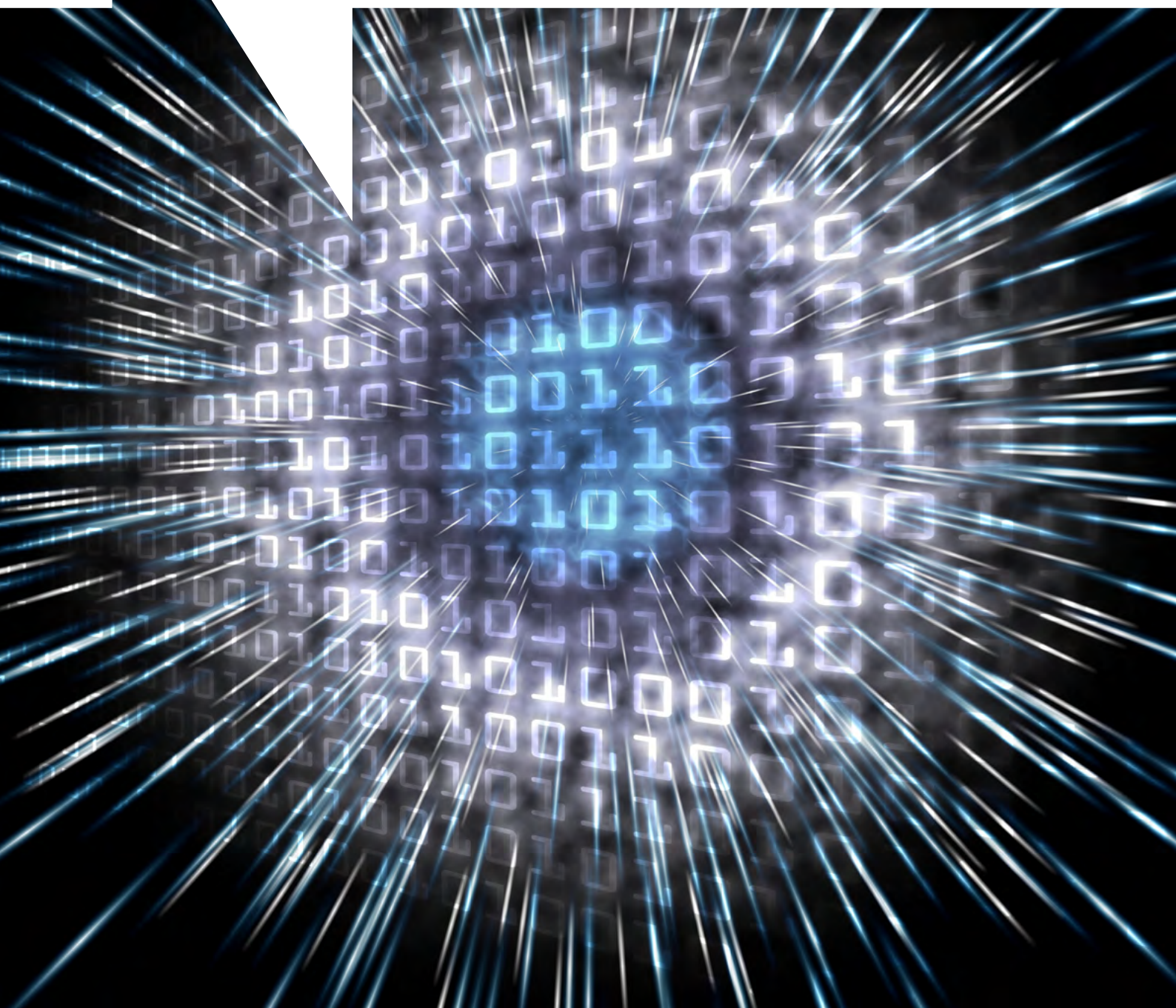
Science and Technology Facilities Council

Scientific Computing Department

Annual Review 2019



Contents

FOREWORD

Tom Griffin, Director SCD 05

SKILLED PEOPLE

People, skills and career development 07

Our graduates and apprentices 08

INSPIRING AND INVOLVING

Public engagement 10

BUILDING INFLUENCE AND FOSTERING COLLABORATION

International connections and activities 14

Collaborative Computational Projects – a success story spanning more than four decades 17

DAFNI: enabling and supporting decision-making 19

DEVELOPING ADVANCED TECHNOLOGIES

AI-driven cloud masking 22

Molecular digital design for the pharmaceutical industry 24

Ontologies for materials modelling marketplaces 26

DELIVERING SOLUTIONS FOR DATA-INTENSIVE SCIENCE

Data and software for 21st Century large-scale facilities science 28

March of the tape robots 30

From Europe to IRIS – key projects 32

SHARING KNOWLEDGE

Enabling open science 33

Towards an open science portal for STFC 34

CONTACTS 36

Scientific computing is fundamental to modern research. This broad and rapidly-advancing field involves exploiting advanced computing capabilities to understand and solve complex problems in science.

