UCT eResearch

ACCELERATING RESEARCH

eResearch report 2017–18

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Introductions

Dealing with the data

Building the infrastructure to meet our needs

Supporting the research endeavour

Introductions

Standing on strong foundations to support data-intensive research in Africa

Professor Michael Kyobe
Deputy vice-chancellor for research and internationalisation

In March 2018, UCT implemented its research data management (RDM) policy in response to a changed research landscape. It has become almost standard for funders to demand the open publication of research data. Soon it will be commonplace for journals to ask for the data behind findings to be made public to ensure replicable, verifiable science. And this is a good thing.

We need no reminding that research is a public good. And just as technology has meant that we can today collect data sets previously unimaginable, it has also allowed us to share that data with colleagues anywhere in the world. And not only that – we can share the software used to analyse the data, or the tools to visualise it. It means we can collaborate better, and it also means greater reliability in our research outcomes.

For many, this step is a frightening one. Technology has moved so fast that it has, in less than a decade, changed the way we do a great many things, from collecting our data to communicating with colleagues. Fortunately, as the pages of this report show, UCT has laid the foundations to support this change.

UCT eResearch was formed when the concepts of big data in research and open science were only starting to emerge. Fortunately for the institution, those who came before me, deputy vice-chancellors Danie Visser and Mamokgethi Phakeng, realised that, soon, these concepts would no longer be niche.

Today UCT is ready for data-intensive research and open science. For this I would like to thank all those who have been working tirelessly to build these strong foundations.

UCT eResearch: The journey

Dr Dale Peters
UCT eResearch director

The Greek poet Constantine Cavafy, addressing perhaps the hero Odysseus on his homeward voyage to the mythical island of Ithaka, has a simple message for all of us about appreciating the value of the journey. The eResearch journey has only recently begun, and yet in our haste and eagerness to satisfy the needs of every researcher, we forget that it is the path that can teach us the most, and is also the most enjoyable.

The annual eResearch Report provides such a moment of reflection, an opportunity to savour the satisfaction of the journey expressed in these pages by an ever-expanding group of colleagues across the university. The rapid development of research infrastructure, both locally and globally, has set the research ecosystem on a path of intense transition to open science. Council approval of the research data management (RDM) policy has focused much of our attention in the past year on developing a coordinated research support effort across all stakeholder groups, comprising the Research Office, Information and Communication Technology Services (ICTS) and UCT Libraries.

The research infrastructure too has taken on a new trajectory with the implementation of Ilifu, aimed at developing the first data node for data-intensive research in the national cyberinfrastructure. Locally, our own high-performance computing (HPC) cluster, seemingly approaching the end of its journey, was reviewed by supportive champions and will be upgraded accordingly. Similarly, the network renewal project has been launched and will offer the reward of unprecedented high-speed connectivity.

New services have evolved in the process, notably scientific communication, a specialist role in science journalism profiling research on big-data science projects. This enables eResearch to tell the stories reflected here, highlighting the scientific achievements and the social impact of research undertaken at UCT.

A thank you to those who made us what we are today

Marilet Sienaert,
Executive director,
Research Office

When Sakkie Janse van Rensburg took up the role of executive director of Information and Communication Technology Services (ICTS) at UCT in 2009, almost no relationship existed between the research endeavour and ICTS. Fortunately for UCT, Sakkie, from his tenure in the same role at the University of the Free State, understood the importance of advanced IT support for research and set us on the path toward establishing UCT eResearch.

> There is no doubt that without Sakkie's initial vision – and his determination to implement that vision – UCT eResearch would not be the leader it is today.

> Under Gwenda's guidance, UCT Libraries transitioned from its traditional role as custodian of books and paper journals to an institution that recognises that libraries are at the heart of the academic community and are needed to provide support to researchers – even in the digital age.

> After making their impact felt at UCT, and changing the research endeavour at this university – and arguably in this country – for the better, both Sakkie and Gwenda are moving on to the next chapters of their lives. While they will both be sorely missed, the work they have begun will be continued and will ensure that our researchers are fully supported to face the challenges of research in the digital age.

Sakkie started setting up a high-performance computing (HPC) service at ICTS in 2009. Over the years, this was expanded into a broader project to support the research endeavour, in the form of a cross-departmental partnership between ICTS, the Research Office and UCT Libraries. This collaboration was formalised in 2014 with the launch of UCT eResearch, which continues to provide a core service to researchers today. There is no doubt that without Sakkie's initial vision – and his determination to implement that vision – UCT eResearch would not be the leader it is today.

I also cannot emphasise enough the vital role UCT Libraries played in this innovative research support space – a role in which the leadership of UCT Libraries executive director, Gwenda Thomas, was pivotal. Under Gwenda's guidance, UCT Libraries transitioned from its traditional role as custodian of books and paper journals to an institution that recognises that libraries are at the heart of the academic community and are needed to provide support to researchers – even in the digital age. Under Gwenda's watch, UCT Libraries upgraded their learning spaces, introduced a collaborative research space to host the eResearch data visualisation wall, upskilled their teams, created an online presence and introduced new data management services to better support researchers. This includes the establishment of the Digital Library Services (DLS), and the development and implementation of the research data management (RDM) policy as a response to the changing international landscape where research data is concerned.

Dealing with the data

Opening up the research enterprise through data publication at UCT

Today, technology offers new ways of not only acquiring but also sharing and storing research data, allowing for greater collaboration among researchers as well as a more rigorous scrutiny. The result is a global movement towards greater openness in science. As part of this, UCT, in March 2018, implemented its research data management (RDM) policy to support effective data sharing and to address the need for data to be findable, accessible, interoperable and reusable (FAIR) to specific quality standards.

“Open science – in particular, making the data on which the science is based freely available – is a response to the notion that university research is a public good and should be publicly available,” says Dr Dale Peters, UCT eResearch director. “In addition, funders are mandating data publication so they don’t repeat-fund research, and journals are mandating it so that results of publications can be verified.”

Beyond the public good, there are other enticing reasons why researchers should want to publish their data openly. The first is for citations, says Niklas Zimmer, headof UCT Libraries’ Digital Library Services (DLS).

“Data is now another thing you can be cited for. Open-access publishing of the data means it will be found and reused by other researchers in your field, and you will be credited for this.”

> Data is now another thing you can be cited for. Open-access publishing of the data means it will be found and reused by other researchers in your field, and you will be credited for this.

The open publication of data does require changes in behaviour. Researchers now need to think through issues of data management they may not have considered before, and spend time on data curation. However, these improved data management practices may impact positively on overall research outcomes, both now and in the future.

Curation is the process of organising data according to logical standards. High-quality curation allows for the long-term preservation and accessibility of data. It can include activities such as regular, secure backup and archiving, or using open formats that will survive software and technological changes and so remain accessible in the long term.

"The advantages of properly curated data are twofold," says Thomas King, data curation officer at DLS. "First, it means your data won't be lost. Second, if it is organised and described in an understandable and logical way, it can be reused by colleagues and students."

An additional advantage of the open publishing of data for reuse is the added value, says Lynn Woolfrey, operations manager at DataFirst, an open data repository at UCT. It has become common to refer to data as the new oil, but, says Woolfrey, this comparison is not accurate.

"To quote Adam Schlosser of the World Economic Forum," she says, "data is not the new oil, because the attitude of scarcity does not apply to it. No one wants to share oil wealth, but data gains value the more it is shared."

