UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Responding to the national cyberinfrastructure challenge

Prof Mamokgethi Phakeng
Deputy Vice-Chancellor for Research & Internationalisation

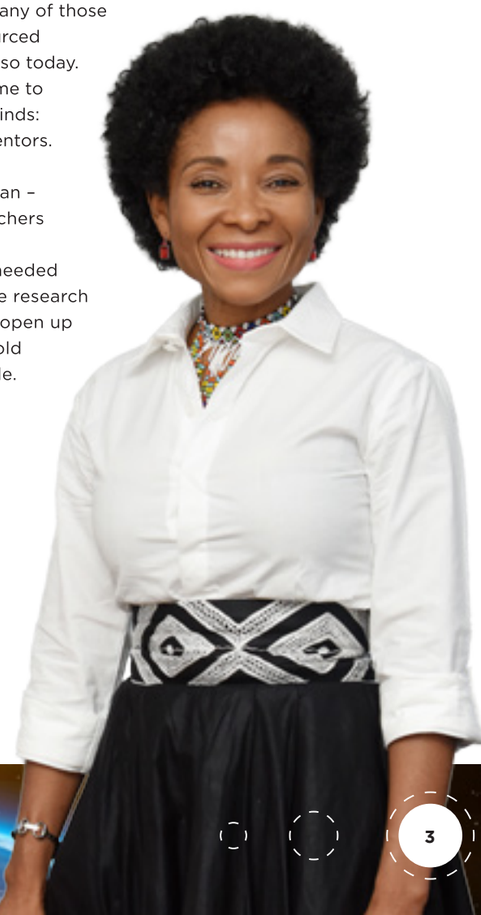
In order for UCT to stay globally competitive as a research institute in a world of big data, we need to ensure our researchers have access to cutting-edge facilities and infrastructure. However, this is a regional, national and continental challenge; UCT needs to work with its neighbours to advance the African research agenda.

The projects outlined in this publication signal a new way of thinking about collaborative systems and services.

The new mindset sets us on a path that will give more access to research infrastructure, software and data to all South African researchers – including those from our under-resourced communities, where access to this kind of infrastructure is presently unimaginable.

South Africa is home to 26 tertiary education institutions; but many of those were chronically under-resourced under apartheid, and remain so today. Yet those institutions are home to some of our nation's great minds: our problem-solvers and inventors.

Ensuring that all South African – and in time, African – researchers have access to the kinds of facilities and infrastructure needed to address the data-intensive research challenges we now face will open up opportunities and bring untold benefits to society as a whole.



Driving regional and national collaboration

Dr Dale Peters
Director UCT eResearch, Ilifu Project Co-Lead

Virtually every field of science, including the natural and human sciences, has been changed by the convergence of exponential increases in computing, storage and bandwidth, enabling global collaboration using distributed information technologies.

Thus the challenges faced by eResearch have not only been in provisioning local infrastructure, but also in negotiating strategic developments in the national integrated cyberinfrastructure, to find sustainable solutions beyond UCT that serve the needs of the country. As a result, we have seen a steady progression towards regional and national partnerships to deliver shared infrastructure and services. UCT eResearch is pioneering the way – amid growing interest from national institutions in what eResearch is and what it can bring to enhance and inform the ways in which scientists carry out their work, the tools and workflows they use, the types of problems they address in response to societal challenges, and the communications resulting from their research.

Research infrastructure needs to continue to be a major driver of collaboration, in a cumulative aggregation across projects and between institutions. The pages of this publication bear testament to both the successes of such collaborations and their future potential.

The launch this year of the Iziko Digital Planetarium is one example of such a success. Researchers from a regional consortium of universities – comprising UCT, the Cape Peninsula University of Technology, and the University of the Western Cape – now have access to a world-class facility, in an ingenious partnership with a public museum. The rapid deployment of the Western Cape Data-Intensive Research Facility, known as Ilifu, is another such success.

While we are certainly proud of the achievements of UCT eResearch in infrastructure and service development since its inception in 2014, none of them would have been possible without these collaborations, and the collective response to the challenge of exponential growth witnessed in data-intensive science.

