Existing Initiatives' In-House Technical Capabilities

Victoria Smith and Jen Rosiere Reynolds
Carnegie Endowment for International Peace

June 7, 2022

Abstract

We drew on mapping exercises conducted by the Partnership for Countering Influence Operations[1] and Disinfo Cloud[2] to identify 84 initiatives whose work focused on analyzing or understanding the information environment. Upon finding contact information, we emailed 52 of these bodies asking recipients to answer four questions about their engineering and data science capabilities.[3] We received responses from 27 initiatives, eight of which[4] were also included in *A Survey of Public Oriented Organizations Analysing Social Media Disinformation.*[5]

We found that 41% of organizations did not have dedicated in-house data science support and 67% did not have engineering support. However, groups did report finding it slightly easier to obtain data science support than engineering support. Six respondents who did not have dedicated support reported that they could draw on capacity from within their larger organization or team to provide data analysis support, while only four respondents said they could do the same for engineering.

[1] Smith, Victoria. "Mapping Worldwide Initiative to Counter Influence Operations". *Carnegie Endowment for International Peace* (14 December 2020). https://carnegieendowment.org/2020/12/14/mapping-worldwide-initiatives-to-counter-influence-operations-pub-83435

[2] "Tracking Propaganda and Disinformation". *Disinfo Cloud* (2022). https://www.disinfocloud.com

[3] The in-house technical capabilities examined in this report do not represent the abilities of all initiaitves who work to understand the information environment. However, the capabilities examined provide insight into the current state of engineering and data science support for research on the information environment.

[4] The eight mentioned are: ASPI; ClemsonHub; CSMaP; DFRLab; Institute for Strategic Dialogue; Stanford Internet Observatory; University of Washington's Center for an Informed Public; and a company that requested anonymity.

[5] Linvill, Darren L., and Patrick L. Warren. "A Survey of Public Oriented Organisations Analysing Social-Media Disinformation". *Clemson University Media Forensics Hub* (2022). https://drive.google.com/file/d/10lQMwsYn3l-0yXg7g54g34sWZm7d3pFJ/view?usp=sharing

Funding: This research was generously supported by Microsoft, Craig Newmark Philanthropies, the John S. and James L. Knight Foundation, and the William and Flora Hewlett Foundation. This work was initiated as part of a joint project between the Partnership for Countering Influence Operations at the Carnegie Endowment for International Peace and the Empirical Studies of Conflict Project at Princeton University. **Author contributions:** VS collected all data and drafted the manuscript. JRR provided review, commentary, and revision. **Data and materials availability:** All data required to evaluate the conclusions are present in the paper or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Table of Contents

Table of Contents **1**

Executive Summary **2**

Introduction **2**

Methodology **3**

Overview of Survey Responses **4**

- **Engineering & Data Science Support 4**
- **Executing Quantitative Analysis 5**
- **Difficult to Access Technical and Engineering Enablers 6**
- **Overview of Initiatives 7**

Conclusion **7**

A.1 Codebook **9**

Executive Summary

1. We gathered publicly available information on the type and quantity of in-house engineering and data science support within 84 initiatives whose work focused on analyzing or researching the information environment.
2. We identified contact details for 52 initiatives and emailed them a brief survey. We received responses from 27 initiatives.
3. 16 respondents (59%) reported that they did not currently have dedicated in-house engineering capabilities. Two said they could draw on their team or wider organization for engineering when required.
4. 18 respondents (67%) reported that they did not currently have dedicated in-house data science support. Six said they could draw on data science contributions from their faculty, wider organization, or external partners and consultants when required.
5. 20 respondents reported access to quantitative analysis support. Of these, sixteen relied on support from their university faculty, including faculty staff, Ph.D. students, post-doctoral researchers, or from within their existing team, three used consultants or external partners,[6] and two could access support from their wider organization.
6. A lack of access to skilled personnel was the most frequently cited difficulty. This seemed primarily due to funding constraints. In general, restricted access to data and funding were the next most cited difficulties.
7. Development of infrastructure and new tools is expensive and can be difficult to justify when funding is restricted to short-term projects.
8. Of the 27 responses received, 44% came from academia, 41% from civil society, 11% from tech companies, and one from a government or intergovernmental initiative.
9. Three-quarters (74%) of the responses were from initiatives based in the United States. Two respondents were based in the UK, and one each was based in Australia, Belgium, Brazil, Canada, and Slovakia.

