Working Paper on Digital Research Infrastructure (DRI)

Transforming scientific knowledge production capabilities and practices by digitally enabling research communities and communication: Virtualisation, visibility, visualisation and valorisation

Prepared for the National Research Foundation (NRF)

Luci Abrahams and Mark Burke

LINK Centre, University of the Witwatersrand, Johannesburg

May 2023

Funded by

Authors

Luci Abrahams

Luci Abrahams (PhD, Wits) is Director of the LINK Centre at Wits University https://www.wits.ac.za/linkcentre/, building research on digital innovation, how digital technologies and processes influence change, and on capabilities-oriented digital transformation. Luci is Corresponding Editor for The African Journal of Information and Communication, indexed in the SciELO Citation Index. She currently serves on the Board of TENET. She has previously served on various Boards and Committees, including the Board of the National Research Foundation and the Ministerial Review Panel on the Science Technology and Innovation Institutional Landscape.

Mark Burke

LINK Research Associate Mark Burke's research focus areas include the drivers and dynamics of digital government and the frameworks, methods and measures for monitoring digital government performance and evaluating its outcomes. He is particularly interested in the impact of digital governance on inclusive economic development processes. Burke has expertise in strategic planning; organisation design and development; programme design; and capacity development.

LINK Centre, University of the Witwatersrand, Johannesburg

The LINK Centre https://www.wits.ac.za/linkcentre/ is an interdisciplinary academic hub based at the University of the Witwatersrand (Wits) School of Literature, Language and Media (SLLM) and the Wits Tshimologong Digital Innovation Precinct. The LINK team teaches, researches and advises on a wide range of digital innovation and digital transformation matters in South African, African and other global south contexts.

This work is based on the research supported wholly by the National Research Foundation of South Africa (**Grant Number UID 149715**).

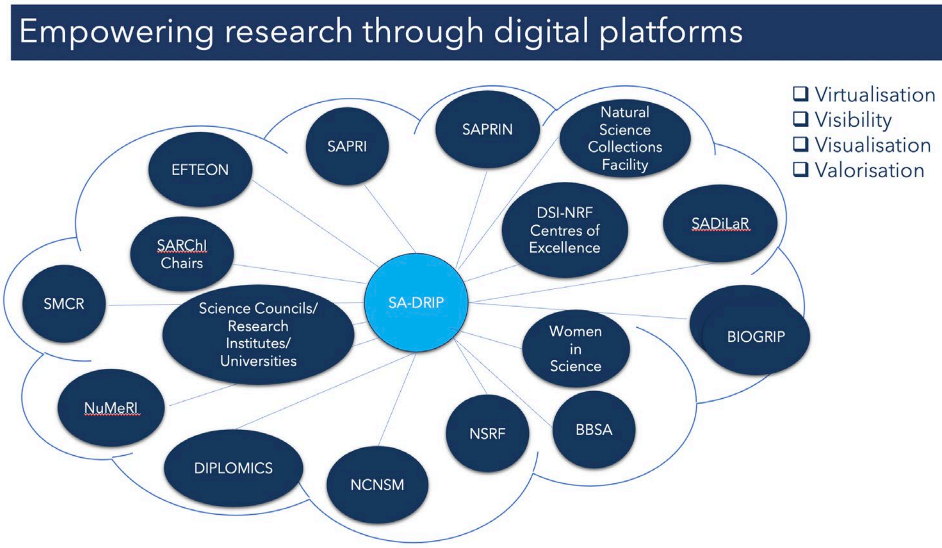
1. Introduction, problem and proposal: Transitioning to/experimenting with digital/data infrastructures for research and innovation

This working paper sets out the logic, argument and considerations with respect to creating digital research infrastructure (DRI), building on the intentions and focus of the South African Research Infrastructure Roadmap (SARIR) (DST, 2016). In the past ten years since adoption of SARIR, those research infrastructures have evolved and progressed. Missing from the landscape is digital research infrastructure, a necessary foundation for advancing commercialisable innovation, the practice of open science and social benefits from innovation. At present, the science and innovation landscape is characterised by highly variable research management capabilities, ranging from early stage digitalisation in some disciplines and fields, to the emergence of digital datasets and databases supported by infrastructure in the nascent stage of development, to established portals and data platforms in a limited number of disciplines and fields. Investments in, and efforts to develop, strengthen and expand such infrastructures are fragmented and isolated, losing out on opportunities for scaling up and accelerating digitally supported research, through the consolidation and concentration of resources.

While the proposed design for this digital research infrastructure draws on a landscape review that primarily sets out the work of the CoEs, RIs and SARChI Chairs (Abrahams & Burke, 2023), the digital research platform is applicable to all prospective creators and users operating in the publicly funded science system, including to those private sector research entities operating at the many points of intersection between the publicly funded and academically funded parts of the national system of innovation (NSI) on the one hand, and private sector R&D on the other hand, see schematic view in Figure 1.

Figure 1

Creating a knowledge cloud through digital platforms


Empowering research through digital platforms
❑ Virtualisation
❑ Visibility
❑ Visualisation
❑ Valorisation
SAPRI
SAPRIN
Natural Science Collections Facility
EFTEON
DSI-NRF Centres of Excellence
SADiLaR
SARChI Chairs
SMCR
SA-DRIP
Science Councils/ Research Institutes/ Universities
Women in Science
BIOGRIP
NuMeRI
NSRF
BBSA
DIPLOMICS
NCNSM

Note. Authors.

Furthermore, bridging digital research infrastructure with other forms of national research infrastructure can take South Africa into fields of science, technology and innovation where it was not previously a player, promoting virtualisation, visibility, visualisation and valorisation, what we term "the four V's". In this paper, virtualisation refers to making artefacts, data, records, published and unpublished work, as well as research instruments, tools and virtual laboratories, available and easily accessible online, in public and in secure user-only access formats. Visibility, including visibility within particular research communities, and more broadly, arises from the practice of making knowledge available in a range of digital formats. Visualisation is made possible through the application of visual design software to any form of artefacts, data or publications, thereby increasing the capacity of researchers to make their work understandable and increasing the capacity of readers to interpret and build on the work. Valorisation is the desired end state of the many processes that apply to making research more widely valuable, beyond the research producer/creator. Each of the four V's alone, and collectively, brings powerful capacities into the science system. Digital research infrastructure can provide the basis for transforming current scientific knowledge production capabilities and practices, by shifting from relatively limited knowledge sharing to extensive data and methodological sharing, thereby empowering the scientific community in many ways, including in applying new methods and techniques of scientific discovery, in generating new research questions, and in answering old and new questions in ways not previously possible.

The full research report (Abrahams and Burke, 2023) provides a framework for decision-making to guide the approach to be adopted in the design, development and application of South Africa's proposed digital research infrastructure platform (SA-DRIP). The deployment of this infrastructure is aimed at strengthening data analytics, virtualisation of research processes and outputs, as well as research management capabilities, across the DSI-NRF centres of excellence (CoEs), the SARIR research infrastructures (RIs) and the complementary research and innovation institutions in the science and innovation landscape, including research chairs, university-based research entities and other science producing entities. The report sets out an approach to the creation of this national DRI, based on a review of the existing research landscape, as the basis for digitally enabling scientific knowledge production, its application and management in South Africa, with attention to women in science and science inclusiveness.