STFC's Scientific Computing Department is one of the UK's leading centres of expertise in data-intensive science, and home to sophisticated high-performance hardware. Our people have cutting-edge skills and expertise in scientific software research and development, and world-leading capabilities in 'big data' storage and analysis, visualisation and simulation, and scientific information management. We support some of the UK's most advanced scientific facilities and provide the tools that enable the scientific community to discover and deliver vital research.

Foreword

Welcome to the 2019 annual review of the STFC Scientific Computing Department (SCD).

I never fail to be gratified and impressed by the variety of work the department is involved with, nor by the wide span of skills, expertise and dedication of our staff. They provide crucial support for the large-scale research facilities in the UK and internationally, and for scientific research globally. Some of this is demonstrated in the following pages, although this is by no means a comprehensive review of everything we do.

Inspiring and encouraging the next generation of scientists is one of our key goals. We employ a number of apprentices and graduates each year and we provide the training and skills they need to develop their careers. We are always delighted when, after completing their training, they decide to stay with us as permanent members of staff, which many of them do. Our public engagement activities continue to grow and engage with hundreds of young people and the wider public every year.

We continue to work with research communities around the world to support global scientific advancement. Our international reach extends widely across Europe, the United States and many other regions and we welcome the opportunity to collaborate and share knowledge.

Growth areas for computational science include Artificial Intelligence and Machine Learning. Our Science for Machine Learning Group (SciML) is working closely with the Alan Turing Institute and in 2019 we announced a collaboration to host and manage the 'PEARL' Machine Learning Service, an advanced system that enables training for complex AI models on very large datasets. And, as you will note from this review, the SciML group is also doing some interesting work around the UK's favourite topic of conversation – the weather!

Although this is the review of 2019, I cannot ignore the COVID-19 situation that is affecting us all. Like so many, SCD has done what it can to support research in this area, from the long running Collaborative Computational Projects in structural biology (CCP4) and electron cryo-microscopy (CCP-EM), to COVID data sharing through the Research Data Alliance (RDA), to providing flexible compute on our cloud and HPC to support research at Diamond and elsewhere. This year has also seen several new initiatives start, and others receive renewed funding. I hope you enjoy reading about just some of our activities in 2019. There is a lot to look forward to in 2020 and beyond.

Tom Griffin
Director

Skilled people

The Scientific Computing Department (SCD) is a strong supporter of skills development through training and mentoring of staff and others. In addition to training and managing sandwich students, graduates and apprentices, SCD staff have organised or led many specialised and technical courses, both in-house and with academic and industry partners.

SCD is privileged to have a very diverse range of talented and dedicated computational scientists and support staff who are passionate about the work they do and who also make time to mentor and encourage others.

Here we feature brief profiles of just a few of our established science staff and some of the graduates and apprentices who have joined SCD to advance their careers.

People, skills and career development

Staff profiles

Valeria Losasso studied for a degree in biotechnology in Modena, "a beautiful, not too big town in Northern Italy. I studied for my BSc between 2000 and 2003. The course structure included the choice of a specialised path starting from the second semester of the second year. I chose Pharmaceutical Biotechnology." She went on to gain two PhDs before joining STFC more than seven years ago.

She is now a Senior Computational Scientist and has, until recently, been working on antimicrobial peptides (used in antibiotics and other medicines). She has just started a project on cryoEM (cryo-electron microscopy) maps of proteins and describes her work as "using powerful computers to save costs and time while searching for new drugs"

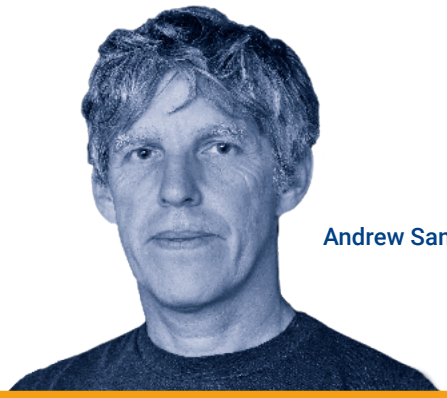
She is also supervising a PhD student on the modelling part of his project in collaboration with Unilever, for whom she has done several projects on drug discovery.