Paving the way for SKA data

Prof Russ Taylor
Joint SKA Research Chair: University of Cape Town and University of the Western Cape,
Director, Inter-University Institute for Data-Intensive Astronomy,
Ilifu Project Lead

The first decades of this century have seen a tremendous advance in information and digital technologies impacting scientific inquiry. Data created by megaprojects in science and engineering, by the ubiquitous sensors tracking the state of the planet, by the connected internet of things, and aggregated in vast and complex collections of data that trace the patterns and trends in human behaviour, are beginning to be creatively mined in ways that fundamentally change our perception of the world and empower global change.

The Square Kilometre Array (SKA) drives one of the most significant big-data challenges of the coming decade. Since South Africa won the SKA bid to co-host this project with Australia, its research-intensive universities have been presented with an opportunity that must be grasped. Our organisations must solve the data challenges of the SKA to lead the SKA key science investigations.

The SKA data onslaught begins with the completion of the South African MeerKAT – the first stage of the SKA – in mid-2018. South Africa must work quickly to rise to the data challenge. Using cloud-computing technologies, Ilifu will build capacity for research on the big data created by MeerKAT. The project will also bring together leading South African researchers, in collaboration with software and system developers and eResearch teams at partner organisations, to co-develop computing software systems and algorithms that will allow South Africa to lead the world in SKA science.

Visualising a universe of data

Cape Town's 30-year-old planetarium has been revamped by Iziko Museums to create a facility that brings data to life through immersive visualisations. The new planetarium – the result of a partnership that includes the University of Cape Town (UCT) – could be a model for others around the world.

By harnessing advanced technologies to create immersive visualisations, the Iziko Planetarium and Digital Dome is in a position not only to help researchers rapidly advance our understanding of the world, but also to make that same information available to the public in an easily accessible visual form. This is particularly true for big data.

Big data refers to the large, complex data sets created and collected through technology. They can range from the data generated by social media or projects such as the Square Kilometre Array (SKA) to the big data produced by genome sequencing.

"In the world of huge data sets, some data can only be understood if you can see it," says UCT Emeritus Professor Danie Visser, patron of the Iziko Planetarium and Alexander von Humboldt Fellow at the Max Planck Institute for Comparative and International Private Law in Hamburg. "This is a powerful tool across all disciplines."

The SKA project is an obvious candidate for the digital dome. Massive data sets are already being created as MeerKAT, the precursor to the SKA telescope, comes online.

This has positive implications for the public, too. It means that, in time, we won't only be reading about the SKA discoveries in headlines; we will also be able to view the secrets of the universe at the planetarium. And, as more researchers from other fields use the facility for data analysis, more groundbreaking science will be visualised and made available to the public.

Using the dome to study the structure of the universe

Professor Tom Jarrett, UCT's Department of Science and Technology/National Research Foundation South African Research Chair in Astrophysics and Space Science, is studying the structure of the universe, trying to better understand the life cycle of galaxies.

Galaxies are not scattered randomly around the universe; they are clustered together in the groups that make up the cosmic web. To study these structures, Jarrett works with data sets that can easily include a million or more galaxies.