> Open inquiry is at the heart of the scientific enterprise. Publication of scientific theories – and of the experimental and observational data on which they are based – permits others to identify errors, to support, reject or refine theories and to reuse data for further understanding and knowledge. Science's powerful capacity for self-correction comes from this openness to scrutiny and challenge.
>
> – The Royal Society Science Policy Centre

Open science support at UCT

UCT offers software and support to researchers throughout their research projects.

DMPonline

Increasingly, funders require the submission of a data management plan (DMP) in the early stages of a research project. DLS hosts DMPonline, a tool developed by the Digital Curation Centre (UK) to enable researchers, data managers and principal investigators to complete their DMPs with appropriate user guidance provided via the platform.

> When you publish data on ZivaHub, a persistent identifier – a digital object identifier (DOI) – is created. This makes the data recognisable as belonging to UCT and can be used to identify the data, irrespective of where it sits.

Open Science Framework

The Open Science Framework is a project management repository – created by the Centre for Open Science – which serves as a collaboration tool.

It allows researchers to either work on their projects privately, with a limited number of collaborators, or to make their projects completely open.

DLS has set up a UCT instance of the Open Science Framework that researchers can freely and securely make use of with their UCT credentials.

ZivaHub

ZivaHub, the institutional data repository of UCT, runs on the cloud-based Figshare platform. Its mission is not to replace existing or discipline-specific data repositories, but rather to offer a service for any staff member or student who needs to openly publish data.

When you publish data on ZivaHub, a persistent identifier – a digital object identifier (DOI) – is created. This makes the data recognisable as belonging to UCT and can be used to identify the data, irrespective of where it sits. The platform also offers metrics, allowing researchers to know how often and where their data has been viewed, downloaded and cited.

An added bonus of publishing on ZivaHub is the institutional support from data curation officers. While this team cannot edit or make changes to the microdata – the data provided by the researcher – they can assist in the describing of the data (metadata) to ensure it is correctly categorised and labelled for maximum discoverability and likelihood of reuse.

DataFirst

DataFirst is an open research data repository based at UCT which also holds socioeconomic data from a number of African governments and research institutions. The data is easily accessible to researchers and policy analysts worldwide.

DataFirst:

- offers subject-specialist support with the anonymisation and special preparation required for microdata sharing
- works with large-scale university projects and government agencies to encourage them to deposit their raw data for further use
- quality-checks and anonymises data
- provides an open data site where researchers can read about and download data
- in the case of sensitive data, allows for the use of the data in their secure centre at UCT
- supports data users
- trains African researchers in data analysis.

Other data repositories

There is also a range of discipline-specific data repositories and alternatives to ZivaHub, for example Zenodo. DLS have set up a UCT community on Zenodo with which UCT-related data publications can be associated.

Researchers are encouraged to use any data repository that best suits their needs.

OpenUCT

Researchers can publish their research in UCT's open-access repository: OpenUCT. They can hyperlink their papers to the data sets published on ZivaHub, and vice versa.

ZivaHub uptake in numbers

92 USERS

Individuals using ZivaHub to publish

UPLOADED **110** ITEMS

Number of items uploaded to ZivaHub by UCT researchers

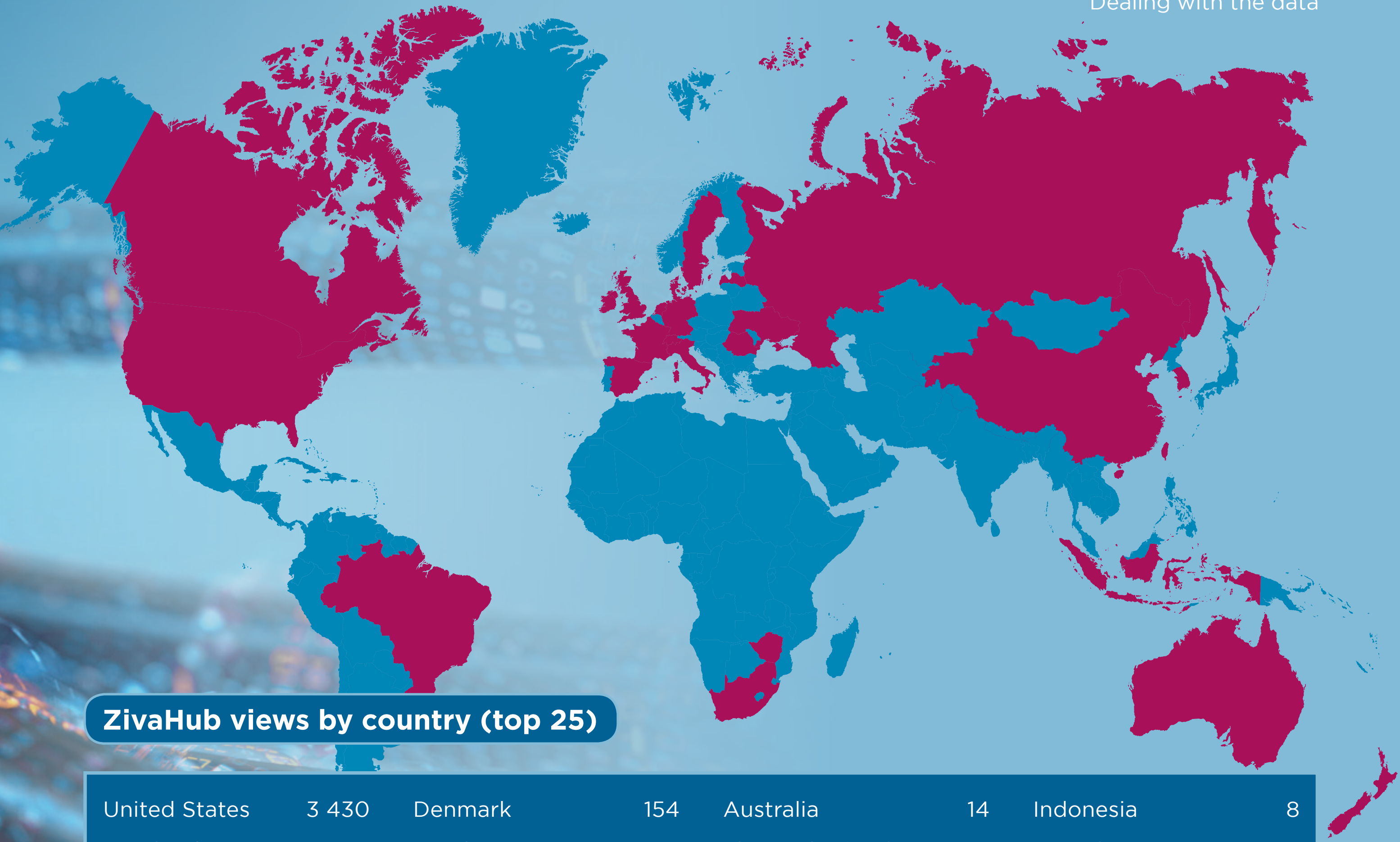
DOWNLOADED **1 874** TIMES

Number of times an item is downloaded

VIEWED **6 887** TIMES

Number of times ZivaHub content is viewed by internal and external users

PRESENTATION	MEDIA	FILE SET	DATA SET	PAPER	POSTER
VIEWED **2 198** TIMES	VIEWED **1 422** TIMES	VIEWED **1 356** TIMES	VIEWED **558** TIMES	VIEWED **374** TIMES	VIEWED **93** TIMES