"The best part of my role is definitely interacting with people from different backgrounds. Finding a common language is challenging in a positive way. Also, having the chance to be constantly learning new things."

Andrew Sansum has been at the Rutherford Appleton Lab since 1986, initially working as a physicist and then moving in 1990 to what was then the Central Computing Department. He has worked in a number of roles over the years, including building the OPAL detector for the Large Electron-Positron, work he describes as "awesome!" and which saw him spending six months each year at CERN during the build.

His role today is just as challenging, although in a different way as he is currently Head of SCD's Systems Division and Technical Director of the IRIS collaboration, a £16M project which entails developing and integrating a national eInfrastructure for STFC.

Valeria Losasso



Andrew Sansum

The best part of his current role? "Supporting and developing staff, watching them develop new skills and gain confidence and independence. It's a fantastic feeling when I see that they are already several steps ahead of me in their analysis of a problem, with their own ideas and plans and I'm left with no more to contribute."

Alin Marin Elena is a computational scientist based at the Daresbury Laboratory and says he loves his job. He was inspired by his teachers and his own stubborn curiosity to follow a career in science, studying computer science and physics in Romania and Ireland, and graduating with a PhD in Physics in 2013.

Alin has worked in SCD for the past five years and has his expert fingers in many pies, from quantum computing projects or collaborations with industry and academia, to teaching and training others in the use of simulation software developed at the Daresbury Lab.

He is also passionate about encouraging and inspiring young people to become our future scientists. He finds public engagement activities particularly rewarding and loves to "meet kids who are interested in science, and to be able to 'myth-bust' about what scientists do".

Asked to describe his work in simple terms he said, "Trying to understand how atoms and molecules behave, but using a computer so there's no danger, pollution or tidying up to do."

Tyrone Rees has always loved science. He recalls his grandparents giving him a chemistry set at the age of 6 or 7 and setting up a 'science lab' in his parents' garage. At university he discovered the world of pure maths and really enjoyed it. Today he leads SCD's Computational Mathematics Group.

"We are well placed in STFC to produce maths that is not just elegant, but is also practical. Working so close to the scientists that make use of numerical algorithms every day means we get more chance than most to both see where the numerical bottlenecks are in current algorithms, and to actually get people to use the maths we develop."

Tyrone uses maths to write computer programs that help give results that are either better than before, or get the same answer quicker than before.

"I have learned a lot about many things in my time at STFC, from software engineering to project management. Possibly the most useful skill has been the ability to be presented with something new - be it a scientific technique, a mathematical result, or a software framework – and quickly be able to make sense of it enough to extend it or use it."

Leandro Liborio grew up in Argentina and was the first in his family to finish secondary school and go to university. He studied for a chemical engineering degree but later, as he started to develop an interest in the science behind the engineering processes, he switched to physics. He says, "When you do physics you make a million approximations. If you know what you did, you can go back at any point in the road, make a different approximation and you end up with a different formula that might work."

Today Leandro is a member of SCD's Theoretical and Computational Physics Group, working closely with the Muons Group at STFC's ISIS Neutron and Muon Source. He builds mathematical models to simulate their experiments, and then helps the scientists compare the data from the real experiment to his model data. This helps them interpret what the experiment is showing.

One of the things he likes best about working at STFC is the physical proximity of the different departments as he is only a short walk from his collaborators. "I can talk to people face-to-face. It's easier, much more dynamic." And, he says, this also means he can interact with the experiments he is modelling in a much more effective way.

Alin Marin Elena

Tyrone Rees

Leandro Liborio

Our graduates and apprentices

Tom Dack has been interested in science and technology for as long as he can remember and was encouraged to study STEM subjects (science, technology, engineering, maths) throughout his school years. Tom found that he particularly enjoyed using computers and programming to enable his research so, after gaining his Physics degree, he made the decision to change direction and completed an MSc in Computer Science.

