective guidance, and strategy for the allocation and prioritization of support resources/personnel?	**6**	Full Community #6
Does your institution have research data governance processes in place to establish data policies for research data?	**6**	Full Community #9
Do researchers have access to consulting and expertise on data wrangling/manipulation and data analysis?	**6**	Full Community #11
Are researchers made aware of research computing and data related resources?	**6**	*Distinct, common with public*
Do researchers have access to dedicated resources (e.g., staff) who can perform data wrangling/manipulation and data analysis?	**6**	*Distinct, common with EPSCoR*
Do researchers have access to guidance or training for cloud computing?	**6**	*Distinct, common with EPSCoR*

Table E-6: Top Priorities for Minority Serving Institutions, and showing overlap with the full community priorities and other demographic groups