Introduction

This review of in-house technical capabilities is part of an effort to evaluate the potential of the Institute for Research on the Information Environment (IRIE). IRIE aims to develop a shared scientific infrastructure to support policy-relevant research on the information environment and its impact on democratic deliberation, politics, and public health. To assess the requirements for this new institute, we want to better understand existing institutions' capabilities and gaps better. To do this, we identified a subset of 84 organizations spanning academia, civil society, government, and tech whose work most closely aligned with the types of activities that IRIE may support or undertake. We solicited feedback from 52 of these identified organizations by emailing representatives a short survey.

Our responses illustrated a field that is quickly adapting to overcome difficult constraints, including a lack of access to data and funding and a high level of competition for skilled personnel. Some respondents described an increasing focus on qualitative, rather than quantitative, analysis–substituting requirements for bulk data collection and analysis with

[6] One respondent reported having access to support from within their wider team and paid consultants.

targeted interviews and manual content analysis. It is unclear whether this shift has occurred largely by choice or due to data analysis constraints.

Two of our responses were from tech companies with greater engineering and data science capabilities than most non-profit initiatives. They shared their responses on the condition that they would remain anonymous and the details of their team composition would not be made public.

Methodology

We drew on sources including the Partnership for Countering Influence Operations' 2020 initiative-mapping exercise[7] and Disinfo Cloud[8] to identify a subset of 84 organizations whose work focused on analyzing or understanding the information environment. We reviewed these organizations' websites to understand their geographic location, staffing levels, work focus, and points of contact. Information published about work and staffing levels varied significantly, so it was difficult to use this information to compare capabilities within and between initiatives.

Four of the organizations on our original list appeared to be no longer operational:

- MIT's Center for Civic Media closed at the end of 2020;
- Oxford University's Computational Propaganda Project's funding ended in 2021;
- Social Science One is closed, although it maintains a website; and
- Harmony Labs' Project Ratio now appears in Harmony Labs' archive of past work.

The extent of independent and dedicated resources was unclear for two of the 84 initiatives:

- The Ethics and Governance of Artificial Intelligence Initiative, a hybrid research effort and philanthropic fund run by MIT Media Lab and the Berkman Klein Center for Internet and Society at Harvard University; and
- NYC Media Lab, a consortium that fosters collaboration between universities and the private sector.

During this process, we used open-source research and our team's personal networks to identify contact details for 52 organizations. We emailed to ask recipients to answer the following survey questions:

1. Do you have dedicated engineering support? If so, approximately how many full-time employee (FTE) equivalents do you have?
2. Do you have dedicated data science support? If so, approximately how many FTE equivalents do you have?

[7] Smith, Victoria. "Mapping Worldwide Initiative to Counter Influence Operations". *Carnegie Endowment for International Peace* (14 December 2020). https://carnegieendowment.org/2020/12/14/mapping-worldwide-initiatives-to-counter-influence-operations-pub-83435

[8] "Tracking Propaganda and Disinformation". *Disinfo Cloud* (2022). https://www.disinfocloud.com

3. If you do not have dedicated support, how do you execute the quantitative analytic tasks necessary for your work (e.g., post-docs, graduate students, etc.)?
4. What are the key technical/engineering enablers you find hard to access (e.g., machine translation, personnel to work on APIs, etc.)?

In total, we received 28 responses.

Overview of Survey Responses

Engineering & Data Science Support

Of the 27 responses, 16 (59%) reported that they did not currently have dedicated in-house engineering support. However, of these, three could access engineering support if necessary.

- One noted that while they did not have dedicated engineering support, their team had a mix of skill sets so they could collectively build the infrastructure needed.
- One respondent said that their parent organization had engineers to provide additional support for specific projects.
- One organization could access support through external partners.

Another initiative that does not currently have any dedicated engineering support said there were plans to recruit two to three engineers in the medium term.