Noting the White Paper (DSI, 2019), the STI priorities and grand challenges, including the push for commercialisable innovation and social innovation; noting also the DSI National Open Science Policy (draft); as well as various international positions on open science, including the UNESCO Recommendation on Open Science (UNESCO, 2021); and the work of the African Open Science Platform (AOSP, no date), the report proposes the establishment of a research infrastructure in the form of a digital platform (born digital), namely the South African digital research infrastructure (SA-DRIP), founded in the mathematical, statistical, computational and data sciences, in order to consolidate the investments and concentrate the necessary human resources capabilities in a lead consortium. The motivation for this proposition is that data science is critical to the long-term sustainability of a research data platform serving multiple research and innovation domains, while the mathematical, statistical and computational sciences are key skills integrated with data science practice and future innovation.

The SA-DRIP must seek to accomplish two key missions. On the one hand, it must establish the necessary capabilities to digitally enable the research capacities of established RIs and other institutions in the science and innovation landscape in South Africa (outward facing mission). On the other hand, it must set out and implement a long-term research and innovation agenda focused on the development of software applications and other digital technologies (inward facing capabilities-oriented mission) for enabling South African research to address the social, economic and environmental challenges that the NSI institutions find to be relevant. Functioning as a digital platform, this DRI is expected to confer advantages of scale economies and standardisation of good data enabled research and innovation practices. This DRI must support virtualisation of science producing activities, promote the visibility of research, and support the visualisation of data, ultimately, to ensure the production of public value from public funding and to promote greater public-private research collaboration. Most importantly, the SA-DRIP does not need to start on a clean slate, as many components of digital research infrastructure already exist and can be aggregated.

2. Methodology applied

The inception discussion indicated a few broad principles and criteria to guide the work, namely:

- Moving beyond the idea of "national equipment" to digitally enabled platforms which constitute national scale digital infrastructure in usage, depth, complexity and scope
- Value/benefits realisation: Deploying digital research infrastructure for gathering, storing, sharing and using research data, content and resources, as well as showcasing evidence that shapes policy, evidence that shapes practice, thus generating public value
- Free-to-use through open access approaches: Data, publications and other research-based knowledge resources that are already open access are the logical starting point for platformisation of publicly funded research present the opportunity to showcase the value of open data and how it promotes open science
- Pay-to-use components: Such an approach would require specific pay-to-use criteria and modalities for use, as well as a combination of public and private investment, for example research on virtualisation in manufacturing where there is commercialisation potential

The methodology (i) refines the problem statement; (ii) presents a literature review and analytical framework; (iii) includes a broad mapping exercise of key concentrations of research production and public funding; (iv) designed and applies a discipline-relevance/value framework to identify key research and innovation domains to commence the platformisation process; (v) includes a scoping exercise based on these six research domains and a selection of cross-cutting themes as the basis for a Phase 1 intervention; (vi) presents a high-level platform design; and (vii) sets out the institutional roles for creating the platform, as the basis for a consultative process. The methodology included a limited exercise in bibliometric analysis, highlighting the levels of research activeness in two of the six domains, and emphasising the need for incorporating bibliometric and scientometric analysis into project design and operationalisation, in order to inform future work on digital research infrastructures and data platforms.

3. Synopsis of observations prevalent in literature: Drivers and layers for data platforms

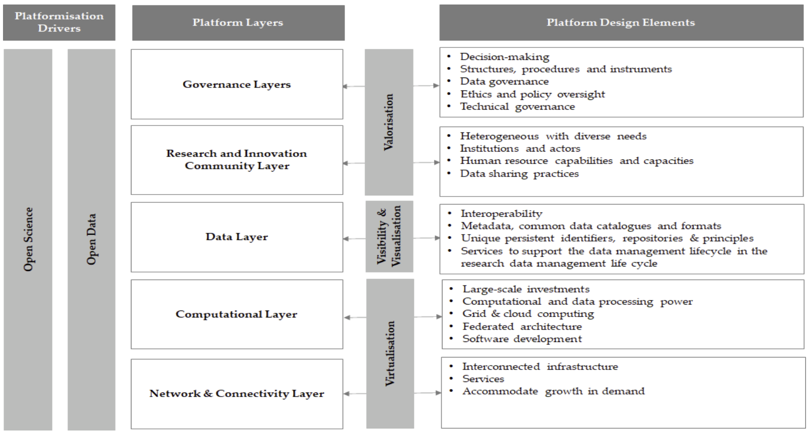
The drivers of digital platformisation in the globalised research and innovation environment are (i) the greater scope for commercialisable innovation, as well as (ii) the trends in open science and open data. This is because the principles and values of openness contribute to the sharing of data and knowledge, which accelerates science production and the application of knowledge, noting that in the 21st century, the volume of research and innovation output in a particular domain is too great for any single institution or entity to optimise value on its own. The value of data is very often in its reuse, for research and innovation purposes that were not initially foreseen, whether commercial or social.

Typically, digital platform layers (layered from the bottom up) are the network and connectivity layer (connecting instruments, researchers, fields of study and enabling data transfer, as well as communication and virtualisation, thus making scaling possible), the computational layer (where data is structured, organised, interpreted, analysed; where the computational and data processing power sits, often using cloud and grid computing and software-as-a-service), the data layer (where structured data is presented and made visible to the research community and in some cases is publicly available; promotes design of metadata and common data catalogues and formats; promotes design of unique identifiers; provides services to support the data management lifecycle), the research and innovation community layer (where researchers/scientists/innovators/ research institutions engage with each other, perform research and collaborate in multiple ways, and form communities of knowledge or practice; the layer where data scientists and disciplinary or interdisciplinary scientists interact; where publications and knowledge resources are available in multiple formats including text, statistics, maps, design libraries, other), and the governance layer (decision-making, manuals and procedures, data governance, ethics and oversight, rule-making, standards setting, other).

The literature review sought to address the question: How are digital infrastructures transforming knowledge production capabilities and practices by reshaping research communities and communication? The review and analysis of the literature resulted in the formulation of the analytical framework, see Figure 2, which is discussed briefly below.

Figure 2

Analytical Framework Diagram


Platformisation Drivers
Platform Layers
Platform Design Elements
Open Science
Open Data
Governance Layers
Research and Innovation Community Layer
Data Layer
Computational Layer
Network & Connectivity Layer
Valorisation
Visibility & Visualisation
Virtualisation
• Decision-making
• Structures, procedures and instruments
• Data governance
• Ethics and policy oversight
• Technical governance
• Heterogeneous with diverse needs
• Institutions and actors
• Human resource capabilities and capacities
• Data sharing practices
• Interoperability
• Metadata, common data catalogues and formats
• Unique persistent identifiers, repositories & principles
• Services to support the data management lifecycle in the research data management life cycle
• Large-scale investments
• Computational and data processing power
• Grid & cloud computing
• Federated architecture
• Software development
• Interconnected infrastructure
• Services
• Accommodate growth in demand

Note. Authors.

3.1 Research infrastructures for enabling science and innovation

Contemporary science and innovation are data-based and data-intensive, characterised by exponentially increasing volumes of data generated through observation, experimentation and simulation (Critchlow & van Dam, 2013). Digitally enabled international and national research infrastructures (RIs) have become critical to processing, storing and curating these large volumes of data as the research cycle and innovation processes have become increasingly digitalised. Research infrastructure is expressed (in the context of science in Europe) as (ESFRI, 2018, p. 11):

> [f]acilities, resources or services of a unique nature, identified by European research communities to conduct and to support top-level research activities in their domains. They include: major scientific equipment – or sets of instruments; knowledge-based resources like collections, archives and scientific data; e-Infrastructures, such as data and computing systems and communication networks; and any other tools that are essential to achieve excellence in research and innovation.

The early development of RIs was characterised by *ad hoc* development in an uncoordinated manner, covering a range of institutionally, geographically or scientifically determined domains and supported by a mix of sponsors, data providers and users (OECD, 2017a). Over the past two decades, the RI landscape in Europe, in particular, has been evolving towards a *consolidated ecosystem* (ESFRI, 2020).