His first encounter with STFC was when his college visited the Rutherford Appleton Laboratory for a Particle Physics Masterclass. Later, Tom discovered that STFC runs a graduate scheme and says it struck him then, that "STFC was not only a cool place to work, but the perfect place to pursue a career in computing within a research environment."

When he finished at university he became a member of the STFC graduate scheme, working in a number of different groups within the Scientific Computing Department to gain the knowledge and skills required for his chosen career.

Tom now has a permanent staff role as a systems administrator in SCD's Distributed Computing Infrastructure group. Here he is developing and managing an Identity and Access Management (IAM) service for use by IRIS, a nationwide collaboration which coordinates e-infrastructure needs and knowledge exchange to enable the exploitation of data outputs from high-energy physics communities.

45 GRADUATE TRAINEES
recruited since 2004

This year we had FIVE APPRENTICES
working in and around SCD

Put simply, Tom says he is "setting up a computer system to make sure the right people have access to the right things!"

Becky Fair joined STFC on the graduate scheme after completing her degree in astrophysics at University of Liverpool and now works in SCD's Theoretical and Computational Physics group.

During her training, Becky worked on a diverse range of projects including user interface development, database monitoring and fluid dynamics. Her work took her to a conference in Spain where she gained experience in public speaking. She has recently published a scientific paper based on her graduate placement work on fluid dynamics, which has applications in naval architecture and offshore oil rig design, amongst others.

Now Becky works in a permanent role on the PACE project, working with the Excitations group at the ISIS Neutron & Muon Source. Her project aims to lower the technological barriers for users to analyse their data. "Before I started this role I knew very little about the experimental method (inelastic neutron scattering) and the thing being observed (phonons) so it's been a steep learning curve," she says. "I feel like my intuition is much better now but I'm still learning new things all the time."

Matthew Richards is a second year apprentice, based in SCD at the Rutherford Appleton Laboratory


Tom Dack


Matthew Richards


Becky Fair

in Oxfordshire. He says that opting out of the traditional university route has allowed him to thrive. "It allows me to study for a degree in the subject while gaining on-the-job experience. I can appreciate how different classroom learning is to what goes on in industry."