ZivaHub views by country (top 25)

United States	3 430	Denmark	154	Australia	14	Indonesia	8
South Africa	414	South Korea	59	The Netherlands	14	Canada	5
United Kingdom	382	Latvia	39	Brazil	13	Ireland	5
China	379	Sweden	38	France	13	New Zealand	5
Germany	322	Zimbabwe	36	Italy	10		
Russia	300	Ukraine	19	Spain	10		
Unknown	270	Romania	18	Switzerland	9		

Numbers from 08 August 2018

Understanding open education in the Global South

Research on Open Educational Resources for Development (ROER4D) in the Global South set out to understand how, and under what circumstances, the adoption of open educational resources (OERs) could address the demands for high-quality and affordable education in the Global South. The project – which ran between 2013 and 2018 – included research from countries in South America, Africa and Asia. UCT served as the project's network hub and host, in conjunction with Wawasan Open University in Malaysia.

Image by World Bank Photo Collection via Creative Commons.

We needed an independent space that was agnostic in terms of where the authors were based and what types of content it could accommodate," she adds. The other key requirement was that the platform could accommodate an open-access licensing imperative.

The research, funded by Canada's International Development Research Centre (IDRC), was the first of its kind in terms of scope. It aimed to set up an empirical baseline to highlight a Global South perspective on OERs.

"A project with this kind of mandate must have the proper curatorial and publishing agenda," says Michelle Wilmers, the project's curation and dissemination manager.

They needed a repository that could host the entire ROER4D collection, and after evaluating their options, decided on Zenodo. "We needed an independent space that was agnostic in terms of where the authors were based and what types of content it could accommodate," she adds. The other key requirement was that the platform could accommodate an open-access licensing imperative.

"The value of Zenodo is that it doesn't discriminate in terms of content, source or type," says Wilmers. "And, importantly, because it is backed by CERN and is part of the Horizon 2020 European Union initiative, there is a strong sustainability factor."

In addition, they placed UCT-authored content in the OpenUCT repository. The project's microdata was published through DataFirst, which provided valuable specialist data curation and publishing expertise.

Once ZivaHub had been established by UCT Libraries' Digital Library Services (DLS), the ROER4D curation and dissemination team collaborated with DLS to harvest the records and actual research objects from Zenodo into ZivaHub in order for these to become part of the formal UCT collection.

"We relied on an important principle from the libraries world – LOCKSS: lots of copies keeps stuff safe," says Wilmers. "But this principle must be applied strategically. It cannot involve just putting content in different places; one must consider aspects such as version control and ensuring that different environments speak to each other."

The goal, she says, is to have a strategic publishing and curation approach, which needs to be carefully planned from the start.

Key lessons learned in building a research database

> The first question for research groups to ask is: What is the strategic value of data in our institution?

As early as 2014, before data was widely recognised as a major challenge to research, the Energy Research Centre (ERC) realised they had a data problem. The group hired Wiebke Toussaint, engineer and data scientist, to manage their data dilemma. Toussaint – with some help from UCT eResearch – built an energy data portal. Now, after a five-year journey of learning about research data management for medium-sized research centres, she has some advice for research groups setting out on a similar path.

“Diverse data assets – from big data to small qualitative surveys – play an important role in scientific research, yet many research centres lack the capacity and technology expertise to build data ecosystems to manage their data sets,” says Toussaint.

She advises that research centres, before investing resources or seeking funds to build a data solution, consider a few key factors.

1. Think strategically

The first question for research groups to ask is: What is the strategic value of data in our institution? They need to decide what role they want to play, as a research group, in the national, continental and global space in terms of data.

Toussaint says there can be a massive strategic advantage in deciding to make data management part of the group’s mandate. “For instance, the University of California, Irvine owns a number of data sets that are used globally for benchmarking. There’s no reason our research groups couldn’t do something similar.”

This is important to decide up front, she says, because if you wish to go this route, you should plan at least five or 10 years into the future, rather than simply a year or two.

Toussaint advises research groups to get expert strategic advice at this early stage, to ensure that they have considered all options and know what they are working towards.

2. Map out the landscape

“Data sharing requires multi-party involvement, and the more partners there are in the endeavour, the greater the chance of future sustainability,” says Toussaint.

She strongly advises against a research group taking on such a project alone. “If there is an existing initiative, join it; otherwise, start building partnerships.”

3. Be ready for a culture change

“Data tends to play a very strategic role in a research centre, with many different touchpoints. Changing the way data is managed often involves a component of organisational change,” says Toussaint.

> Researchers must recognise that this will affect how they work with their data and must be open to changing some practices.

An important factor to consider, therefore, is the mandate of the person or people employed to implement sustainable data practices. Researchers must recognise that this will affect how they work with their data and must be open to changing some practices.

4. Know what skills you need

For a database such as the one the ERC built, a group would effectively be hiring for three different skill sets, which may not come neatly rolled up in one person.

First, says Toussaint, you need a person with the technical skills to build the database – somebody with an engineering mindset who is happy to tinker and problem-solve.

Once it is built, you need to employ a person long-term to maintain it: somebody with a librarian mindset to focus on the fine details around curating and archiving. Finally, you need someone to do the "data science" – the visualisation and storytelling from the stored and curated data set. Toussaint is extremely excited that researchers are starting to grapple with these problems: it is new and difficult, but she believes it is important to send the message that this is not a space to be feared.

"Data is only as valuable as the value people add to it, and if yours is inaccessible, there is other data out there that researchers will use, while your untouched data grows stale," she says.

How the portal was built

Toussaint was tasked with building a data portal where the ERC's energy data – currently being gathered from a range of sources and scattered across different storage options – could be stored in one place. In addition, the data had to be made open to researchers worldwide.

An early pioneer, before ZivaHub was available, Toussaint relied on web-based open-source software called Comprehensive Knowledge Archive Network (CKAN), created specifically for storing and distributing open data.

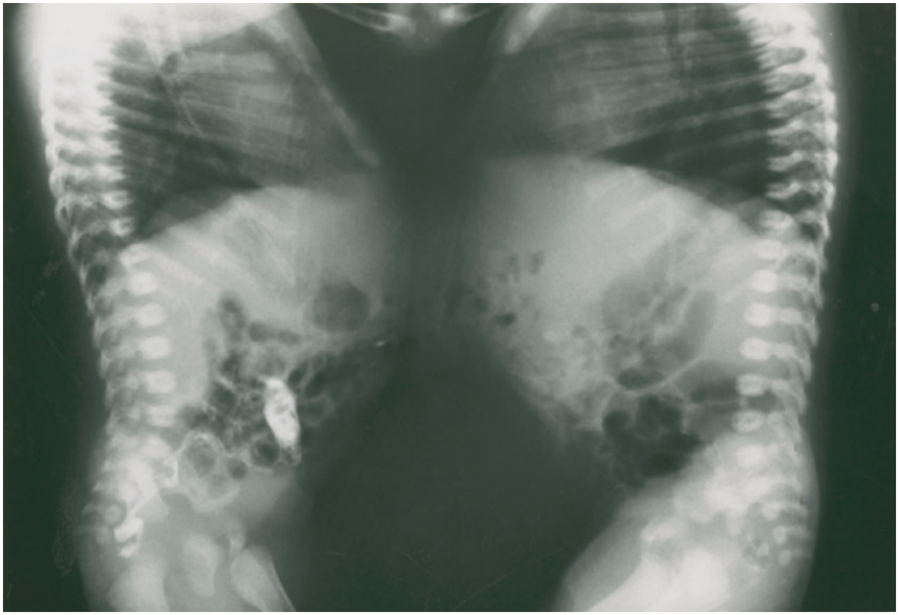
"CKAN is a great resource," says Toussaint. "It's not an out-the-box solution, as setting it up requires technical know-how, but it's free."