3.2 New scientific and innovation possibilities through open science and open data

The emergence of *open science*, over the past two decades, is anticipated to open new avenues for addressing the manifold global challenges facing humanity (Cudennec et al., 2022). Open science is premised on the notion that "good data enables good science, and digital technologies provide the means for acquiring, transmitting, storing and analysing and reusing massive volumes of data" (Lipton, 2020, p. 16). *Open data* is viewed as fundamental to open science (Gabrielsen, 2020) and represents a necessary condition for reproducibility and scientific progress (Burgelman et al., 2019). Open data speeds up the research process and innovation (Borgerud et al., 2020), gives credit to data creators (Burgelman, 2019), and enhances transparency and accountability (Gabrielsen, 2020). Without the necessary data infrastructure on which open science depends, the sustainability of open data remains an "open question" (Paic, 2021).

3.3 The emergence of a digital research infrastructure ecosystem

The role of digital technologies in enabling the evolution of research infrastructure to meet the demands of contemporary science and innovation has been critical. This is evident in the role of these technologies in the different layers that comprise the digital research infrastructure ecosystem.

3.3.1 The network and connectivity layer

National Research and Education Networks (NRENs) are the main vehicles for connecting research communities across the globe (RISCAPE, 2019). More than 120 NRENs have been established around the world (Foley, 2016). These networks are diverse and comprise a broad range of infrastructure and communications technologies (GÉANT, 2022). Typically, NRENs provide the following *categories of services*: network and connectivity; network management; performance and analytics; trust, identity and security, cloud services and applications; real-time communication and multimedia (Abramov, 2021; Foley, 2016).

3.3.2 The computational layer

High-performance computing (HPC) refers to supercomputing facilities used as a critical tool in fields such as climate research, numerical weather prediction, particle physics and astrophysics, earth sciences and chemistry, and has recently become a cornerstone for most scientific fields ranging from biology, life sciences and health, energy, geosciences, material sciences, to social sciences and humanities (PRACE, 2018).

The growth in demand for HPC resources has led to the initiation of several global *large-scale investments and initiatives* (ESFRI, 2020). There is a shift away from traditional supercomputing dedicated to handling computationally intensive tasks with a focus on computational performance, programmability, scalability and energy efficiency towards pursuing *both computational and data processing power* (PRACE, 2018). The emergence of *cloud computing* offers a new paradigm in which researchers and institutions do not have to maintain physical infrastructure but rather acquire infrastructure services from dedicated providers (RISCAPE, 2019). The key features of cloud-based computational environments are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured services (Yang et al., 2014).

Software is critical to navigating the processes of data collection, storage and analysis and building and testing models in the research and innovation journey (Carver et al., 2022). It controls the instrumentation to record data and is invaluable to making research infrastructures work (UKRI, 2020). There is limited direct support for software development despite the recognition of the significance of software in reproducibility and replication (Carver et al., 2022), an area that will require attention in the immediate future.

The continued digitalisation of the research and innovation cycle through digital research infrastructures has contributed to increasing *virtualisation* and the establishment of virtual research environments (VREs) (Connaway & Dickey 2010). Virtualisation is a method of deploying computing resources (Jiang et al., 2020) aimed at efficient computing resource utilisation through the creation and operation of virtual versions of servers, storage devices, networks and operating systems (Radchenko et al., 2019).

3.3.3 The data layer

Data-intensive science and innovation works with large, heterogeneous, distributed data that requires specialised research and data infrastructure for its collection, storage, processing and visualisation (Otto et al., 2022). *Research data management* (RDM) is at the heart of the research process since the accuracy and validity of data bears directly on the conclusions to be drawn, promotes reproducibility and replicability, and establishes the methods to obtain, handle, transfer and archive the data (Zozus, 2017). The *data management life cycle* is a subset of steps structured by the data lifecycle, which is tied to the research lifecycle (Bratt, 2022).

Laborious *data preparation* and *cleaning* are ways of imposing "order and intelligibility on a dataset" (Boumans & Leonelli, 2020, p. 94). A key challenge to address in this process is the *harmonisation* of data through a *standard schema* (Broeder et al., 2017). The role of *metadata* is critical in this regard since it is fundamental for data organisation (Trumpy et al., 2015), requires the establishment of *common data catalogues* (Bailo et al., 2017), and harmonisation of metadata to promote *interoperability* across scientific domains (Kindade & Sheperd, 2022).

Research data curation facilitates discovery, retrieval quality and value management and supports long-term availability and reusability (Lee & Stvilia, 2017). The development of *data management plans* (DMPs) and practices, driven by the funding agencies as a requirement, is an evolving approach in data-intensive research to support good data management (EU, 2020). DMPs are expected to encourage researchers to reflect on their data management practices (Devriendt, et al., 2022). *Data storage* and *archiving* play a crucial role in making data accessible (EU, 2020). *Research data repositories* are an essential part of the research infrastructure of open science (OECD, 2017). The development of the *FAIR Data Principles* serves as a guide to assist data stewards and publishers with evaluating choices for rendering digital research artefacts **f**indable, **a**ccessible, **i**nteroperable, and **r**eusable (Wilkinson et al., 2016).

Making data findable in terms of the first requirement of the implementation of the FAIR principles means that it needs to be visible. *Visibility* and accessibility of data can drive research questions (DARE UK Consortium, 2021). The combination of computational virtualisation and data visibility provides the building blocks for data *visualisation*. Data visualisation can be understood as "cultural artefacts with distinct semiotic, aesthetic, and social affordances" (Kennedy & Engebretsen, 2020, p. 24).

3.3.4 The research and innovation community layer

Connecting research communities through digital research infrastructure is a prerequisite to stimulating the exchange of ideas, data and results. A range of *different categories of institutions* play a role in supporting research data management, including higher education institutions, research organisations, research funding agencies, science councils, scholarly publishers, third-party service providers, and international organisations (Khair et al., 2020). Communities are drawn from, and constituted by, university departments, research groups and individual researchers; data centres, institutional repositories, national and international data archives; and special interest groups, international communities of interest, standards organisations, and professional societies (Meghini et al., 2017). A key concern of many countries is the development and retention of the requisite *human resource capability and capacity* (Australian Government, 2021; CFI, 2015; DARE UK Consortium, 2021).

Digital research infrastructures, as part of a national system of innovation, serve to increase the value of the economy and responsiveness of society. Virtualisation, visibility and visualisation through digital research infrastructures paves the way for the valorisation of data, for the benefit of the economy and society.

3.3.5 The governance layer

The governance layer integrates the network and connectivity, computational, data and community layers of the digital research infrastructure ecosystem by guiding *decision-making* (Australian Government, 2021). The governance of digital research infrastructures addresses the *adoption of structures, procedures and instruments* necessary for steering relationships and interdependencies between the actors involved (Crompvoets, et al., 2018). Robust governance allows for greater *coordination* and *alignment* among components and actors, facilitating timely access to appropriate services and resources (CFI, 2015) and reconciling collective and individual needs and interests from different stakeholders to achieve common goals (Crompvoets, et al., 2018). Data governance in the context of digital research infrastructures forms an integral part of the overall governance process. It is understood as the management and maintenance of data assets and related aspects, including data access, privacy and security, and incorporates the mechanisms for decision-making and processes of data (Curry, et al., 2022), noting that attention should also be paid to standards for research ethics.