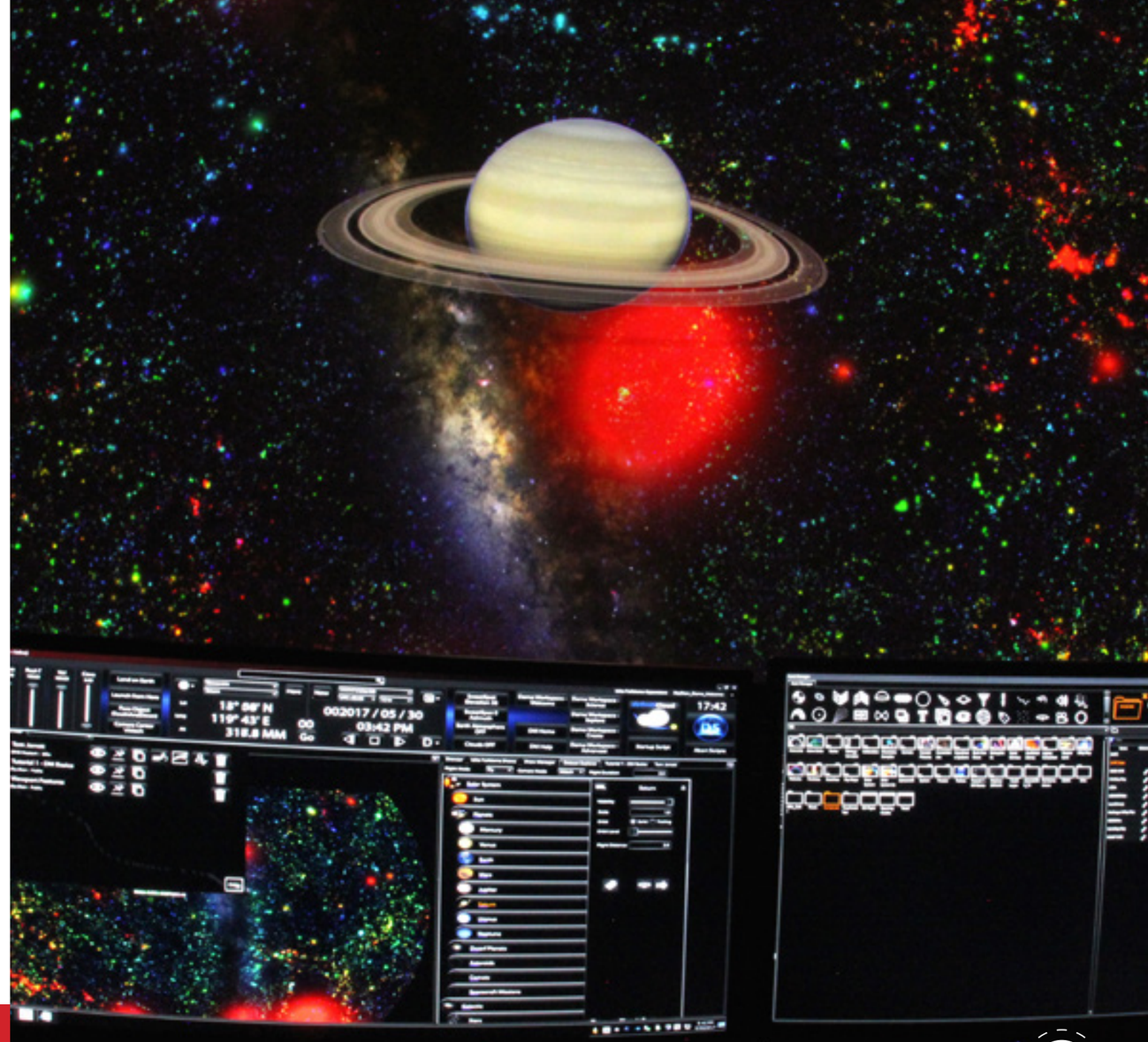
"These cosmic structures are really big; it's difficult to study them on a computer screen," he says. Each data set is loaded into a 3D catalogue in which each galaxy is mapped according to its coordinates in space. Because Jarrett has the coordinates, he can project them onto the dome and fly around the galaxy structures.

"Within these galaxy catalogues, I am trying to find new structures of galaxies," he explains. "With the immersive dome, I can actually fly into the data set. I can spy and isolate particular structures and accurately measure the gaps and filaments between the galaxy clusters."

So far, Jarrett is the only researcher to use the digital dome for research. Part of his role is to ensure the computing support and necessary software for this facility is available for researchers to use. With his colleague Professor Michelle Cluver, Associate Director of the Inter-University Institute for Data-Intensive Astronomy, he is also running a series of workshops and presentations for researchers in the Western Cape, to showcase how the facility can be used for research purposes.

"To learn how galaxies are born, evolve and grow, we need to study their environment – where they draw fuel from to grow, and gravity to shape their development. This is the cosmic web that we consider the context for galaxy evolution," says Jarrett.

"Because of the immersive nature of the dome, this is an excellent facility to help researchers really get deep into their data. It offers both the breadth and the 3D capacity to allow you to study your data from all angles, which are so limited on a flat screen."



View from Iziko Digital Planetarium and Digital Dome control booth. Control panel display screens visible with Saturn projected on the dome's foreground and galaxies in the background.