UCT eResearch assisted by providing a virtual server. Toussaint says her interactions with UCT eResearch – and their support – facilitated her work.

Toussaint adds that she hopes to collaborate more with eResearch in the future: "As the project comes to an end, and I reflect on what I've learned, I have a greater idea of what's actually possible. I would like to engage with eResearch at the next level, and say: You're doing great work, and changing the way things are done; what else can you do for me?"

Rediscovering a common history: Digitising a clinical photography collection

UCT's medical school, which is also South Africa's oldest, has for more than 100 years trained doctors, treated patients and advanced medical science. While the walls of this pre-eminent school and its teaching hospital will never speak, a collection of recently uncovered images offers fascinating insights into not only the history of medicine in South Africa, but also our society across decades. Working with UCT Libraries' Digital Library Services (DLS), researchers at the Pathology Learning Centre (PLC) are ditgitising this collection to make it public so it can serve as research material for a variety of disciplines.



This X-ray is of a pair of conjoined twins, who were later successfully separated in the first operation of its kind in South Africa (Red Cross Chidren's Hospital, 1966). The twins were joined at their abdomens. Before the surgery could be done, many investigations were needed to establish whether internal organs were shared and whether separation would be possible.

The collection was discovered by Dr Jane Yeats, director of the PLC, when she requested records from the Department of Surgery related to another project. Along with the material she requested, she received boxes of files packed with cards displaying photographs of body parts, organs and surgeries; X-rays; illustrations from period textbooks; and, mainly, patient photographs – all annotated. In among these were X-rays of conjoined twins, photos taken four years apart of a woman who had undergone oesophageal reconstruction, and a photo of a young child – his face in shadow – with a parasitic cyst on his spine. When the full collection was recovered from a storeroom, the cards totalled around 7 000. They detailed surgical procedures, documented methodologies and showed the people and pathologies moving through the medical school and hospital between the 1920s and 1970s.

> This collection will be of value to disciplines such as medical history, medical anthropology, sociology, arts, politics and more.

Yeats realised the value of what she was looking at and decided the images needed to be digitised. “This collection will be of value to disciplines such as medical history, medical anthropology, sociology, arts, politics and more,” she says.

In addition to Yeats, the collection has a champion in Michaela Clark, a visual studies graduate from the University of Stellenbosch, who was appointed as a research assistant to do the archiving.

Getting the ball rolling

The problem was clear to Yeats: How do we give other researchers access to this resource, which has so much research potential?

The solution seemed to lie in a digital repository that could allow researchers – from any discipline and anywhere in the world – to access the information.

One of the difficulties was around confidentiality. Some of the patient photographs are revealing, and most of the people are identifiable, says Clark.

“I approached eResearch, and they put us in touch with DLS,” says Yeats. “That got the ball rolling.”

A solution of two parts

Kayleigh Lino and Erika Mias, digital curation officers at DLS, recommended a dual-system solution using the web-based open-source applications Access to Memory (AtoM) and Omeka.

The entire collection is described in AtoM@UCT, but, for confidentiality, visual examples are only provided when they do not reveal the patients’ identities. The collection is searchable and indexed, which means researchers can query, group and organise the collection in different ways. This is key, says Clark, as it allows researchers to identify patterns in the data.

Omeka then adds another layer to this archive: curation. “Omeka gives us a blank slate to talk about the pictures and show the research possibilities,” says Yeats. It is on Omeka that the PLC presents the curated selection of images as exhibits.

This collection of historical clinical photographs has fascinating stories to tell, says Yeats. “By carefully framing these images in exhibitions on Omeka, and describing them in AtoM@UCT, we hope to offer a respectful engagement with our material to showcase its value beyond the medical field.”

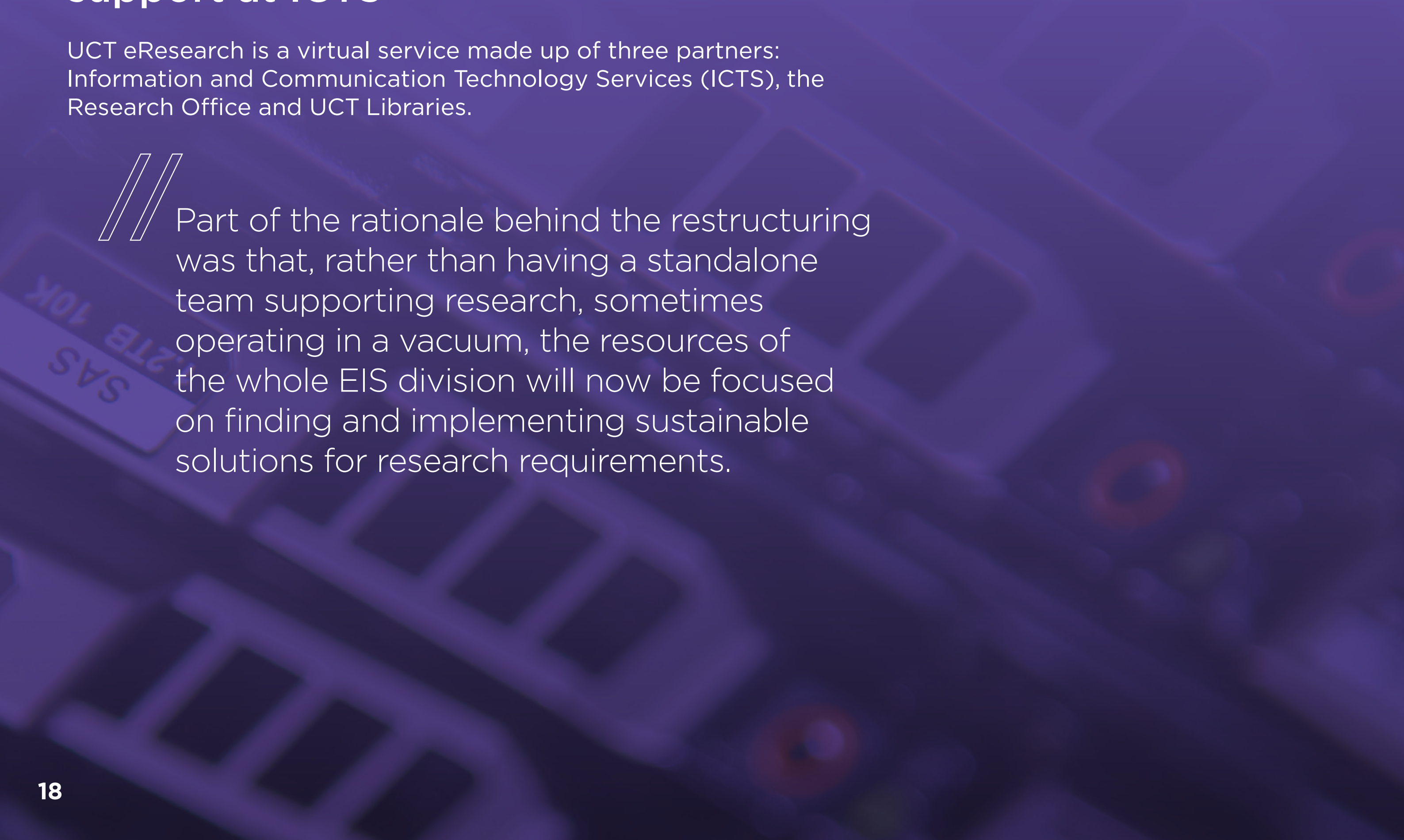
See the online exhibitions at surgeryclinicalphotos.uct.ac.za

