2023-09-15

Emerging Trends in Research - Survey Findings

Felipe Pérez-Jvostov, Seppo Sahrakorpi, and Qian Zhang

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

Table of Contents

1. Introduction

Over the past three years the Digital Research Alliance of Canada (the Alliance) has extensively consulted the research community to gather their perspectives on the current state of digital research infrastructure (DRI) in Canada, its challenges and opportunities. This input has been instrumental in shaping the organizational priorities and investments for the period 2023-2025.

To plan towards the years 2025-2030 and to identify future opportunities not previously recognized, the Alliance launched a call for input on Emerging Trends in Research that may impact how researchers interact and use DRI. The survey call was aimed at researchers across sectors and was open from July 7 to August 18, 2023. A total of 51 researchers (47 participants in English and 4 in French) responded to the following questions:

1. *What are the emerging research trends in your discipline that you believe will have a significant impact on how you use digital research infrastructure (DRI)?*

2. *How do you envision researchers in your field will interact with DRI in 10 years, considering the emerging research trends, future technologies and policy advancements?*

3. *Are there any concerns or technical challenges that may preclude Canadian researchers in your discipline in partaking in cutting edge research?*

4. *Is there anything that you believe will influence DRI and research in general that you haven't shared with us during our previous consultations or in the questions above?*

Summary key findings are discussed below and organized across eight themes: Big Data; compute and storage at scale; Artificial Intelligence; integrated digital research infrastructure; equity diversity, inclusion, and accessibility; data security, open policies, and governance; and quantum, with an additional section on key concerns.

2. Findings

2.1. Big Data

Researchers discussed how research data is rapidly increasing in size and complexity–a trend that will continue for the foreseeable future. Many researchers alluded to the future need for compute and storage to be much higher than what is currently available, as well as robust research software to analyze such large datasets. High availability and reliability of storage instances was also commonly discussed in the context of large analyses.

Many researchers mentioned that manual data management will be impossible when dealing with vast amounts of data. Automation–and artificial intelligence (AI)-powered tools will likely be needed to support researchers in research data management-related tasks. Similarly, AI was often referred to as a tool needed to facilitate the integration of large, heterogenous datasets and to increase data interoperability.

Open data and open access were also commonly referred to. As Open Science is rapidly becoming adopted across disciplines, many researchers acknowledged the need for broader recognition and use of discipline-specific data standards. Researchers working with regulated or sensitive data (e.g., protected health information, PHI; or personally identifiable information, PII) further emphasized that future DRI must enable secure and compliant sharing of well curated and annotated data in order to be able to capitalize on the large amounts of Canadian health data and tackle increasingly complex challenges in medical research.

"…*the world community of BioImaging scientists is facing the challenges of very large datasets. Their management and analyses require automation and cloud computing*…"

"…*storage and compute requirements in genomic continue to grow exponentially…"*

"…*high performance computing for particle physics needs networked storage on a global scale. Storage instances need to be of high availability, reliability, and performance*…"

2.2. Compute and storage at scale

Many researchers highlighted that compute and storage resources will need to be more capable to deal with Big Data. Rapid growth and the need for active storage and high-performance computing (HPC) resources were also frequently mentioned. The growth in demand for compute and storage was estimated to range from two- to ten-fold during this decade.

Broad reference to different computing architectures was commonplace, with mentions of CPU, GPU, and memory intensive computing needs. Some researchers explicitly mentioned the need for exascale computing capability or high-throughput computing (HTC), likely indicating that these technical terms are not commonly referred to by DRI end-users.

"…*there remains and will continue to remain large demand for traditional HPC resources, with CPU cores and memory being the key ingredients to large numbers of important fundamental and applied problems in physical sciences*…"

"…*computational resources required to conduct cutting edge research will be extremely demanding*…"

"…*we need data storage and backup infrastructure capable of handling these next-generation data problems…high-capacity/throughput computing needs; high-capacity data storage and backup requirements…"*

2.3. Artificial Intelligence

AI and related technologies (e.g., machine learning, ML; and large language models, LLMs) were seen as key future trends for advancing research. Many researchers highlighted and discussed how AI technologies will be important tools and enablers of research in their fields of research. The respondents did not discuss developing AI methodologies or training AI models per se, rather they focused on how AI-based tools could be applied to help them in doing their research.

Some specific application examples of AI in research were provided, for example for image and sound analysis, and (large) dataset analysis. Other potential use cases included automated translation, transcription of handwriting, natural language processing, AI-aided and guided

simulations, and applied problem solving. AI was frequently mentioned as relevant in health and medical fields, and in genomics. Particle physics, materials sciences, and biomechanics were also mentioned as fields where AI could help advance and accelerate science. AI assisted analysis of medical images and data was highlighted by multiple researchers, indicating a high priority deliverable in the cross section of AI technologies and fields of research.