Meeting our research infrastructure needs

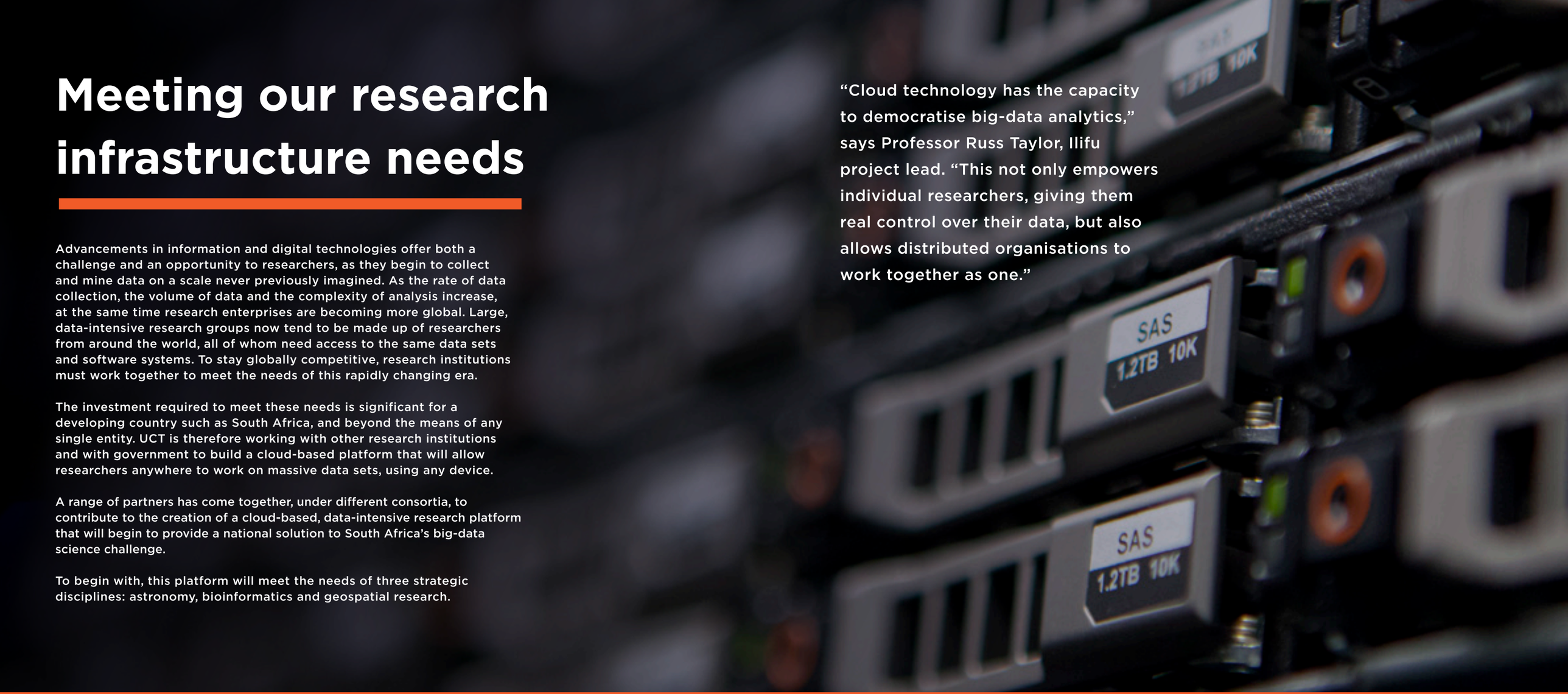
Advancements in information and digital technologies offer both a challenge and an opportunity to researchers, as they begin to collect and mine data on a scale never previously imagined. As the rate of data collection, the volume of data and the complexity of analysis increase, at the same time research enterprises are becoming more global. Large, data-intensive research groups now tend to be made up of researchers from around the world, all of whom need access to the same data sets and software systems. To stay globally competitive, research institutions must work together to meet the needs of this rapidly changing era.

The investment required to meet these needs is significant for a developing country such as South Africa, and beyond the means of any single entity. UCT is therefore working with other research institutions and with government to build a cloud-based platform that will allow researchers anywhere to work on massive data sets, using any device.