3.4 From digital research infrastructure to the platformisation of science and innovation

The emergence of *digital research infrastructure* as a focus area for improving and evolving established research infrastructures resulted from the increasing reliance on the digital infrastructure (Stührenberg, et al., 2021). It is useful to think of infrastructure not as static but as dynamic and evolving so that the use of the analytical concept 'infrastructuring' may be useful to the extent that it shifts the focus from structure to *process* (Pawlicka-Deger, 2022). From a process perspective, digital research infrastructures operate as an *ecosystem* in a constant process of engagement, adjustment and readjustment as the parts of the system interact, shift and change (Anderson, 2013).

Digital platforms and ecosystems have become a dominant form not only in social and economic organisation in the digital age (Gawer, 2021), but increasingly also visible in the *platformisation of science* (Chiarini & Netto, 2022). Platforms are, by design, a *central agent* of a network (Gawer, 2021). Research platforms organise by bringing bodies and brains into relation and encourage *collaborative practices* with the potential for the invention of new institutional forms (Kanngieser, et al., 2014). Research platforms serve as the *interface* between the data and users (OECD, 2017).

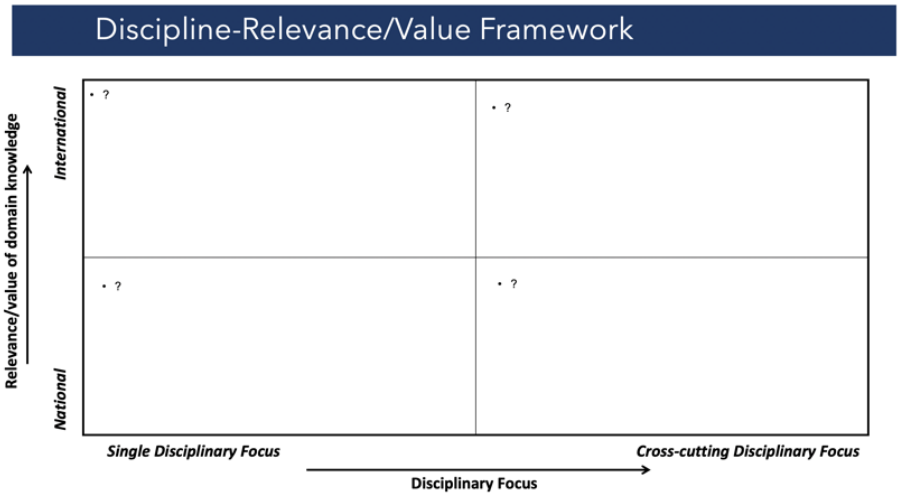
While the public costs are significant, future research platform advancement opens new possibilities of *cross-domain and cross-sector integration* which require new approaches to infrastructure provisioning for collection, storage, distribution, analysis, exchange, preservation and re-use of research data (Otto, et al., 2022). This can enable data journeys through which research data platforms play a significant role in *reshaping institutional, disciplinary and social boundaries* by continuously constructing, destroying and re-making those boundaries (Leonelli, 2020).

4. Mapping of research/innovation concentrations and domain selection and rationale

The foundational mapping exercise included all RI's, all CoEs', a selection of the approximately 200 SARChI Chairs and a few other key institutions such as the broader environmental observation network SAEON. This mapping identified a logical approach to the task of constructing digital research infrastructure, because it showed (i) overlapping fields of research that are relatively powerful but not currently well connected (ii) broad fields of research that are marginalised in the innovation system (iii) areas for R&D and innovation that could be much more actively pursued and (iv) where the "low hanging fruit" for transitioning to digital platforms lies, in the sense that these domains can contribute to and benefit from a data platforms logic. We applied the discipline-relevance/value framework, see Figure 3 below, to more carefully express the logic apparent for the six domains.

Figure 3

Design of the Discipline-Relevance/Value Matrix

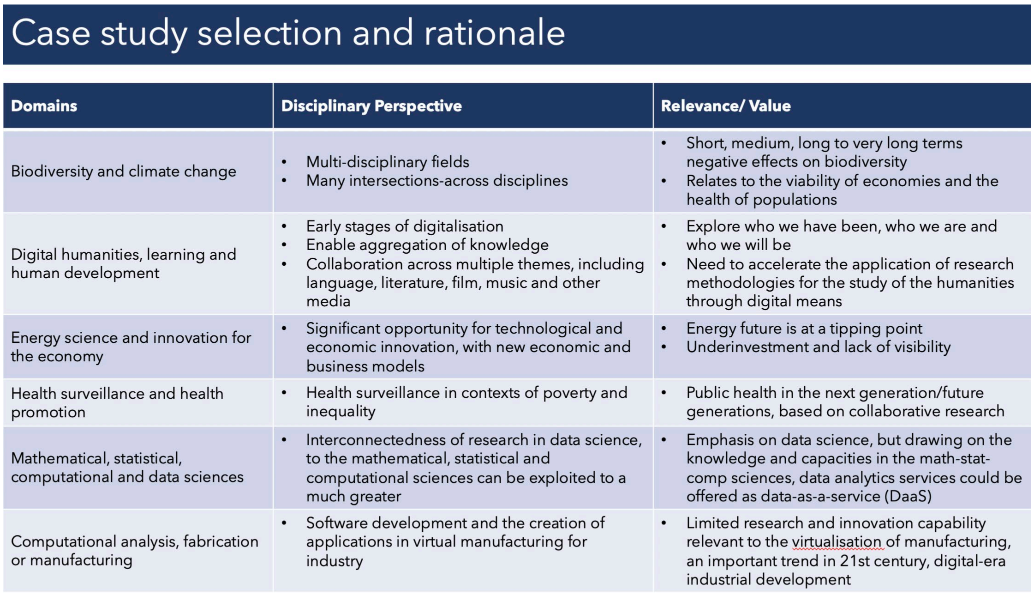

Discipline-Relevance/Value Framework
• ?
• ?
• ?
• ?
Relevance/value of domain knowledge
International
National
Single Disciplinary Focus
Cross-cutting Disciplinary Focus
Disciplinary Focus

Note. Authors.

The six initial interdisciplinary domains are (i) biodiversity and climate change (ii) digital humanities, learning and human development (iii) energy science and innovation for the economy (iv) health surveillance and promotion in contexts of poverty and inequality (v) mathematical, statistical, computational and data science and (vi) virtualisation in computational analysis, fabrication or manufacturing, see the domains, disciplinary perspective, and relevance/value summarised in Figure 4 below.

Figure 4

Case Study Selection: Domains, Disciplinary Perspective, Relevance/Value



Case study selection and rationale

Domains	Disciplinary Perspective	Relevance/ Value
Biodiversity and climate change	• Multi-disciplinary fields • Many intersections-across disciplines	• Short, medium, long to very long terms negative effects on biodiversity • Relates to the viability of economies and the health of populations
Digital humanities, learning and human development	• Early stages of digitalisation • Enable aggregation of knowledge • Collaboration across multiple themes, including language, literature, film, music and other media	• Explore who we have been, who we are and who we will be • Need to accelerate the application of research methodologies for the study of the humanities through digital means
Energy science and innovation for the economy	• Significant opportunity for technological and economic innovation, with new economic and business models	• Energy future is at a tipping point • Underinvestment and lack of visibility
Health surveillance and health promotion	• Health surveillance in contexts of poverty and inequality	• Public health in the next generation/future generations, based on collaborative research
Mathematical, statistical, computational and data sciences	• Interconnectedness of research in data science, to the mathematical, statistical and computational sciences can be exploited to a much greater	• Emphasis on data science, but drawing on the knowledge and capacities in the math-stat-comp sciences, data analytics services could be offered as data-as-a-service (DaaS)
Computational analysis, fabrication or manufacturing	• Software development and the creation of applications in virtual manufacturing for industry	• Limited research and innovation capability relevant to the virtualisation of manufacturing, an important trend in 21st century, digital-era industrial development

Note. Authors.