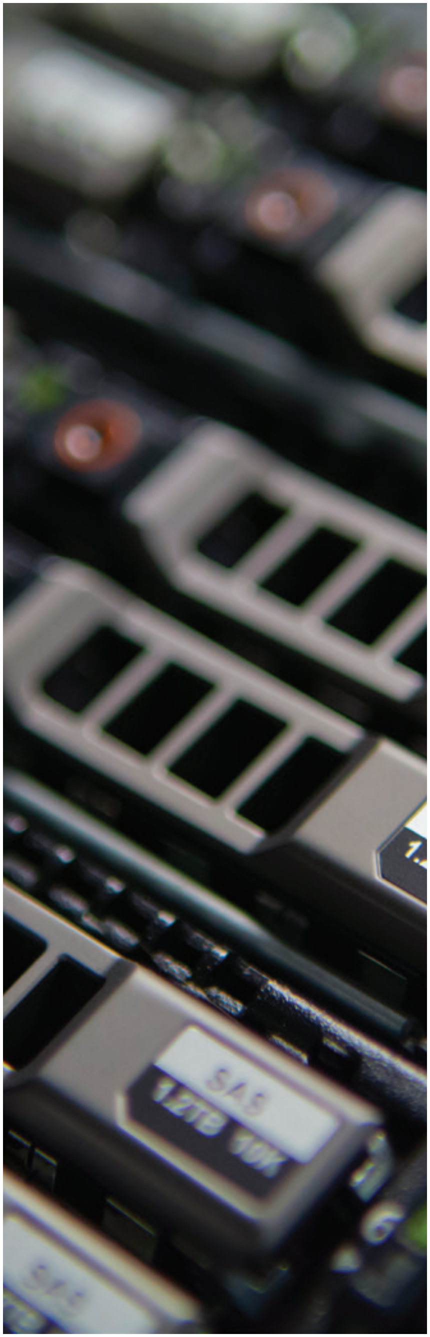
Building the infrastructure to meet our needs

EIS: A holistic approach to research support at ICTS

UCT eResearch is a virtual service made up of three partners: Information and Communication Technology Services (ICTS), the Research Office and UCT Libraries.

> Part of the rationale behind the restructuring was that, rather than having a standalone team supporting research, sometimes operating in a vacuum, the resources of the whole EIS division will now be focused on finding and implementing sustainable solutions for research requirements.





> Rather than implementing parallel solutions for individual research groups, we are now better able to ensure UCT has the infrastructure, tools and training available to underpin the full cycle of research data.

ICTS has, through eResearch, supported researchers with their advanced IT requirements since 2014, when UCT eResearch was first established. As this advanced IT support for research grew, so did demand. In 2017, to ensure that ICTS could keep up with the demand coming from the research endeavour, among other things, the divison which housed the IT arm of eResearch was restructured into what is now known as Enterprise Infrastructure Services (EIS).

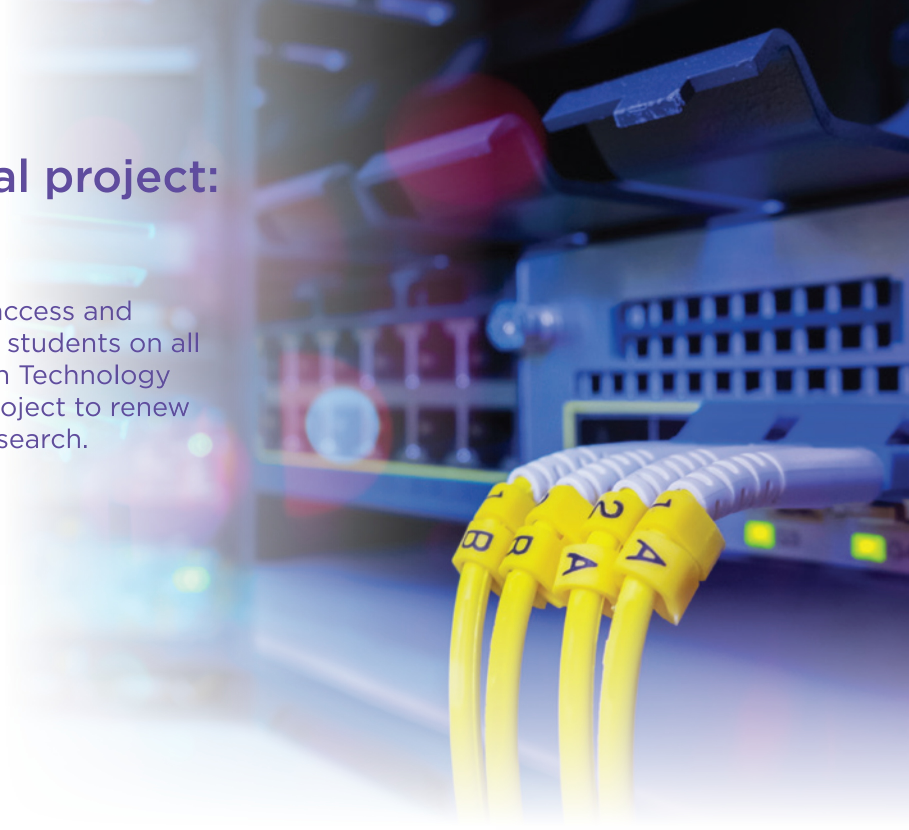
Part of the rationale behind the restructuring," says Andre le Roux, director of EIS, "was that, rather than having a standalone team supporting research, sometimes operating in a vacuum, the resources of the whole EIS division will now be focused on finding and implementing sustainable solutions for research requirements."

Dr Dale Peters, eResearch director, explains that the restructuring is aimed at ensuring that research solutions are scaleable and support the greater research endeavour. "Rather than implementing parallel solutions for individual research groups," she adds, "we are now better able to ensure UCT has the infrastructure, tools and training available to underpin the full cycle of research data."

In terms of the new EIS strategy, researchers will continue to engage with the director and eResearch analyst for their needs. The researcher requirements will then be picked up by the appropriate unit in the EIS division and the full resources of that unit will be devoted to providing the needed infrastructure, services or support.

The ICTS network renewal project: Benefits for research

The UCT network brings storage, internet access and other services – such as WiFi – to staff and students on all campuses. Information and Communication Technology Services (ICTS) is embarking on a major project to renew the network – with a positive impact for research.