A range of partners has come together, under different consortia, to contribute to the creation of a cloud-based, data-intensive research platform that will begin to provide a national solution to South Africa's big-data science challenge.

To begin with, this platform will meet the needs of three strategic disciplines: astronomy, bioinformatics and geospatial research.

"Cloud technology has the capacity to democratise big-data analytics," says Professor Russ Taylor, Ilifu project lead. "This not only empowers individual researchers, giving them real control over their data, but also allows distributed organisations to work together as one."



AFRICAN RESEARCH CLOUD (ARC)

The ARC, a collaboration between UCT, the Inter-University Institute for Data-Intensive Astronomy (IDIA) and North West University, is the prototype for a cloud-based service to researchers working in data-intensive disciplines. Established in 2016, the ARC is testing different models of data management, storage and transfer through radio astronomy and genomics projects.

“The initiative is a first for Africa, and will be a real benefit to researchers on the continent,” says Sakkie Janse van Rensburg, UCT’s executive director of Information Communication Technology Services (ICTS).

SOUTH AFRICAN DATA-INTENSIVE RESEARCH CLOUD (SADIRC)

Given the success of the ARC prototype, the next step is the expansion of the ARC to include a greater number of research institutes, including both universities and organisations such as SKA South Africa and the South African National Space Agency (SANSA). A memorandum of understanding was in development at the time of writing, which will formally constitute SADIRC.

In time, it is hoped, SADIRC will expand to offer access to storage for massive data sets, as well as the tools and software to properly collaborate on, analyse and visualise the data – to all South African researchers, including those based at our most under-resourced institutions.

THE ILIFU PARTNERS

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA · UNIVERSITEIT VAN KAAPSTAD

Cape Peninsula University of Technology

SKA AFRICA
SQUARE KILOMETRE ARRAY

SOL PLAATJE UNIVERSITY

UNIVERSITEIT STELLENBOSCH UNIVERSITY

UNIVERSITY of the WESTERN CAPE

ILIFU

The establishment of the ARC meant that UCT was perfectly placed to lead a consortium of institutions in the Western Cape province of South Africa to put in a bid to the National Integrated Cyberinfrastructure System (NICIS), supported by the Department of Science and Technology. The goal of this bid was to build a data-intensive research facility in the Western Cape that would cater explicitly to the needs of researchers working in astronomy and bioinformatics. The bid was successful; and today, this project is known as ‘Ilifu’ (‘cloud’, in isiXhosa).

Ilifu will receive funding from the Department of Science and Technology for a period of three years. It will bring together the existing infrastructure and expertise of the various partner institutions, and build on that to create a hub for data-intensive research systems, platforms and tools in the Western Cape. A further mandate for Ilifu is the development of a research data management system (see the following chapter of this publication).

Within the three-year funding period, Ilifu is set to continue as a self-sustaining facility.

THE SUM IS GREATER THAN ITS PARTS

While the investment in infrastructure is segmented, the offering itself is greater than the sum of its parts. Working together, Ilifu, the ARC and (in time) the SADIRC will provide researchers access – through an online portal – to the entire tiered infrastructure system, as one entity (see box below).

A researcher will thus be able to log in to the online portal, from any device in any location, to access the stored data sets and run the necessary programs to analyse and visualise the data.

NATIONAL INTEGRATED CYBERINFRASTRUCTURE SYSTEM

The NICIS is a national initiative of South Africa’s Department of Science and Technology. The strategy defines different tiers of research infrastructure. Tier III provides institutional infrastructure, and tier II is regional – and, as is the case with Ilifu, involves the collaboration of several universities. Tier I refers to national-level infrastructure.

Research data management:
towards a national solution

Robust, reproducible research outputs are a key measurable of tertiary education institutions, says Dr Dale Peters, eResearch Director and Ilifu Project Co-Lead. Research data must be properly managed, disseminated and measured to ensure the verification of research results; this also allows for new research discoveries from existing data sets. This is why UCT, as lead institution in Ilifu, is working with national partners to develop a solution that will meet the research data management (RDM) needs of all South African tertiary education institutions.