"…*the use of AI will become increasingly important in data analysis and representation methods*…"

"…*obviously, ML/AI will be a major factor for genomic research as it will be in other fields*…"

"…*the amount of [diagnostic medical] data received back from these tools and that usage of said data with early AI tools is solving decades old problems*…"

2.4. Integrated Digital Research Infrastructure

Researchers envisioned that in future having integrated DRI (data + compute + software) systems will be key. Researchers wanted to easily access resources via comprehensive interoperable platforms rather than separately for each individual component. Potential integrated solutions mentioned included an integrated domain-specific platform, a common open data repository integrated with large-scale analysis and compute capability, and an accessible Software-as-a-Service (SaaS) cloud-based instance that facilitates both analysis and collaboration.

"…we are hoping to see a much more integrated data ecosystem spanning research and clinical data.... Overall, I'm also really hoping that DRAC will realize the potential for DRI integration across infrastructure/software/data. We don't need resources separately in these 3 areas, we need domain-specific platforms spanning all 3…"

"…absolutely, researchers are increasingly expecting integrations between systems...."

"…the ideal would be significant integration potential (and interoperability) across different platforms that hold / govern varied sources of data...."

2.5. Equity, Diversity, Inclusion, and Accessibility

Researchers in the Humanities identified important gaps in accessibility to DRI that must be addressed in the future. Issues related to inequities in funding allocation for infrastructural development and operational resources between disciplines were highlighted as barriers for broader access and use of DRI. Others emphasized that accessibility would be improved by adopting easy-to-use interfaces and software.

Researchers working with marginalized groups, visible minorities, and Indigenous Peoples discussed the ethical considerations of using data about such groups, emphasizing the need to find a balance between Indigenous data sovereignty and the protection of privacy with open scholarship and data linkage, particularly in the context of AI and Big Data.

"…*DRI in Canada must not leave any disciplines or communities behind*…"

"…I hope that user interfaces become more accessible to researchers without strong computer backgrounds…"

"…*another important technical challenge is enabling Indigenous data sovereignty on existing platforms…It is important to strike a balance between as much open scholarship and connectivity as possible, and the need to protect the privacy of Canadian scholars and respect Indigenous data sovereignty*…"

2.6. Data security, (open) policies, and governance

Many researchers emphasized the importance of the Open Science movement and the value of linked and open data, while at the same time recognizing the importance of data security and privacy. Many researchers raised the problem around "as open as possible and as closed as necessary" to foster data sharing, discovery, access, linkage, and reusability in order to accelerate research, while safeguarding the intellectual property (IP) rights, and privacy of the subjects (particularly PHI-compliant data). Another important aspect mentioned was the proper

federation, storing, retention, and curation of long-term data. On the policy front the development and adoption of best practices was also highlighted, while on the legal front the difficulties arising from jurisdictional differences in PHI and data legislation were a concern and indicated as an impediment for effective interjurisdictional research.

"…*data transfer agreements [are] needed to send the [health] images. The major stumbling block has been the part of the agreement about governing law and jurisdiction. Both sides want their own law and jurisdiction in the agreement*…"

"…*the open scholarship movement, and the positive engagement it promises, has been foundational for my recent work*…"

"…*there is increasing interest in the ability to federate data, and take analysis to data rather than moving data to a single location…while data federation is a clear future direction, it will not address all use cases in the health and health-related research space. There will likely still be a need for DRI that can host pan-Canadian platforms and approved data sharing, including data that are not collected by or governed by researchers / Universities*…"

2.7. Quantum

Only a few researchers mentioned quantum computing as a potential future trend in their research. It is well acknowledged that quantum computing has potential to drastically change the DRI landscape once the technology and applications are more mature and better understood. Given the early R&D phase of quantum computing, it is understandable that the research community does not yet have a clear vision for how to leverage such technology in future.

A potential use case, however, was how quantum computing, through increased computing power, will have potential to allow researchers who are not experts in theory and computing to execute complex simulations. Similarly, other examples of future applications were simulations of quantum theory-based models, e.g., for running quantum theory-based particle physics models, and pattern recognition of complex experimental signatures.

"…*quantum computing has the potential to change some parts of particle physics research very dramatically since most of the CPU-time we require is for the simulation of quantum processes and pattern recognition of complex detector signatures where large-scale availability of quantum processors could completely revolutionize the way we work*…"

2.8. Miscellaneous

Many researchers highlighted issues that touched on various aspects of DRI. Among them, were issues related to easy-to-use infrastructure that would use various technologies to facilitate accessibility, use, and productivity. Others highlighted the need for better integration of existing systems in the future, but that this would continue to be at a minimum-level of support. For Canada and Canadian researchers take full advantage of modern and future technologies, there must be an important increase in the level of funding to provide more robust, integrated, secure and automated infrastructure that is tailored to researchers with various levels of technological proficiency and from different fields. Finally, some researchers emphasized that future research will continue to be more international, and Canadian infrastructure must facilitate such collaborations while protecting Canadian intellectual property and interests.

"...new technologies are ultimately less important than smart deployment of demonstrated technologies. ... Our primary mode shift will be a return to requiring data and processing being located in single facilities, i.e., the server model, which is in contrast with our current model where researchers can use their local computing to manage their data. As such, well established data centres with discipline-adapted science platforms are going to be required to meet research goals…"

"…hiding the complexity of the DRI ecosystem and letting researcher do their research (not be tech experts) will be critical to moving research forward. There is an enormous gap between supporting the bare minimum in DRI (what we do today - keep the systems running, software is installed, knock your self out) and enabling researchers (by removing technical barriers) to excel at their science."