The network renewal project will enable ICTS to ensure a continued high level of network quality, even as demand grows, as well as improved network performance and security.

As network renewal is a vast and complicated undertaking, the work has been divided up into various phases. The first phase kicked off in 2018, and the project is expected to be completed by 2020.

Researchers will experience a number of benefits thanks to the project:

- **Enhanced security**: With major data breaches becoming more common, security is no longer an afterthought. World-class threat detection and security mechanisms will be built into all layers of the new network, helping to safeguard UCT data, devices and systems.
- **Better performance:** Increasingly, research relies on a strong network connection as data sets become more voluminous and cloud technology is used more often for both storage and software. The new network will cater for both current and future needs.
- **Improved efficiency:** Standardisation and automation will boost efficiency, while the new network will also lay the foundation for future smart technologies.

High-performance computing upgrade

> I can honestly say that it is thanks to the support of the HPC team over the years that my research reached the level where it qualified for such significant funding from Pfizer.

In 2009, Information and Communication Technology Services (ICTS) took the first step towards dedicated support for research through establishing a centralised high-performance computing (HPC) resource to provide a reliable, scalable and economic computing facility for UCT's researchers. Nearly 10 years later, ICTS is embarking on a major project to upgrade the facility – with significant investment from research funding.

"Hex, the original HPC cluster that has served us for nearly seven years, was no longer fit for purpose," says senior technical specialist Andrew Lewis. "And we were experiencing a dip in researcher uptake."

The decision was taken in 2017 to upgrade Hex. The new cluster is currently being installed and is expected to be fully operational by November 2018. It will be split into several partitions, featuring newer, faster nodes; graphic processing units (GPUs); high-memory nodes; and the older Hex nodes, which will be available for teaching and low-priority jobs.

The project aims to provide an HPC cluster geared towards high-profile researchers with major computing needs, but which is still accessible and useful to most researchers. The cluster will have to be both scalable and economical, especially in terms of power consumption and heat dissipation.

A research investment

Associate Professor Michelle Kuttel, who has been using the HPC facilities for years, has made a significant investment in the cluster with a funding windfall she received from Pfizer for her work on molecular modelling for vaccines. She has bought GPU nodes which will be integrated into the new cluster. "I am a big believer in the shared model for computing resources," says Kuttel.

In this model, researchers can outsource the management of the facilities – a time-consuming, highly skilled and expensive endeavour – while still enjoying the benefits of access. Kuttel also notes that, no matter how data-intensive the research, a machine allocated to a single project is likely to have a lot of downtime, which could be used by other researchers.

The support that comes with the centralised HPC facility is invaluable, she adds.

"I can honestly say that it is thanks to the support of the HPC team over the years that my research reached the level where it qualified for such significant funding from Pfizer."

Connecting rural campuses

The Faculty of Health Sciences boasts a number of rural sites where important research and teaching are undertaken. Up to now, these sites have had to make do with the levels of connectivity in the region. In 2018, Information and Communication Technology Services (ICTS) concluded the UCT segment of the Rural Campuses Connection (RCC) project, part of a national initiative to connect the country's rural research facilities.

"The RCC project was very rewarding in terms of customer satisfaction," says Bruce Fielies, senior manager of Workplace Services at ICTS. "Researchers are used to working on campus and take this connectivity for granted. Then they go to rural research sites and they're confronted with major challenges around internet speeds and instability of connection."

Fortunately, the lack of access to adequate broadband in rural areas was identified by the Tertiary Education and Research Network of South Africa (TENET). They procured a grant from government to connect the 53 mostly rural higher education research and teaching sites in South Africa to the South African National Research Network (SANReN).

Once TENET secured the funding, universities submitted their rural sites for approval as beneficiaries. TENET then covered the installation costs and two years' maintenance costs to keep these sites on the network.

At UCT, ICTS has connected a total of nine rural sites so far, with four still in progress. Researchers can now easily access the internet and collect and upload data, with little or no connectivity interruptions. Students at these campuses can also access their lectures through video conference facilities.

"We have had very positive feedback from researchers and students," says Riedewaan Jacobs, senior technical officer at ICTS. "The speeds offered by the new infrastructure are significant, which enables us to deliver the same services available on upper campus."

Cloud computing for data-intensive research

Ilifu, which means cloud in isiXhosa, is a shared big-data cloud infrastructure for data-intensive research. The goal of Ilifu is to enable South African researchers to be global pioneers in the strategic science domains of astronomy and bioinformatics. Operated by a consortium of universities and research organisations in the Western Cape and the Northern Cape, Ilifu is a regional node in the national data infrastructure, partly funded by the Department of Science and Technology, to support the National Integrated Cyberinfrastructure System (NICIS) of South Africa. Ilifu brings together the existing infrastructure and expertise of the partner institutions and builds on that to create a regional hub for data-intensive research.

The universe in a cloud

The MeerKAT telescope, a precursor to the Square Kilometre Array (SKA) telescope in South Africa, will enable astronomers to try to answer age-old questions about our universe. However, in order to do that, we have to solve the accompanying big-data challenge. To this end, the Inter-university Institute of Data-Intensive Astronomy (IDIA) built a research cloud, which is the first building block of Ilifu.

IDIA – a partnership between UCT, the University of the Western Cape (UWC) and the University of Pretoria (UP) – is an important role player in Ilifu. The IDIA research cloud served as a prototype of the technical foundation of the Ilifu facility.

The IDIA data-intensive research cloud is tailored to process the huge data expected from the MeerKAT radio telescope. Beyond the mere storage of raw data, the cloud-computing facility allows researchers to use their own tools to process data on an unprecedented scale.

And as the cloud software systems are all built on open-source technologies, these tools can be reused and recreated in other disciplines.

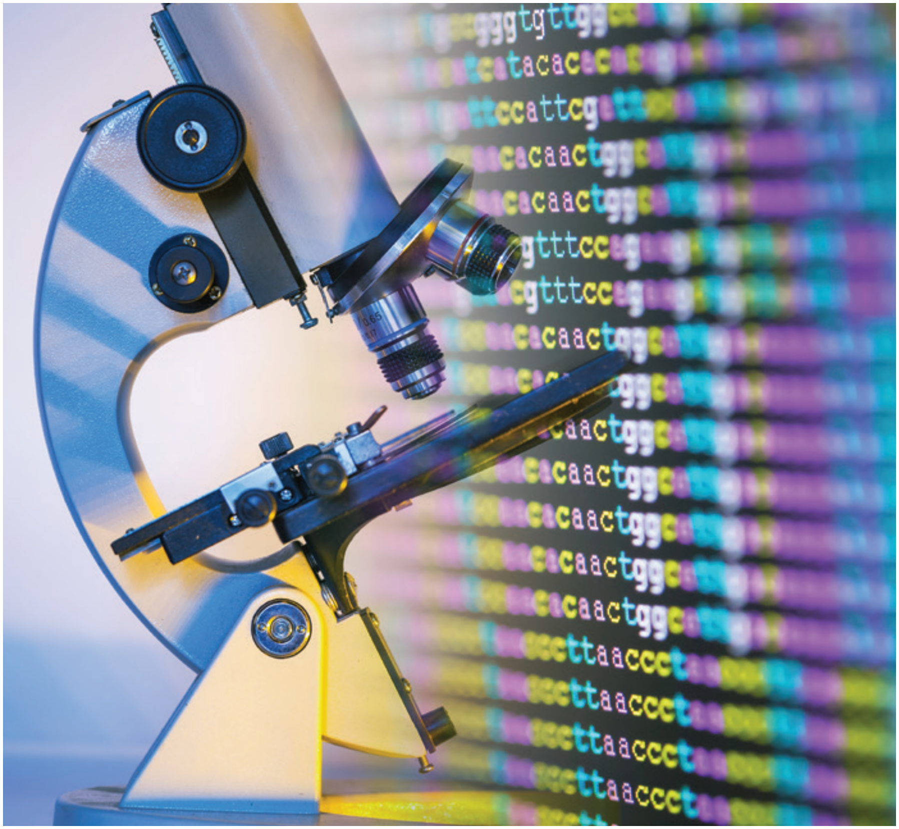
The original IDIA cloud infrastructure will now be expanded into a single, larger data-intensive research cloud that is Ilifu.