5. Scoping of six domains for first phase of digital platformisation

It is noted here that the six domains are reasonably well aligned to, though not an exact fit with, the six domains of SARIR, as that would require an even greater undertaking, in terms of scope, than what is proposed here for the first phase of platformisation.

5.1 Domain 1: Biodiversity and climate change

Prospective creators and users would include six DSI/NRF Centres of Excellence or RIs, at least 14 SARChI Chairs and institutions such as SAEON and SAIAB. There are significant datasets and resources that can be shared across a wider research community, rendering greater visibility and greater ease of access (examples Wind Atlas for South Africa Times Series Data; South African Estuaries Information System). The building blocks of digital infrastructure and platforms already exist (examples Climate Information Portal; SAEON's Open Data Platform and Observations Database) but are not accessible on the same interoperable platform. There are important early features of platform governance including guidance, rules and standards setting.

5.2 Domain 2: Digital humanities, learning and human development

Key institutions (prospective creators and users) would be the CoE Human, SADiLaR, a few SARChI Chairs, and CSIR voice computing, but this important domain of socially oriented research and innovation is under resourced and there is very limited sharing of knowledge across the digital humanities, which is (i) strongly emerging as a domain of research across Africa and globally and (ii) is closely connected to additional pathways of learning and human development in the 21st century. This domain has some specific areas of focus with respect to datasets and resources, noting in particular the work in natural language processing and voice computing, which can enhance the capacity for human communication; the capacity for heritage preservation including the national archives, national sound film and video archives and the many scattered heritage collections; and opportunities in the visual arts. In this domain, content is overwhelmingly analogue and therefore largely inaccessible to scholars and to the society, which presents a high risk for advancing the field.

5.3 Domain 3: Health surveillance and health promotion in contexts of poverty and inequality

Key institutions (prospective creators and users) would include the South African Population Research Infrastructure Network (SAPRIN), the Medical Research Council (MRC), the African Centre of Excellence for Inequality Research and the HSRC, engaged in various forms of demographic surveillance in resource-constrained environments, yet where there is very limited interconnection across these fields of study, thus limiting the value of the research to policymakers and to citizens. There are already extensive datasets and resources that can be made visible for much more extensive analysis, some of which are already open access. This includes the SAPRIN Individual Surveillance Episodes Dataset and the HSRC National HIV Prevalence, Incidence, Behaviour and Communication Survey. Early-stage digital infrastructure includes the work of DataFirst through its open data infrastructure. There are opportunities for advanced research through computational modelling and data visualisation, which can be provided as a service by the data science community.

5.4 Domain 4: Energy science and innovation for the economy

Prospective research and innovation creators and users would include CIMERA, CRSES, the Energy Research Centre and the Photovoltaics Research Group at CSIR, the Energy Research Center UCT, SANEDI, SARETEC and the Solar Thermal Energy Research Group. According to available information, there is as yet only limited or no formal or semi-formal cross-institutional research collaboration, and no extensive knowledge sharing. The publicly available data is mainly informational and mostly accessible through websites, with no visible research and innovation platform infrastructure and services. Platform design and governance could be structured to promote energy virtual laboratories and research environments (EVLREs), while the platform shared services (see high level platform visualisation below) can offer data-analytics-as-a-service (DAaaS).

5.5 Domain 5: Mathematical, statistical, computational and data sciences

Prospective research and innovation creators and users in this domain would include the School for Data Science and Computational Thinking (SU), the Unit for Data Science and Computing (NWU), the Centre for Applied Data Science (SPU), the Centre for Applied Data Science (UJ), the Data Science Centre for Business (UKZN), the Wits Institute of Data Science (WIDS), CODATA, the Statistics South Africa open data portal and interactive data site, the work of CREST in bibliometrics and scientometrics, the work of CeSTII in STI indicators and innovation surveys, the DataFirst Open Data Portal and civic technology and open data specialists such as OpenUp. Currently available resources and datasets include the SAKnowledgebase, CeSTII's innovation survey datasets, the NACI STI indicators portal, as well as datasets and repositories held in each of the six domains. Since they conduct data science, this group would be the ideal lead consortium for the creation of SA-DRIP.

The institutions that provide the computational environment for research (in general) and for data science (in particular) include DIRISA, the NICIS SANReN, the NICIS CHPC, the NITheCS and TENET. Each of these institutions can make a significant contribution to advancing the computational power, data storage and data management environment for platform design. The proposition made here would need to be more carefully assessed in discussion with these institutions.

5.6 Domain 6: Virtualisation in hybrid computational analysis, fabrication or manufacturing (virtual manufacturing)

Prospective research and innovation creators and users would include iThemba LABS, the NCNSM materials characterisation facility (RI), and the CoE in Strong Materials. This is an opportunity for significant engagement with industry to actively foster industrial innovation through automated fabrication and virtual manufacturing, to engender more effective competition with South Africa's main industrial competitors. Datasets and resources would include future data libraries, modelling libraries, software libraries and design libraries, work on artificial intelligence (AI) applications and collaborative robotics (co-botics, where humans and machines collaborate), and availability of an accessible online API (application processing interface) laboratory.

6. Cross-cutting themes and preliminary insights

A few brief statements on cross-cutting themes and preliminary insights, more extensively addressed in the full report:

6.1 Cross-cutting theme 1: Women in Science

Platform design must consciously and explicitly promote participation and visibility of women in science, as researchers and innovators. Furthermore, it must promote the practice of women in science participating as colleagues in shaping platform and applications design and platform governance, advancing gender inclusiveness in practice. Herein lies significant opportunity for women's participation in software and hardware development.

6.2 Cross-cutting theme 2: Science practice and inclusion

The nature and characteristics of platform design, while using the prominent work in domains as the foundation, must promote science inclusiveness through open science and other means to promote participation from researchers/innovators beyond those entities and institutions where activities are currently concentrated. This is really important because the design of digital infrastructure could either entrench existing exclusion or foster greater inclusiveness. Amongst the means to promote science inclusiveness including women in science and rural-facing science, while advancing all forms of demographic and geographic equity, must be to apply the principles of science inclusiveness and value realisation, by creating a value realisation matrix at the outset of the project design, and a value realisation register that enables the NRF/science community to monitor and measure value and inclusiveness.

6.3 Preliminary insight 1: Platformisation (digital)

The NRF can create the conditions for conducting "science at scale" or scaling up collaborative science and innovation production or alternatively knowledge sharing, by front-loading the required financial investment in digital research infrastructure. A major benefit would be reducing the marginal cost of digital research infrastructure for the NSI, over time. Initial steps in digital platform design must include a strong focus on open data and open access publishing as a means to populate content, while simultaneously encouraging usage for open science from inception. A key challenge will be to address the institutional barriers that may exist or may arise. The initial steps must also create virtual closed space for innovation with commercialisable potential, for example in the field of virtual or hybrid manufacturing.