2.9. Key Concerns

Researchers shared several concerns around the future of DRI: security and privacy concerns, lack of DRI funding, highly-qualified personnel (HQP) staffing levels, and concerns around diversity and services for Humanities and Social Sciences (HSS).

The lead concern among the respondents was how security and privacy-related restrictions could impede advances in DRI and research in general. Multiple researchers gave examples of instances where privacy issues and jurisdictional incompatibilities (e.g., EU's GDPR, or variety of PHI regulations) had already blocked or slowed down research activities. While multi-factor

authentication was acknowledged to increase security, it was found to be cumbersome, or even preventing international collaboration. Another leading concern was the lack of or insufficient funding for DRI in Canada. This included concerns regarding funding of either the current advanced research computing (ARC) refresh cycle, or funding of future growth in DRI infrastructure. Multiple respondents were also concerned about potential lack of diverse types of infrastructure going forward, emphasizing the need for custom infrastructures, workflows and technical expertise in underserved disciplines. The concern of insufficient number of HQP resources was also raised, since some use cases will require substantial additional HQP resources–particularly in the case of HSS.

In addition to the above main themes, the researchers were also concerned about costs of using DRI, bureaucracy of resource allocation process, difficulties in data reuse, too much focus on HPC, lack of long-term storage, and influence, hacking and IP theft by foreign adversaries while working towards open science with friendly nations and institutions.

"*…there is no support for the mass of researchers. Money and infrastructure is lacking*…."

"*…new laws and regulations, for example personal data protection, are being rapidly (and with good intentions), leaving researchers/institutions without the required expertise or resources to completely re-structure how research is done*…"

"*…resources for HSS scholarship and infrastructure have always been limited compared to funding for STEM…HSS infrastructure is more likely to be human-based infrastructure requiring technical expertise…these projects face challenges competing with well-funded industries that can offer higher salaries to the types of people who would be strong candidates for these roles*…"

3. Conclusions

Respondents shared a vision of a future Canadian DRI that is more robust, integrated, secure and automated. One that facilitates research collaborations across borders, while protecting data and the rights of data owners. The future Canadian DRI should take full advantage of modern and emerging technological advances in automation and Artificial Intelligence, and substantially increased HQP staffing, to make research more efficient and connected, ensuring research and research infrastructure funded through public funds is trustworthy and can have impacts beyond scholarly research.

Many researchers highlighted the need for additional compute and storage resources, estimating the growth to range from two- to ten-fold during this decade. Many researchers expressed the need for a variety of types of compute infrastructure (i.e., CPU, GPU, and memory intensive architectures). Researchers recognized that research data is rapidly increasing in size and complexity. Many alluded to the future need for compute and storage much larger than what is currently available, as well as robust research software to analyze large datasets. High availability and reliability of storage instances was also commonly discussed in the context of large analyses.

AI and related (e.g., ML, LLMs) technologies were seen as key future trends for advancing research by many researchers, highlighting how AI technologies will be important tools and key enablers for research. Manual data management will be impossible when dealing with vast amounts of data. Automation and AI-powered tools will likely be needed to support researchers in research data management–related tasks. Similarly, AI was often referred to as a tool needed to facilitate the integration of large, heterogenous datasets and increase data interoperability.

Open data and open access were also commonly referred to. As Open Science is rapidly becoming adopted across disciplines, many researchers acknowledged the need for broader recognition and use of discipline-specific data standards. Many respondents emphasized the importance of open science movement and value of linked, open data while at the same time recognizing the importance of data security and privacy. The problem of how to make data “as open as possible and as closed as necessary” to foster sharing, discovery, access, linkage, and reusability and accelerate the research, while safeguarding the IP rights, and privacy of the subjects (particularly PHI-compliant data) was raised in many responses.

Researchers in the Humanities identified important gaps in accessibility to DRI that must be addressed in the future. Issues related to inequities in funding allocation for infrastructural development and operational resources between disciplines were highlighted as barriers for broader access and use of DRI. Others emphasized that accessibility would be improved by adopting easy-to-use interfaces and software.

Researchers working with marginalized groups, visible minorities, and Indigenous Peoples discussed the ethical considerations of using data about such groups, emphasizing the need to find a balance between Indigenous data sovereignty and the protection of privacy with open scholarship and data linkage, particularly in the context of AI and Big Data.

Notably only a few respondents discussed quantum computing as a potential trend for their research in the midterm, indicating the early development stage of this potentially paradigm shifting technology.

Concerns over security and privacy-related restrictions impeding research were raised. Multiple respondents gave examples of instances where privacy policies, jurisdictional incompatibilities, and multi-factor authentication had already blocked or slowed down research activities and international collaboration. Another leading concern was the insufficient funding for DRI in Canada. This included concerns regarding funding of either the current ARC refresh cycle, or sustained funding of sufficient future growth in DRI infrastructure.