A bioinformatics investment to expand Ilifu infrastructure

H3ABioNet, the Pan African Bioinformatics Network for H3Africa (Human Heredity and Health in Africa), is already a strategic science project on Ilifu. However, for the project's principal investigator, Professor Nicola Mulder, the value of the shared resource is so great that when H3ABioNet developed a need for extra computing power – and significant funding to pay for it – she and her team decided to invest that money in Ilifu to further expand the infrastructure.

"H3ABioNet designed a genotyping array for African populations, which has been manufactured by Illumina," says Mulder. "Our project now offers a service called imputation, which, in a nutshell, adds value to their data, but this is very computationally intensive."

> You need a lot of resources to manage a system such as this, and you need them long term," she says. "We don't have those resources, and also, with the shared facility we know that when we are not using the machines, they will be used by others.



Their work requires a very reliable computing resource, she says – one that does not have a lot of downtime and, importantly, where her team of bioinformaticians do not have to worry about the maintenance of the machines.

"You need a lot of resources to manage a system such as this, and you need them long term," she says. "We don't have those resources, and also, with the shared facility we know that when we are not using the machines, they will be used by others."

Mulder says it is also great for her team to collaborate with astronomers around infrastructure for big-data needs, because the astronomy community has been grappling with big data for so long. It means other disciplines don't need to reinvent the wheel.

Supporting the research endeavour

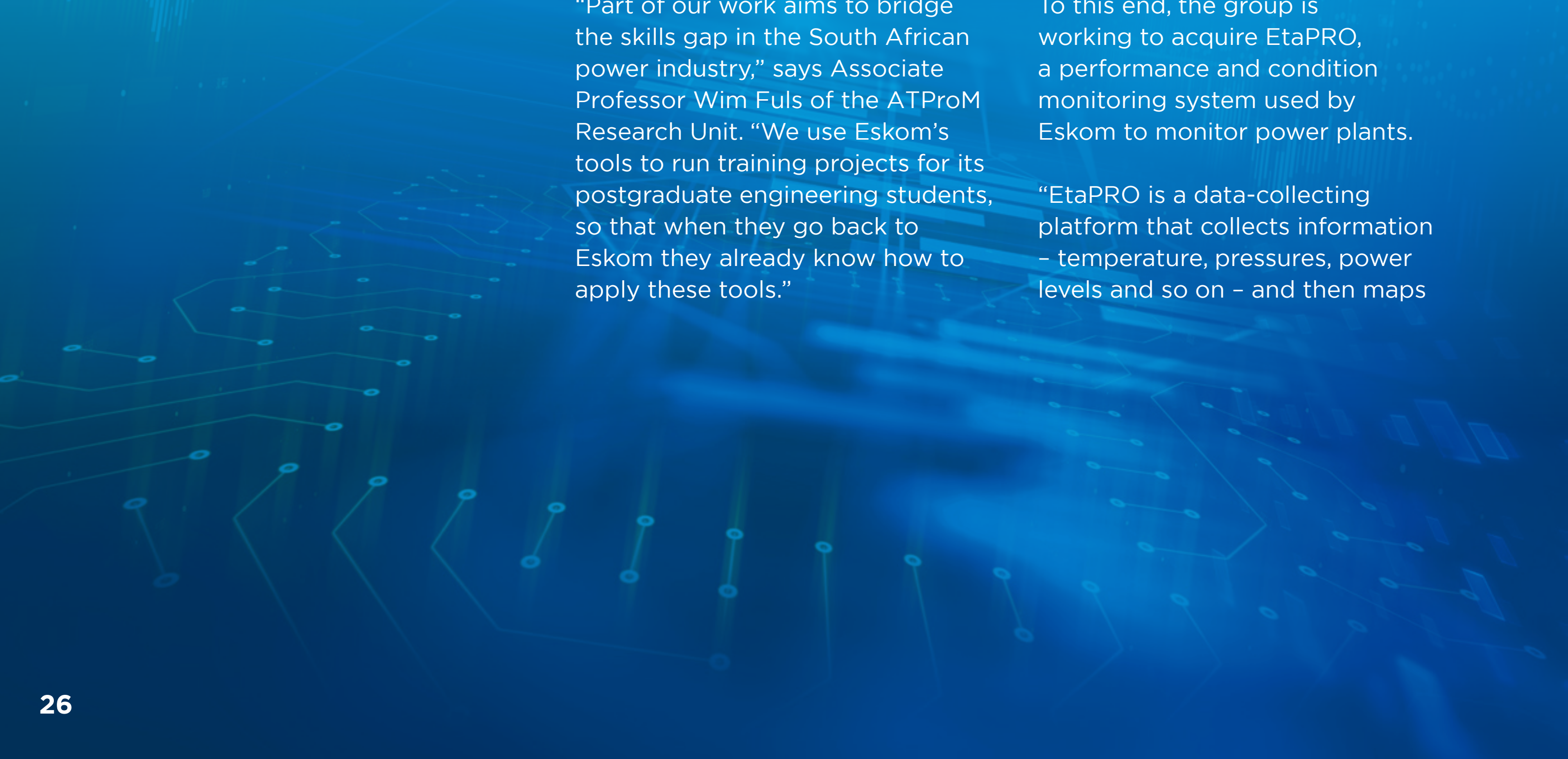
Data for learning: Training the next generation of engineers

The Eskom Specialisation Centre for Energy Efficiency, hosted by the Applied Thermofluid Process Modelling (ATProM) Research Unit at the Mechanical Engineering Department, trains and mentors engineers for South Africa's power utility, Eskom. To do this they need access to the tools and platforms Eskom uses. Some of these tools require extra technical support, and it is here that eResearch assisted.

"Part of our work aims to bridge the skills gap in the South African power industry," says Associate Professor Wim Fuls of the ATProM Research Unit. "We use Eskom's tools to run training projects for its postgraduate engineering students, so that when they go back to Eskom they already know how to apply these tools."

To this end, the group is working to acquire EtaPRO, a performance and condition monitoring system used by Eskom to monitor power plants.

"EtaPRO is a data-collecting platform that collects information – temperature, pressures, power levels and so on – and then maps



this against what the conditions of the power plant theoretically should be," explains Fuls. "It gives the operator and engineers an indication of the condition of the plant and early warnings if it looks like things may go awry." EtaPRO also facilitates pattern recognition, using data collected over years to identify trends.