6.4 Preliminary insight 2: Human resource capacity for DRI and data platforms for all domains

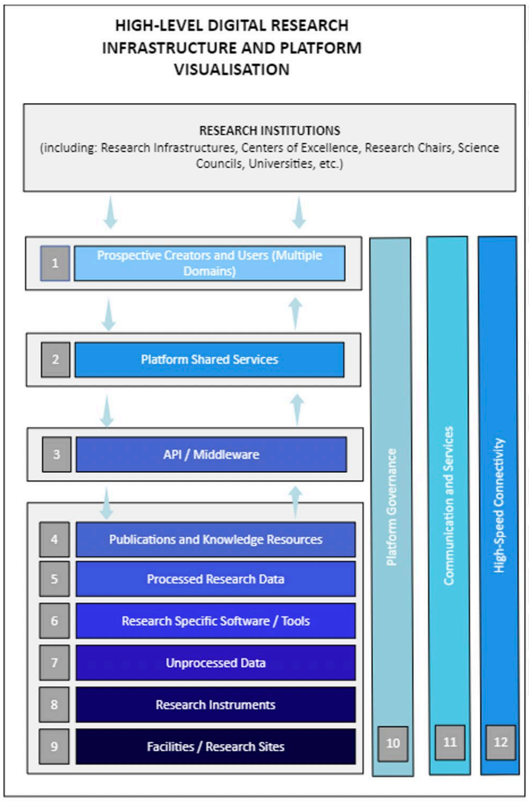
Human resource development in the broad field of the mathematical, statistical, computational and data sciences is a key feature of this particular initiative. Investments would be required for human resource capacity at the platform level, specifically Domain 5, but also in each of the other domains. This is because data science is critical to the long-term sustainability of a research data platform and the mathematical, statistical and computational sciences are key skills integrated with data science practice and future innovation; and because a skills balance is required for the DRI and for its initial six user domains (noting that data science would be engaged in platform governance/execution as well as being a user domain) and for any future domains to be added. While greater capacity would be available at the DRI through inter alia its shared services, more capacity will also increasingly be required in the major user domains.

7. High-level research platform visualisation

The diagrams below highlight the main platform features in the form of a 12-layered high-level design. In Figure 5, we see the 12 layers including the infrastructure and services that enable virtualisation, visibility, visualisation and valorisation. This is the structure of the proposed SA-DRIP, which will have its formal platform governance arrangements (layer 10).

Figure 5

Functional Scoping and Technical Explanation: South Africa's Digital Research Infrastructure Platform (SA-DRIP) for the NSI


HIGH-LEVEL DIGITAL RESEARCH INFRASTRUCTURE AND PLATFORM VISUALISATION
RESEARCH INSTITUTIONS
(including: Research Infrastructures, Centers of Excellence, Research Chairs, Science Councils, Universities, etc.)
1 Prospective Creators and Users (Multiple Domains)
2 Platform Shared Services
3 API / Middleware
4 Publications and Knowledge Resources
5 Processed Research Data
6 Research Specific Software / Tools
7 Unprocessed Data
8 Research Instruments
9 Facilities / Research Sites
Platform Governance
Communication and Services
High-Speed Connectivity
10
11
12

1 **Users** across multiple research domains are able to access research data that have been made available by various researchers across South Africa. This includes both "processed" and "unprocessed" data. Access to the information is facilitated via the Platform Shared Services and enabled by Application Programming Interfaces (APIs) and/or Middleware (software linking two or more systems together).

2 The **Platform Shared Services** serves as "hub" or "nexus" between Research Institutions and Research Projects in South Africa. The Platform Shared Services offers a variety of research coordinating services, detailed in the Platform Shared Services schematic.

3 **API's or Middleware** provides a mechanism to access data across multiple standalone research systems, in order to surface the same via the Platform Shared Services to Research Institutions.

4 **Publications and knowledge resources** refer to published work, based on "processed" research data. This information, alongside "processed" and "unprocessed data" are surfaced (with permission) via the Platform Shared Services to Research Institutions.

5 **"Processed" research data** refers to data that has been analysed in some way. The argument is that research opportunities exist by surfacing this data via the API/Middleware layer to the Platform Shared Services, and ultimately to the Research Institutions as a whole.

6 **Research specific software** refers to: a) software used to interface research specific data and instruments, b) software that is used to generate "processed" research data.

7 **"Unprocessed" research data** refers to data generated by research instruments which has not yet been analysed or "processed". The argument is that this type of data could benefit concurrent or future research projects across multiple domains. Access to this type of data is enabled via the API / Middleware layer and facilitated via the Platform Shared Services.

8 **Research instruments** refer to a) research equipment (electron microscopes, radio telescopes, etc.) and the associated firmware (embedded software used to configure or tune the instruments) as well as b) intangible research instruments including questionnaires, surveys, interviews, etc. which generates research data.

9 **Facilities and research sites** include buildings and specific locations where research activities are conducted.

10 **Platform governance** refers to to decision-making, steering and oversight of platform development and expansion.

11 **Communication and services** are crucial across the platform. This function is faciliated by means of the Platform Shared Services.

12 **High-speed connectivity** remain a crucial catalyst for the flow of data across the research platform.

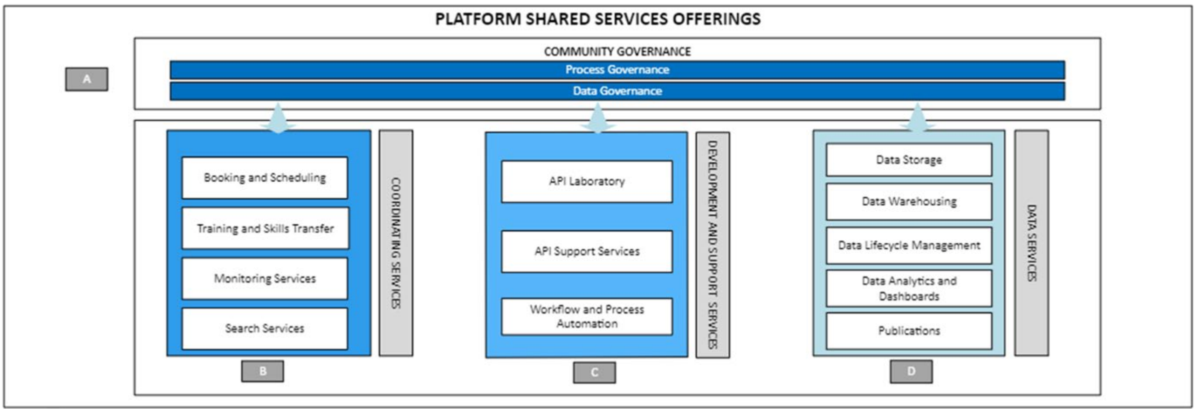
Notes
The development of the API/Middleware tier, requires a comprehensive assessment of existing research data repositories and research specific software and tools. It is anticipated that the Research Platform would be realised over a 5-7 year period.
A phased approach is recommended to realise the full potential of the platform which includes governance elements, the establishment of a shared services capability (office, etc), assessment of existing research data and systems, the provisioning of storage services, and the development of the API or middleware.

Note. Abrahams, Burke and Du Preez, 2023.

In Figure 6, we see the platform shared services offerings, including the co-ordinating services (column B), the development and support services including the API laboratory (column C), and the data services (column D), as well as research community governance specifically for the platform shared services, as one of the forms of DRI governance.

Figure 6

Functional Scoping and Technical Explanation for the Platform Shared Services



A

The platform follows a community governance approach which is typically characterized by an emphasis on participation, inclusivity, and transparency, with decisions being made through a democratic process that involves input and feedback from all members of the research community. Specific emphasis will be given to the following:

- **Process Governance –** The platform aims to guide the defining of standards and rules for how platform processes are created, documented, reviewed, and maintained, with the aim of ensuring consistency, compliance, and quality.
- **Data Governance –** The platform aims to ensure that data is managed as a strategic research asset, with proper oversight and accountability to ensure that data is trusted and reliable, and that it is used effectively to support platform operations and decision-making.

B

- **Booking and Scheduling –** Booking and scheduling of research instruments and access to research sites is provided via the Shared Services offices.
- **Training and Skills Transfer –** The platform shared services provides the necessary skills transfer capabilities required to access, configure and operate specific research instruments.
- **Monitoring Services –** Monitoring services are provided to demonstrate the efficacy of the platform as well as the availability of data and research instruments.
- **Search Services –** Search services across the platform is provided via a centralised portal.