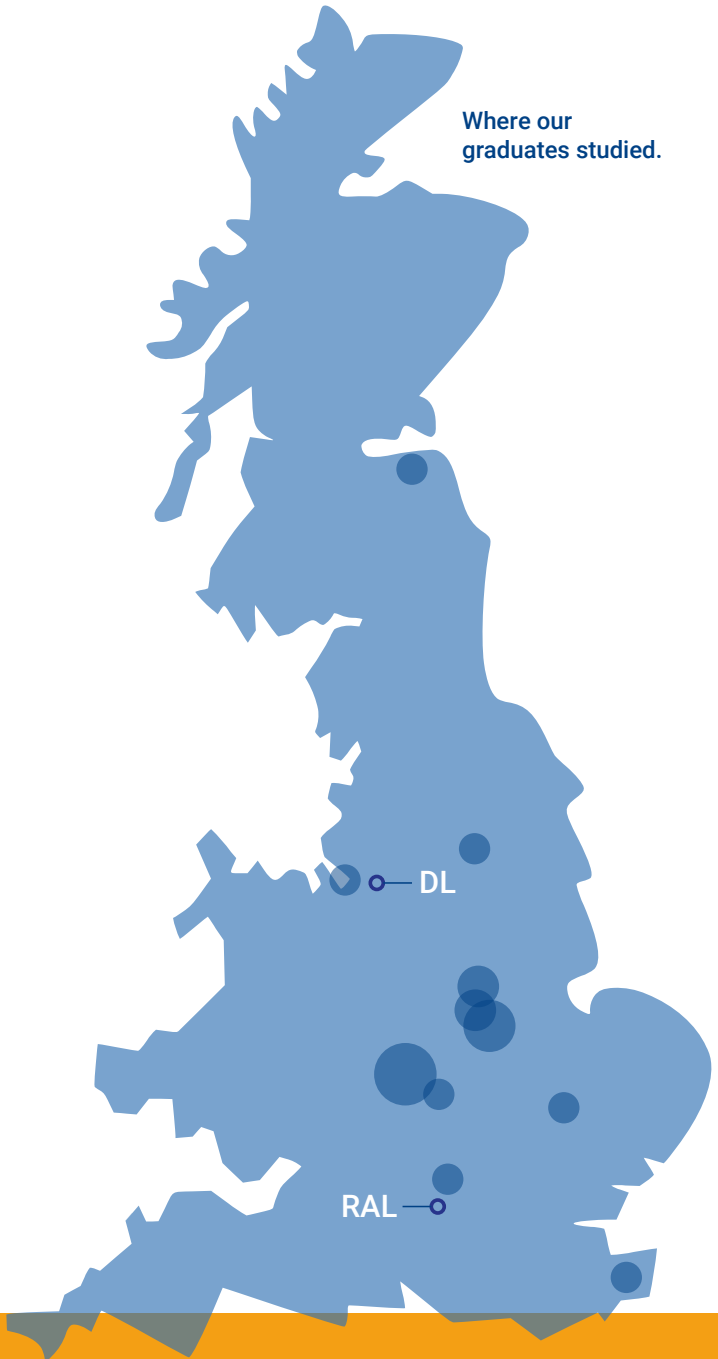
8/10

graduates who completed the scheme in the past 10 years ARE STILL EMPLOYED BY SCD

As a computing apprentice he rotates roles every six months, trying out new departments and building his skillset. "I enjoy the wide variety of work," he says. "There's always things I'm more interested in doing due to gaining experience and confidence levels with the technologies used, but it's good to be exposed to new things." He spent one rotation working with high performance servers in the SCD 'batch farm', and another developing and maintaining a tool to update operating systems used on STFC's cloud service. More recently he has worked on testing tiles of the Large Pixel Detector (LPD), a large x-ray laser which can image ultrafast chemical reactions in exceptional detail.

Matthew appreciates the opportunities he has to keep learning and to explore areas that he didn't know would interest him. "I've discovered you can work in this industry without initially thinking you're interested in science but having worked here, the science that goes on throughout the campus is really interesting and diverse. Seeing the amount of work that goes into making a scientific facility such as ISIS is amazing. There's so many aspects to the facility and an incredible amount of skilled engineers have built it."

Where our graduates studied.



Degrees studied before joining SCD.

9% Other including Maths and Computational Biology

39% MSc/MPhys/MEng Physics

4% BSc Physics

13% MSc/MEng Computer Science

35% BSc/BA Computer Science

Authors:
Marion O'Sullivan and Evelyn Greeves, SCD Impact Group

Public engagement

SCD has a very active Public Engagement Committee which met regularly during 2019 to plan and develop activities to enthuse and inspire children, teachers, families and members of the public. The committee has made key contributions to the National Laboratories Public Engagement Programme.

In addition to a week of public engagement events to celebrate the anniversary of the Apollo Moon Landings, which saw 12 staff members delivering over 280 hours' worth of engagement with students and teachers, we have also reached over 4000 people over the course of the year at smaller events. New resources and workshops have also been developed, and are now a core part of the public engagement programme.

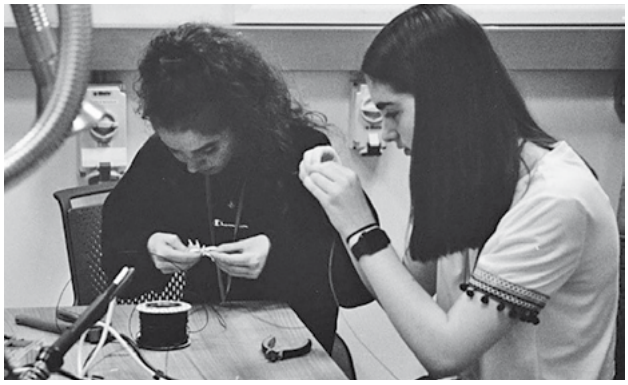
Inspiring the next generation

This year, SCD teamed-up with the Technology department to run a week of work experience at the Daresbury Laboratory (DL) for 16 Year 10 students. This exciting new project was borne out of a collaboration between our staff in Scientific Computing, but soon brought in support from a range of people on site.

The brief given to the students was to "control the temperature of a water bath" in order to be able to develop photographic film. This innovative project was developed specifically with students in mind who were interested in STEM subjects, but who might not know exactly what sort of work they would like to pursue. As well as working on their main project the students listened to talks given by Scientific Computing staff from the engineering, chemistry, physics and the scientific machine learning groups.

The students found the week rewarding, with one commenting that "from talking to one of the data analysis scientists it has allowed me to see a new interesting area of programming to look into."