The need for a national research data management solution was brought sharply into focus by the introduction of a data management mandate by the National Research Foundation (NRF) – South Africa's government-mandated research and science development agency, and the very lifeblood of the national academic enterprise.

To meet this mandate, UCT – through the Ilifu consortium – is rolling out a two-phase process as an interim measure, using the cloud-based platform Figshare. The long-term vision is to provide a sustainable national platform for the gathering, publishing and archiving of research data provided by the Data-Intensive Research Initiative of Southern Africa (DIRISA), part of South Africa's National Integrated Cyberinfrastructure System.

FIGSHARE

Figshare is internationally known and trusted software that allows institutions to manage, disseminate and measure the impact of research outputs.

"Figshare will allow us not only to trace all institutional data, but also to measure how research data is being used and reused, and gain insight into the social impact of our institutional research data," says Peters.

BENEFITS OF A FIGSHARE DATA MANAGEMENT PLATFORM

- immediate provision of a robust and stable data management platform
- a citable digital object identifier (DOI) for each institution that enables monitoring of and reporting on compliance with funder requirements
- delivery of data management services to international standards of practice
- resource efficiencies in information technology and library staffing costs associated with alternative open-source software development.

"South African universities have to meet these mandates, no matter what their resources. This solution allows us to work collaboratively to save costs, rather than duplicating effort," says Peters.

PHASE I

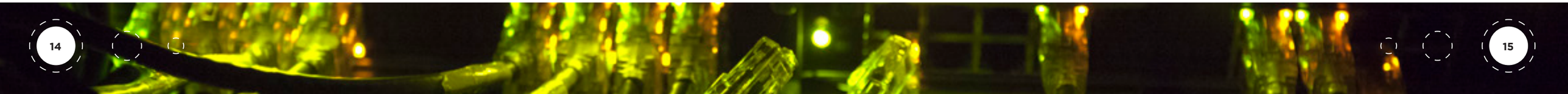
Included in the NICIS tier II proposal – the successful consortium bid to build a Western Cape Data-Intensive Research Facility (now called 'Ilifu') – was a requirement for the development and implementation of research data management systems and tools. A collaborative Figshare licence meets this requirement. Initially, the consortium will comprise most of the Ilifu consortium partners and will be known as the Figshare Consortium.