C

- **Application Programming Interface (API) Laboratory –** The API's or middleware provide a mechanism to access data across multiple standalone research systems. These API's are developed in the API laboratory following assessments of the distributed research systems and the development of integration standards.
- **API Support –** Existing API's are serviced by development and support services provided as part of the shared services offerings.
- **Workflow and Process Automation –** Platform processes are continuously reviewed for efficiency and those requiring attention are re-engineered and automated. This forms part of a continuous process optimisation cycle.

D

- **Data Storage -** The platform provides data storage services to house data made available by research institutions across South Africa. Access to "processed" and "unprocessed" data is enabled via the API/Middleware layer.
- **Data Warehousing –** Data warehousing capabilities allows for consolidated views of data across research domains.
- **Data Lifecycle Management –** Thesaurus and metadata services are provided as part of data lifecycle management services.
- **Data Analytics and Dashboards –** The data warehousing capabilities of the platform provides a vehicle for intelligence reports and dashboards pertaining to research activities across the platform as well as the platform itself.
- **Publications -** Publications and resources, generated by the research platform, are stored as part of the data services offerings provided by the platform.

Note. Abrahams, Burke & Du Preez, 2023.

The institutions in the data science domain (Domain 5) would be at the core of building the South African digital research infrastructure platform (SA-DRIP). Furthermore, three working groups could be established to foster collaborative creation, including (i) Working Group on Science Observation and Intelligence (ii) Working Group on Research Data Management Standards and (iii) Working Group on Cloud Computing, Data Storage and Processing, Networking and Connectivity. Each working group would contribute their knowledge to advancing the establishment and work of the SA-DRIP.

8. Concluding remarks

As expressed in the introduction, digital research infrastructures can, where designed for effective use and continuous evolution, provide the basis for transforming scientific knowledge production capabilities and practice in the health sciences, in the natural sciences and in the social sciences and humanities. It can promote science inclusiveness of many kinds, including a much stronger push for women in science inclusiveness. Digital research infrastructures can empower the scientific community by enabling the application of new methods and techniques of scientific discovery, by enabling greater aggregation of knowledge and by enabling greater scientific collaboration. But it can only do so through the active engagement of scientists through their science.

References

Abrahams, L. & Burke, M. (2023, March). *Transforming scientific knowledge production capabilities and practices by digitally enabling research communities and communication: Virtualisation, visibility, visualisation and valorisation.* Guiding report to the National Research Foundation and Appendix A. LINK Centre, University of the Witwatersrand, Johannesburg.

Abramov, A. (2021). Service portfolios of leading national research and education networks and implementation on the basis of the National Research Computer Network of Russia. *Lobachevskii Journal of Mathematics*, *42*(11), 2481–2492.

African Open Science Platform (AOSP). The African Open Science Platform. https://aosp.org.za/

Anderson, S. (2013). What are research infrastructures? *International Journal of Humanities and Arts Computing, 7*, 4–23. https://doi.org/10.3366/ijhac.2013.0078

Australian Government. (2021). *National research infrastructure roadmap 2021.* https://www.education.gov.au/national-research-infrastructure/resources/2021-national-research-infrastructure-roadmap

Bailo, D., Ulbricht, D., Nayembil, M., Trani, L., Spinuso, A., & Jeffrey, D. (2017). Mapping solid earth data and research infrastructures to CERIF. *Procedia Computer Science*, *106*, 112–121. https://doi.org/10.1016/j.procs.2017.03.043

Borgerud, C., & Borglund, E. (2020). Open research data, an archival challenge? *Archival Science 20*, 279–302. https://doi.org/10.1007/s10502-020-09330-3

Boumans, M., Leonelli, S. (2020). From dirty data to tidy facts: Clustering practices in plant phenomics and business cycle analysis. In S. Leonelli, & N. Tempini (Eds.), *Data journeys in the sciences* (pp. 79–102). SpringerOpen. https://doi.org/10.1007/978-3-030-37177-7

Bratt, S. (2022). *Research data management practices and impacts on long-term sustainability: An institutional exploration*. [Doctoral dissertation, Syracuse University]. Surface. https://surface.syr.edu/cgi/viewcontent.cgi?article=2544&context=etd

Broeder, D., Du Toit, L., Meyer, J., & Bot, J. (2017). Data and computing landscape characteristics and scientific communities' environment & requirements. EUDAT. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5b77b4095&appId=PPGMS

Burgelman, J., Pascu, C., Szkuta, K., Von Schomberg, R., Karalopoulos, A., Repanas, K., & Schouppe, M. (2019). Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. *Frontiers in Big Data, 2*, 43. http://doi.org/10.3389/fdata.2019.00043

Canada Foundation for Innovation (CFI). (2015). *Developing a digital research infrastructure strategy for Canada: The CFI perspective.* https://www.innovation.ca/sites/default/files/Funds/cyber/developing-dri-strategy-canada-en.pdf

Carver, J. C., Weber, N., Ram, K., Gesing, S., & Katz, D. S. (2022). A survey of the state of the practice for research software in the United States. *PeerJ. Computer Science*, *8*, e963. https://doi.org/10.7717/peerj-cs.963

Chiarini, T., & Da Silve Neto, V. (2022). The platformisation of science: Towards a scientific digital platform taxonomy. *Minerva.* https://doi.org/10.1007/s11024-022-09477-6

Connaway, L., & Dickey, T. (2010). *Towards a profile of the researcher of today: Common themes identified in an analysis of JISC virtual research environment and digital repository projects*. https://www.webarchive.org.uk/wayback/archive/20140613220103/http://www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf

Critchlow, T., & van Dam K. (2013). What is data-intensive science? In T. Critchlow & K. van Dam (Eds.), *Data-intensive science* (pp. 1–14). CRC Press.

Crompvoets, J., Vancauwenberghe, G., Ho, S., Masser, I., & de Vries, W. T. (2018). Governance of national spatial data infrastructures in Europe. *International Journal of Spatial Data Infrastructures Research*, *13*, 253–285. https://doi.org/10.2902/1725-0463.2018.13.art16

Cudennec, C., Lins, H., Uhlenbrook, S., Amani, A., & Arheimer, B. (2022). Operational, epistemic and ethical value chaining of hydrological data to knowledge and services: A watershed moment. *Hydrological Sciences Journal*, *67*(6), 2363–2368. https://doi.org/10.1080/02626667.2022.2086462

Curry, E., Scerri, S., & Tuikka, T. (2022). Data spaces: Design, deployment, and future directions. In E. Curry, S. Scerri, & T. Tuikka (Eds.), *Data spaces* (pp. 1–17). Springer. https://doi.org/10.1007/978-3-030-98636-0_1

DARE UK Consortium. (2021). *UK data research infrastructure landscape*. https://doi.org/10.5281/zenodo.5584696

Department of Science and Innovation (DSI). (2019, March). White paper on science, technology and innovation. https://www.dst.gov.za/images/2019/White_paper_web_copyv1.pdf

Department of Science and Technology (DST). (2016, October). *South African research infrastructure roadmap*. First edition. https://www.dst.gov.za/images/Attachments/Department_of_Science_and_Technology_SARIR_2016.pdf

Devriendt, T, Shabani, M., & Borry, P. (2022). Policies to regulate data sharing of cohorts via data infrastructures: An interview study with funding agencies. *International Journal of Medical Informatics*, *168*, 104900, https://doi.org/10.1016/j.ijmedinf.2022.104900