Thanks to all DL staff who happily volunteered their time to ensure the work experience was as rich an experience as possible, with a special thanks extended to Dawn Geatches and Tim Franks.



Two images taken by students and developed as part of their project.

284

the number of people we showed around

RAL SCIENTIFIC DATA CENTRE

1 in 7 of our staff mentored a work experience student

Exchanging best practice

The International Conference on Computing in High Energy and Nuclear Physics (CHEP) addresses the computing issues for the world's leading data-intensive global science experiments. The Conference is a major event in the field, featuring plenary sessions, parallel sections and poster presentations as well as publishing peer-reviewed proceedings.

This year, CHEP started with a public event "Universal Science", allowing department staff to engage the public with our work in the areas of Particle Physics and High Performance Distributed Computing.

After this, the conference dedicated one of nine tracks to Collaboration, Education, Training and Outreach. Greg Corbett presented how the department has worked to support a culture of Public Engagement (PE) and highlighted the STFC PE Strategy and Evaluation Framework as examples for use by other science institutions. Within the track, many ways of engaging the public were presented, some of which we will integrate into our PE programme in the coming year, such as the "Unplugged Computing for Children" activities developed at CERN.

We engaged with
4,000
members of the public via
SCD led events and activities

CERN

The CHEP public day in full swing.

Summer coding with Boulby Underground Laboratory

In August, the Rutherford Appleton Laboratory (RAL) welcomed 20 young coders to the lab for a week of coding as part of "Summer Coding". Summer coding is an annual event where, over the course of a week, children aged between 7 and 15 complete a series of computing based challenges, culminating in a demonstration of their work in a final presentation to peers, friends, family, and staff.

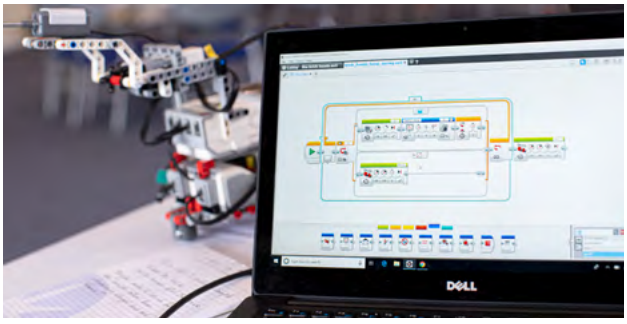
This year the children were tasked with writing code to control a LEGO Mindstorms "rover", which would run remotely on a rover located remotely at STFC's Boulby Underground Laboratory. The Boulby laboratory is situated 1.1km underground sharing its tunnels with an active mine and conducts research into topics that require extremely low background radiation, such as Dark Matter.

28 Number of work experience students mentored by our staff

Over the course of the week, with the assistance of SCD staff members as mentors, groups programmed their Mindstorms to complete "landing surveys", undertake "geological surveys" and take temperature readings. These programs were tested above ground in RAL's visitor centre, before being tested and then ran on site at Boulby. Groups also analysed the data obtained from their robots' sensor readings by drawing graphs and performing simple calculations.

Lessons learnt from the event are being used in the planning of Remote[3] (Remote sensing by Remote schools in Remote environments), a new Spark award funded project to engage children from schools in remote Scotland. This project is being organised by the STFC PE teams, the Boulby Underground Laboratory and the University of Edinburgh. SCD staff are continuing to support and facilitate this activity with Will Furnell and Laura Murgatroyd acting as mentors for two of the schools, and Tom Dack participating in the organisation and running of the project as a whole.

We delivered **58** hours OF QUALITY **PUBLIC** ENGAGEMENT



Author:
Greg Corbett, Distributed Computing Infrastructure Group

The schools involved in Remote[3].

Building influence and fostering collaboration

International connections and activities

SCD is committed to working internationally with researchers and computational experts to support global scientific advancement. During 2019 our staff visited more than 32 countries around the world, visited colleagues and collaborators in other science institutes, and also participated in many international meetings via video-conferencing. Although much of our work has been focused in Europe and the USA, we have also been active in countries from Russia in the north to Australia in the south.

Our staff were present at many conferences and workshops, some of which we had co-organised with an international partner, at others we were teaching or mentoring, and some we were attending to share and gain knowledge, keep up with new practices and forge new collaborations.