Giving students access to such data on power plants will be very valuable for research and training: it means they can start playing with the data and developing useful algorithms to streamline processes while still in the university environment.

UCT eResearch is providing valuable support in the form of a server to house the EtaPRO platform and data. Once licensing is finalised, students will be able to access the platform by simply logging in.

"It is not necessarily a big-data environment, but the server will be critical for the EtaPRO platform and display," Fuls explains. Looking forward, Fuls says the group will likely be moving into the big-data space as they work more on pattern recognition and machine learning. This may well mean great collaboration with eResearch down the road.

Understanding community currencies in rural Africa

Around 2 000 community currencies are used around the world, providing an alternative way of exchanging goods and services when, for whatever reason, the national currency is unavailable. Will Ruddick, founder of Grassroots Economics and PhD researcher, is collecting survey data to evaluate the impact of these alternative currencies. UCT eResearch assisted with the storage of that data.

> eResearch is a service with a purpose, and that purpose is to support and accelerate research at UCT. It doesn’t matter who is conducting the research or how big it is. For some research projects, such as this one, a simple server can make all the difference.

“Across Africa, rural villages sometimes just run out of money,” says Ruddick. “So the villagers find themselves in a situation where they have goods and services to trade, but no medium of exchange.”

To address this, the NGO Grassroots Economics, founded by Ruddick in 2012, developed a paper voucher system that functions as a complementary or alternative currency. These vouchers are security-printed in the same way a national currency is, and can be exchanged for goods and services within a specific community network. The system enables those who have no money to enter the local economy, and also allows business owners to save the national currency for business upgrades or other strategic expenses.

This alternative currency has been implemented in six locations in Kenya and in two in South Africa.

In 2016 Ruddick decided to undertake a PhD to study this alternative currency system. Using economic survey data, he seeks to better understand how these vouchers are used and what impact they have on the social and economic welfare of those using them. Using the data, he also plans to develop a model to simulate, understand and predict the effects of community currencies over time, within a range of demographics.

To do this, Ruddick has been using the open-source software Open Data Kit, designed to collect, manage and use data in resource-constrained environments by means of mobile phones.

“Open Data Kit has a really cool Excel interface for creating surveys; it also allows you to input additional data on top of it, such as GPS data, or images from your mobile phone. Then you just click ‘send’ and it uploads to the server.”

The trouble, however, lay in storing the data. Previously, Ruddick used a commercial cloud service for data storage, but this became quite expensive, as he had to pay for even the smallest amounts of data. UCT eResearch set up a server to store the data at UCT, and because his needs are so small, this comes at no additional cost.

“eResearch is a service with a purpose, and that purpose is to support and accelerate research at UCT,” says Dr Dale Peters, UCT eResearch director. “It doesn’t matter who is conducting the research or how big it is. For some research projects, such as this one, a simple server can make all the difference.”

Needle in a haystack: Wrangling data to identify host biomarkers of TB progression

Up to 80% of the adult South African population is estimated to be infected with Mycobacterium tuberculosis (Mtb), the bacterium that causes tuberculosis (TB). However, about 90% of healthy individuals believed to be carriers of the bacteria will never get sick or infect anyone else. If researchers could pinpoint what causes TB to progress in some individuals and not in others, they could break the back of the TB epidemic.

The South African Tuberculosis Vaccine Initiative (SATVI) is working towards predicting an individual's likelihood of developing TB, but the size of the data involved is a challenge in itself. UCT eResearch has been working with SATVI from the point of data acquisition and will continue to do so until its eventual publication.

Researchers at SATVI, co-led by SATVI director Professor Mark Hatherill and immunology lead Professor Thomas Scriba, set out to identify what differentiates healthy individuals infected with Mtb who ultimately develop TB from those who remain healthy. A decade ago, SATVI recruited a large cohort of around 6 300 healthy adolescents, about half of whom were infected with Mtb. This cohort was followed up every six months for two years.

"While we recruited all the adolescents at the same point, healthy and without symptoms," says Dr Virginie Rozot, a postdoctoral researcher at SATVI, "at the end of the study, we had two clear groups: those who were susceptible and developed TB, and those who stayed healthy."

The researchers' goal was to identify differences in immune responses between the two groups, and key blood markers that could be used to distinguish which individuals would develop TB, so that they could be treated pre-emptively.

As part of this project, Rozot developed the first mass cytometry (CyTOF) platform in Africa. This technique combines two experimental platforms – flow cytometry and elemental mass spectrometry – to allow researchers to study more properties of individual blood cells than was previously possible.

Rozot and her colleagues rapidly ran into difficulties due to the amount of data they generated. Every day, the cytometer would analyse and produce data for a few dozen samples, each with a million-odd cells, resulting in a high number of combinations of markers. The resulting data files were massive and needed to be stored until the completion of the project – and beyond – for analysis. The challenge was that the controller computer attached to the cytometer could not store the data generated by the equipment.

Ashley Rustin, senior technical specialist at UCT eResearch, helped the group with their data requirements by ensuring that the data from their instruments, such as the CyTOF, was automatically backed up to the research data central storage repository.

> Systems biology and big data are key to unravelling the complex interplay between Mtb and the human host.

This ensured that the data was secure and that SATVI researchers could access the data from their computers over the network or from anywhere in the world.

"The data sets generated on the CyTOF are massive, and I soon realised that the network was a bottleneck," says Rustin. "I arranged for the network link between the controller computer and the building switch to be upgraded. This significantly improved the speed of the backup of the data sets to the research data repository located in the Upper Campus data centre."

Curating the data for open publishing

As the funder of this project, the Bill & Melinda Gates Foundation requires that the data sets be published in a reliable open-access repository. Thus far the team has identified UCT's institutional repository ZivaHub as a good option for sections of the data, while the NIH-supported ImmPort repository will house the remainder. The curation of the data is not completely straightforward, though, says Dr Mbandi Kimbung, a SATVI researcher who is working on preparing this data set so that it can eventually be published and shared with the public. Using one raw data set, the SATVI researchers are undertaking a number of different research projects. The outcome is a range of processed data sets, each looking at different aspects of the same blood samples.

"This has resulted in very rich sets of data, with a range of layers that can now be integrated," explains Kimbung. "It is an attempt to model the immune system, to see whether that can help us to understand the development of TB."

In addition to curating the data, Kimbung and the team also had to factor in the requirements of ethics committees around the use of samples from human participants in research. The data needed to be completely de-identified before it could be published, for instance.

"The question of who would control the data was important to the ethics committee," says Kimbung. ZivaHub was an attractive option for hosting the clinical database, as it assigns a UCT digital object identifier (DOI) to the data that establishes its ownership. "Also, with ZivaHub, we can edit the data during the lifecycle of the project and have complete control over when the data is made public."

Large gaps remain in our understanding of the complex interactions between the TB bacterium and the human immune system. Improving this is critical to developing better interventions that will halt TB transmission.

"Systems biology and big data are key to unravelling the complex interplay between Mtb and the human host," says Scriba. "The recent advances in technology and data science are already bearing fruit and I am very excited about the new biological insights and innovative medical interventions that are emerging."

UCT eResearch // ICTS-on-Main building // 7 Main Road // Mowbray // Cape Town // South Africa

{email} eresearch@uct.ac.za // {web} www.eresearch.uct.ac.za