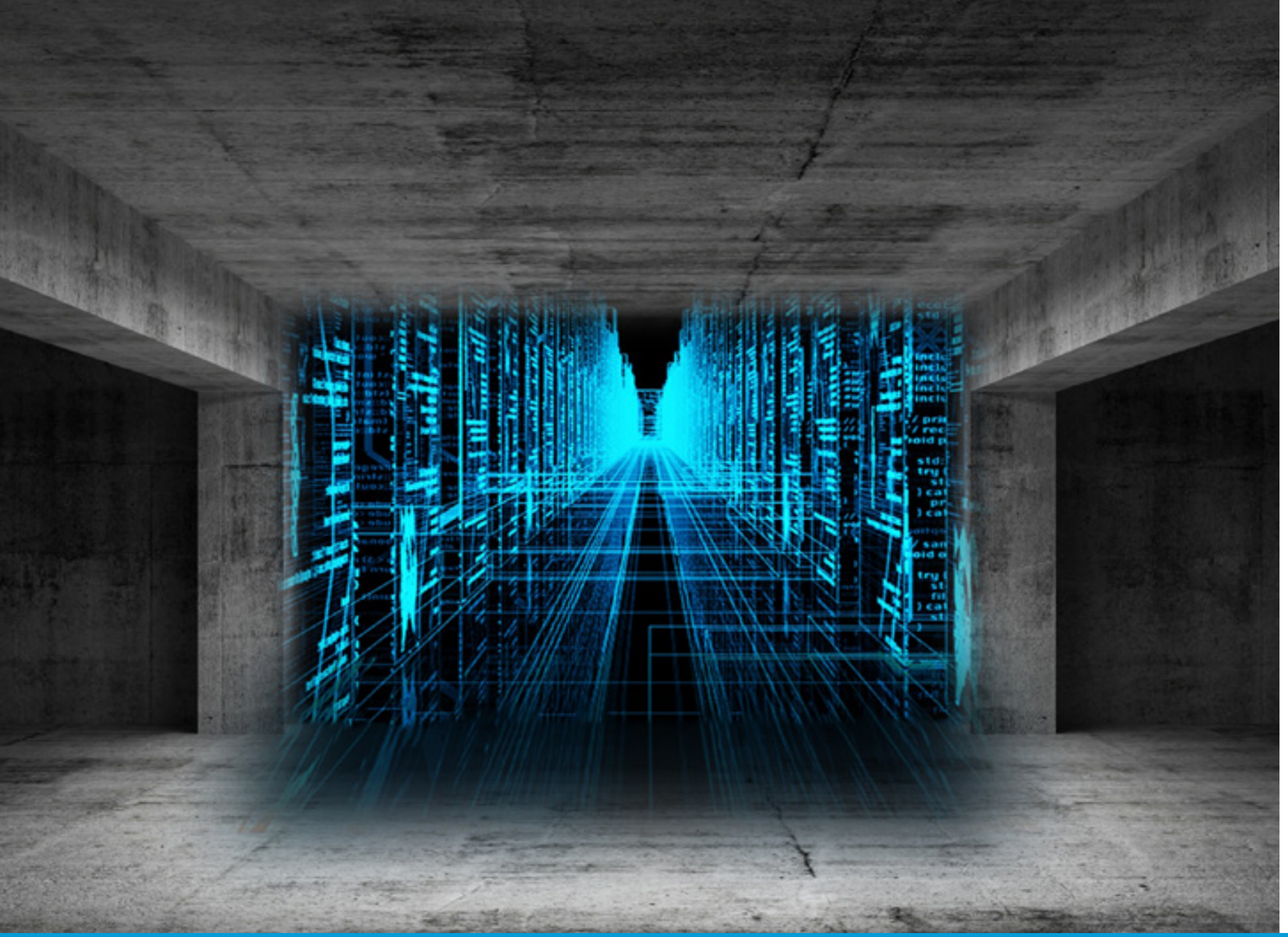
PHASE II

In the second phase of implementation, the Figshare Consortium will be opened up to other national institutions for participation. Collaborating at a national level is the most cost-effective way to produce a tool for South African universities that ensures research data outputs can be effectively managed, in line with funder mandates, within a very short space of time.

THE LONG-TERM SOLUTION

While Figshare offers an excellent short-term solution, provision for effective research data management is to be made within the national cyberinfrastructure system. UCT eResearch and the Ilifu consortium have been working with DIRISA to ensure a sustainable national solution in the long term. The interim solution is intended to last for three years, while the national system is implemented. DIRISA will then be recognised as the repository of first resort for South African research data.





Preparing the next generation for the big-data challenge

Demand for skills at the interface between technology and information is growing, and demand already far exceeds supply. This is particularly the case in Africa. In response to that demand, UCT has launched two new postgraduate programmes, to foster a generation equipped with the skills to lead the continent in meeting this data science need.