Eleven respondents reported some level of dedicated engineering support:

- One respondent had less than one full-time engineer.
- Five reported one full-time engineer. Of those five, one planned to hire a second; another said they could also draw on the support of external partners as required.
- One had two full-time engineers.
- Two had three dedicated full-time engineers.
- Of the two tech companies surveyed, one reported 14, and the other 35, full-time engineers on staff.

Regarding dedicated data science support, 18 (67%) organizations reported that they did not have dedicated support. However, of these:

- One hoped to recruit a data scientist in the short term.
- Five said that additional support could be found in their faculty or team if required.
- Two had access to data science skills in their wider organization. One could access support from external partners.

Six of these organizations that do not have dedicated data science support have developed ways to access it:

- One university reported hiring a data science consultant by the hour to fulfill specific tasks.
- One initiative could either hire a data science consultant or access support through an organizational partnership.
- Another initiative said they had an ongoing organizational partnership that provided them with support that varied on a project-by-project basis.

- A large organization said that while there was no dedicated support, the organization had about 10 full-time data scientists who were not linked to any specific projects but could provide support when required.
- One university said that all faculty members had some level of sophistication with statistics and data science.
- Another university relied on Ph.D. and postgraduate students to provide support as required.

Of the initiatives with in-house data scientists, the team size varied.

- Five entities reported the dedicated contributions of one full-time data scientist. One of these five organizations hoped to recruit a second. Another noted that they could also draw on the skills of existing team members. Finally, one initiative reported that their one data scientist spent "most of their time in meetings"; thus the organization only had "about 10% of a full-time data scientist." It is unclear whether this situation was because the organization does not need full-time support.
- One initiative shared that they had the support of two data scientists on their team.
- The tech companies had the most data science support. One company reported seven data scientists and the other three.
- Finally, two groups reported that their access to data science support varied according to the work they were undertaking. One organization explained that they paid computer scientists by the hour as needed. Another organization received data science support through a strategic partnership with a tech company on a project-by-project basis.

Executing Quantitative Analysis

Twenty-three respondents answered the question asking how their initiative executes the quantitative analysis tasks necessary for their work:

- Sixteen relied on support from their university faculty, including faculty staff, Ph.D. students, post-doctoral researchers, or from within their existing team.
- Three used external partners or consultants.[9]
- Two reported using staff from other areas of their wider organization.
- Finally, three respondents said they did not do quantitative analysis.

Most respondents did not specify exactly how many staff members with quantitative analysis skills they had access to. However, three were more specific:

- Stanford Internet Observatory said they had access to one technical postdoctoral fellow supplemented with part-time work by undergraduate and graduate student research assistants.
- CSMaP said they have six full-time postdocs with computational social science backgrounds and were, for the most part, able to perform the necessary quantitative analytics tasks.
- DFRLab said that each of the approximately 20 FTEs on their research team was capable of quantitative analysis. They added that "with very few exceptions, most (of these 20 FTEs) would categorize themselves as data journalists or online researchers. The vast

[9] One respondent reported access to support from within their team and from paid consultants

> majority of our quantitative analysis is conducted using third-party tools with varying access to platform APIs. On a few occasions, we've built our own tools. On more occasions, we've built or curated our own data sets for further analysis or comprehensive analysis on a given topic."

Difficult to Access Technical and Engineering Enablers

Many respondents described difficulty accessing more than one technical or engineering enabler. Access to skilled personnel was the most frequently cited, followed by data access and general funding constraints.

Personnel issues included a lack of:

- Personnel to work on APIs or develop alternatives to APIs.
- Personnel to develop tools to better analyze and visualize data.
- Personnel willing to take on repetitive, lower-skilled tasks.

Some of the reasons cited for personnel issues were:

- Challenges of recruiting and retaining skilled labour, given competitive salaries within the private sector.
- Funding limitations restricting the number of staff that can be recruited.
- Limited capacity to train the required number of students/staff.

The next most frequently cited difficulties were data access and general funding constraints, both cited by seven respondents. Data access constraints included:

- Reliance on limited data made available through second or third parties, such as via platform APIs.
- Not meeting the requirements for certain platform data.