European Strategy Forum on Research Infrastructure (ESFRI). (2018). *Strategy report on Research Infrastructures roadmap 2018*. https://research-and-innovation.ec.europa.eu/system/files/2018-10/esfri-roadmap-2018.pdf

European Strategy Forum on Research Infrastructure (ESFRI). (2020). *Roadmap 2021: Strategy report on research infrastructures*. https://roadmap2021.esfri.eu/media/1295/esfri-roadmap-2021.pdf

European Commission (EU). (2020). *Landscape of EOSC-related infrastructures and initiatives: Report from the EOSC Executive Board Working Group (WG) landscape: Version 2*, Directorate-General for Research and Innovation. https://data.europa.eu/doi/10.2777/132181

Foley, M. (2016). *The role and status of national research and education networks (NRENs) in Africa*. https://documents1.worldbank.org/curated/en/233231488314835003/pdf/113114-NRENSinAfrica-SABER-ICTno05.pdf

Gabrielsen, A. (2020). Openness and trust in data-intensive science: The case of biocuration. *Medicine, Health Care and Philosophy*, *23*, 497–504. https://doi.org/10.1007/s11019-020-09960-5

Gawer, A. (2022). Digital platforms and ecosystems: Remarks on the dominant organizational forms of the digital age. *Innovation*, *24*(1), 110–124. https://doi.org/10.1080/14479338.2021.1965888

GÉANT. (2022). *Compendium report 2021 of national research and education networks in Europe*. https://resources.geant.org/wp-content/uploads/2022/07/Compendium-2021-web.pdf

Jiang, Y., Li, X., & An, B. (2020). Function virtualisation in high performance computing: Opportunities and challenges. *Procedia Computer Science*, *174*, 210–215. https://doi.org/10.1016/j.procs.2020.06.076

Kanngieser, A., Neilson, B., & Rossiter, N. (2014). What is a research platform? Mapping methods, mobilities and subjectivities. *Media, Culture & Society*, *36*(3), 302–318. https://0-doi-org.innopac.wits.ac.za/10.1177/0163443714521089

Khair, S., Dara, R., Haigh, S., Leggott, M., Milligan, I., Moon, J., Payne, K., Portales-Casamar, E., Roquet, G., & Wilson, L. (2020). *The current state of research data management in Canada*. Research Data Alliance of Canada. https://alliancecan.ca/sites/default/files/2022-03/rdm_current_state_report-1_1.pdf

Kennedy, M., & Engebretsen, M. (2020). Introduction: The relationships between graphs, charts, maps and meanings, feelings, engagements. In M. Kennedy & M. Engebretsen (Eds.), *Data visualisation in society* (pp. 19–34). Amsterdam University Press. https://library.oapen.org/bitstream/id/da811a79-99fd-4ad7-b52e-45076a32beb4/9789048543137.pdf

Kinkade, D., & Sheperd, A. (2022). Geoscience data publication: Practices and perspectives on enabling the FAIR guiding principles. *Geoscience Data Journal*, *9*, 177–186. https://doi.org/10.1002/gdj3.120

Lee, D. J., & Stvilia, B. (2017). Practices of research data curation in institutional repositories: A qualitative view from repository staff. *PloS One*, *12*(3), e0173987. https://doi.org/10.1371/journal.pone.0173987

Leonelli, S. (2022). Learning from data journeys. In S. Leonelli, & N. Tempini (Eds.), *Data journeys in the sciences*. SpringerOpen. https://doi.org/10.1007/978-3-030-37177-7

Lipton, V. (2020). *Open scientific data: Why choosing and reusing the right data matters*. Intechopen. https://doi.org/10.5772/intechopen.87201

Meghini, C., Scopigno, R., Richards, J., Wright, H., Geser, G., Cuy, S., Fihn, J., Fanini, B., Hollander, H., Niccolucci, F., Felicetti, A., Ronzino, P., Nurra, F., Papatheodorou, C., Gavrilis, D., Theodoridou, M., Doerr, M., Tudhope, D., Binding, C., & Vlachidis, A. (2017). ARIADNE: A research infrastructure for archaeology. *Journal on Computing and Cultural Heritage*, *10*(3), 1–27. https://doi.org/10.1145/3064527

Organisation for Economic and Cooperation and Development (OECD). (2017). *Business models for sustainable research data repositories* (OECD science, technology and innovation policy papers, No. 47). https://doi.org/10.1787/302b12bb-en

Organisation for Economic Cooperation and Development (OECD) (2017a). *Digital platforms for facilitating access to research infrastructure* (OECD science, technology and innovation policy papers, No. 49). https://www.oecd-ilibrary.org/docserver/8288d208-en.pdf?expires=1674912494&id=id&accname=guest&checksum=121B4AC7043A0784E878CD8C9783F8A3

Otto, B., ten Hompel, M., & Wrobel, S. (Eds.) (2022). *Designing data spaces: The ecosystem approach to competitive advantage*. Springer. https://doi.org/10.1007/978-3-030-93975-5

Paic, A. (2021). *Open science – Enabling discovery in the digital age* (OECD Going Digital Toolkit notes, No. 13). https://doi.org/10.1787/81a9dcf0-en.

Partnership for Advanced Computing in Europe (PRACE). (2018). *PRACE in the EuroHPC era*. [Position Paper]. https://prace-ri.eu/about/position-papers/#PositionPaper

Pawlicka-Deger, U. (2022). Infrastructuring digital humanities: On relational infrastructure and global reconfiguration of the field, *Digital Scholarship in the Humanities*, *37*(2), 534–550, https://doi.org/10.1093/llc/fqab086

Radchenko, G., Alaasam, A., & Tchernykh, A. (2019). Comparative analysis of virtualisation methods in big data processing. *Supercomputing Frontiers and Innovations*, *6*(1), 48–79. https://doi.org/10.14529/jsfi190107

RISCAPE. (2019). *International research infrastructure landscape report 2019*. https://zenodo.org/record/3539254/files/RISCAPE_consolidated_301219.pdf?download=1

Stührenberg, M., Schonefeld, O., & Witt, A. (2021). Digital research infrastructure. In C. Koschtial, T. Köhler, & C. Felden (Eds.). *e-Science: Open, social and virtual technologies for research collaboration* (pp. 67–76). Springer. https://doi.org/10.1007/978-3-030-66262-2_5

Trumpy, E., Coro, G., Manzella, A., Pagano, P., Castelli, D., Calcagno, P., Nador, A., Bragasson, T., Grellet, S., & Sidiqi, G. (2015). Building a European geothermal information network using a distributed e-infrastructure. *International Journal of Digital Earth*, *9*(5), 499–519. https://doi.org/10.1080/17538947.2015.1073378

United Kingdom Research and Innovation (UKRI). (2020). *The UK's research and innovation infrastructure: Opportunities to grow our capability*. https://www.ukri.org/wp-content/uploads/2020/10/UKRI-201020-UKinfrastructure-opportunities-to-grow-our-capacity-FINAL.pdf

United Nations Educational, Scientific and Cultural Organization (UNESCO). (2021, November). *UNESCO recommendation on open science*. https://en.unesco.org/science-sustainable-future/open-science/recommendation

Wilkinson, M., Dumontier, M., Aalbersberg, I., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., da Silva Santos, L., Bourne, P., Bouwman, J., Brookes, A., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C., Finkers, R., …Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), https://doi.org/10.1038/sdata.2016.18

Yang et al., (2014). Cloud computing in e-science: Research challenges and opportunities. *Journal of Supercomputing*, *70*, 408–464. https://doi.org/10.1007/s11227-014-1251-5

Zozus, M. (2017). *The data book: Collection and management of research data*. CRC Press.