MASTER'S IN DIGITAL CURATION

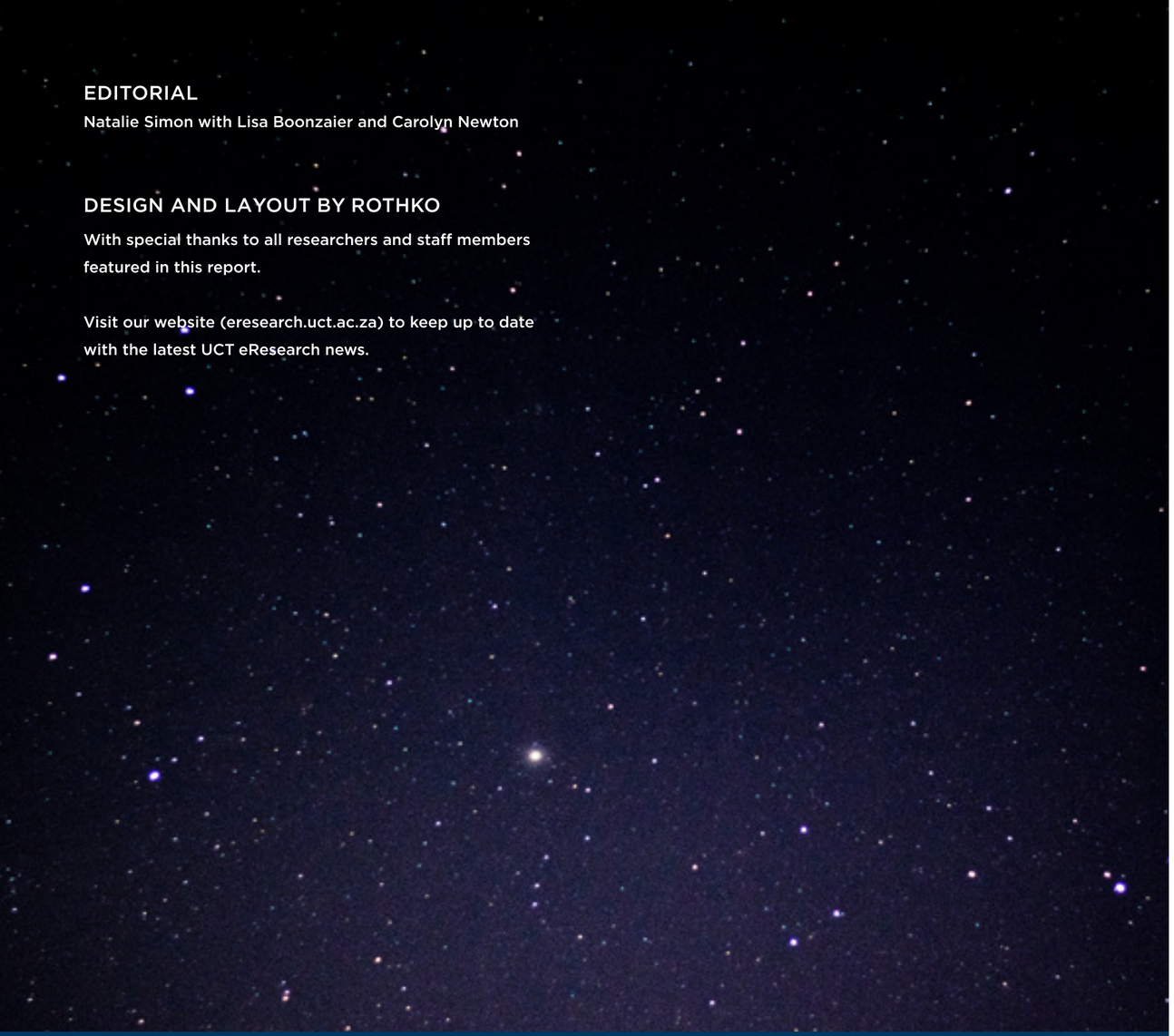
UCT is the first university in Africa to offer a master's-level course in digital curation: that is, the selection, maintenance and archiving of digital data repositories. The course, offered by UCT Libraries and Information Studies Centre, provides its graduates with a comprehensive set of digital curation skills applicable to any sector.

The programme accommodates evolving technology, addresses the ethics and governance of curatorship, and fosters critical leadership in this emerging field in Africa and beyond.

The one-year coursework component of the programme allows for specialisation, following a core course on the principles, theory and philosophy of digital curation.

MASTER'S IN DATA SCIENCE

This interdisciplinary master's degree aims to furnish graduates with the statistical and computing skills needed to deal with big data from the fields of astronomy, physics, medicine and commerce. The university departments collaborating on this master's programme are statistical sciences, computer science, astronomy, physics, computational biology and the African Institute of Financial Markets and Risk Management. The programme comprises two equally weighted components: coursework and a dissertation on a research topic related to data science in astronomy, bioinformatics, computer science, physics or statistical sciences. Students have the choice of two streams for the programme: a general stream and a stream specialising in financial technology.



EDITORIAL

Natalie Simon with Lisa Boonzaier and Carolyn Newton

DESIGN AND LAYOUT BY ROTHKO

With special thanks to all researchers and staff members featured in this report.

Visit our website (eresearch.uct.ac.za) to keep up to date with the latest UCT eResearch news.