In ongoing consultations in the counter-influence community since 2019, researchers have frequently cited data access as an issue. However, data access likely means different things to different researchers. Some, who can store and analyze bulk data at scale, want unrestricted access to large volumes of data. However, this access typically comes with high costs, including the funds and staff to build and maintain the infrastructure to store and process the data and the technologies to analyze it (such as natural language processing, image/video analysis, machine translation or data visualization). For other researchers, data access means access to better quality data than they currently have; this could be data that they can search and filter to narrow their sample size or structure to help better compare it against information from other platforms. Until researchers and platforms can find a way to resolve the tension between user privacy and data access, these issues are likely to remain largely unresolved.

The lack of funding affected some respondents' workload and staff recruitment and retention. Respondents also mentioned the difficulty of building and maintaining infrastructure in an environment where funding is project-based. Others cited the costs of specialized technologies, like machine translation, which they said can be prohibitively expensive.

Other resources cited as difficult to access were automated image analysis, cited by two respondents, and machine translation, video analytics, training data sets, standardized reporting for threat sharing, and natural language programming in central European languages, which were all cited once.

One respondent found no enablers difficult to access, while two explained that the question did not apply to their work. Two respondents emphasized qualitative research methods rather than quantitative analysis in their work. One group stressed the need to recognize the importance of qualitative approaches and that civil society researchers may need support in this area, including personnel to code, shared codebooks, or strategies to minimize the negative impacts of reviewing harmful content. CDT reported that their main consideration for using quantitative versus qualitative methods was implementing the best methods to address the chosen research question.

Finally, one respondent highlighted bureaucratic obstacles as a significant impediment to international cooperation on these issues, even among allies, but did not provide additional details.

Overview of Initiatives

Out of the 84 bodies originally identified, we classed 38% as academic, 43% civil society, 11% technology companies, 5% government or intergovernmental, and 2% media.

Of the 28 responses received, 44% came from academia, 41% civil society, 11% technology companies, and one response from a government or intergovernmental group. Three-quarters (74%) of the responses were from initiatives based in the United States. Two respondents were based in the UK, and one each were from Australia, Belgium, Brazil, Canada, and Slovakia.

Conclusion

Our research found that most of the initiatives that responded to the survey did not have in-house engineering or data science support. Still, it was easier for these groups to find data science, rather than engineering, support for ad-hoc requirements.

Respondents reported that financial restrictions also significantly impact a range of issues in the community. Competitive salaries in the private sector make staff recruitment and retention difficult. Development of infrastructure and new tools is expensive and can be difficult to justify when funding is restricted to short-term projects. Access to data, a frequent refrain among the counter-influence operations community, was also raised as an issue by several respondents in this survey. However, ongoing consultations led by the PCIO have found that different researchers can mean different things when describing data access; some want a larger quantity of data, while others want access to better quality data.

Two of the initiatives stated that they favor qualitative research questions rather than quantitative analysis. While one respondent said the choice between adopting quantitative or qualitative research methods was driven by the chosen research question, they did concede that quantitative research methods could incur additional costs related to the secure and ethical collection and retention of large volumes of data.

Appendix

A.1 Codebook

Variable	Description
Host Organisation	Name of the parent organization to which the initiative belongs (if applicable)–for example, the name of a university or think tank
Initiative Name	Name of the initiative: This could be an initiative in its own right, a university department, or project housed at a think tank
Organisation Type	Academia: A university department or project Civil Society: Think Tanks and other non-governmental organizations Government & Intergovernmental: Run by a local or national government or housed at an intergovernmental organization such as the UN or EU Tech: A company such as a platform or data analysis company
Country	Geographic location of the initiative
Engineering Support (FTEs)	Number of in-house full-time employees (FTEs), answers provided by survey respondents
Data Science Support (FTEs)	Number of in-house FTEs, answers provided by survey respondents
How do you execute the quantitative analytics tasks necessary for your work?	Answers provided by survey respondents
What are the key technical/engineering enablers you find hard to access	Answers provided by survey respondents
Additional Information	Answers provided by survey respondents