We also held the UK's annual HPC and Big Data Conference, *Computing Insight UK (CIUK) 2019*. This conference is a real home-grown event as it is organised and produced by SCD and held in Manchester each December. It is fast becoming the UK's premier 'supercomputing' conference and attracts speakers and delegates from around the world. This conference, along with a small selection of our other activities, are reported here.

Map produced by Peter Oliver, SCD

UK

Although CIUK has been running for only a few years in its present form, it started life almost 30 years ago as a very popular Machine Engineering Workshop where delegates could find out all about the latest computing technologies and learn how to use them. Delegates at CIUK can still find out about the latest technologies and share knowledge and experiences at the technical talks and vendor workshops, the research and industry exhibits, and through the many networking opportunities the conference now provides.

For the first time we included a 'Research Zone' in the exhibition hall where academic and research centres could demonstrate the outcomes and potential of research projects that use HPC hardware and storage systems. This delighted some of our delegates who were keen to find out more about facilities like the national supercomputing service ARCHER, the JASMIN 'super-data computer' and DiRAC (Distributed Research Utilizing Advanced Computing), all of which had expert staff on hand to explain the relationship between researchers and the machines they rely on for their work.

Another 'first' for CIUK was a talk delivered remotely by one of our speakers who had been injured in an accident and was unable to travel. The absence of an actual person on the stage was of no consequence to the audience who packed into the conference room for SCD's Jens Jensen's presentation about the development of an app which includes cloud-based Machine Learning, to assist travellers at airports.

And we heard a talk from Demi Pink, a PhD student at Kings College London, who was the winner of our first 'Jacky Pallas Memorial Award' competition. Jacky was an active member of the CIUK Science Advisory Committee and this award will be a regular feature at CIUK conferences to commemorate her research work and her passion for encouraging young people in science.

CIUK 2019 was our most successful conference to date, attracting around 400 delegates and over 50 exhibitors. The conference theme was 'Computing the Future' with a sub-theme for each of the two days. Day one focused on computing today, with talks and presentations on Cloud Computing, Software Development, and Data Science – the convergence of AI and HPC. Day two looked to the future with sessions including Hardware Towards Exascale, Software Development Towards Exascale and Quantum Computing.

Europe

We saw the successful completion of the European Open Science Cloud (EOSC) pilot project led by Dr Juan Bicarregui, our Head of Data, who was subsequently elected to the executive board for the EOSC project going forward. EOSC aims to offer researchers and businesses across Europe a virtual environment to store, manage and analyse data, making science more open and more productive. This was highlighted at the World Economic Forum in Davos by European Commission President Ursula von der Leyan, who stressed that developing the EOSC will ultimately lead researchers to new insights, new findings and new solutions, and that Europe is the first in the world to do that.

We celebrated, together with colleagues across Europe, the 50th anniversary of CECAM (Centre Européen de Calcul Atomique et Moléculaire). Originally formed to promote advanced computational research for science and technology, there are now 17 CECAM nodes across Europe.

We host the UK node at our Daresbury Laboratory in Cheshire. SCD's Leon Petit is on the Board of Directors for CECAM. He said: "Computational materials science delivers fundamental components for many of the things we rely on every day, such as electronic semiconductors, magnetic materials for refrigeration, and nuclear fuels. At STFC our research focuses on using computer simulations to develop and test these materials, and we also train and support researchers working in these areas."

The Daresbury CECAM node's workshop and training program leverages the support that SCD provides to the UK scientific community through CoSeC and the Collaborative Computational Projects, at the same time accentuating its international dimension.

We also marked the finish of the three and a half year EU Horizon 2020 NLAFET Project, developing linear algebra software for extreme scale computers. A final review meeting was held in Luxembourg in the summer of 2019.

Members of our department were invited speakers at many European conferences, such as the Cephalocon conference in Barcelona, Spain, where we discussed the challenges of managing a rapidly growing Ceph cluster; the International Metadata and Semantics Research Conference in Rome, Italy; and the Applied Mathematics ICIAM 2019 meeting in Valencia, Spain.

SCD and the Hartree Centre took a joint exhibition booth at ISC2019, Europe's largest Supercomputing Conference, which is held each summer in Frankfurt. Our staff presented a research poster about the JASMIN super-data-cluster, both for the Conference poster session and the Women in HPC session; proposed and organised a session on *Software Engineering and Reuse in Computational Science and Engineering* and attended many of the technical sessions and workshops. These are invaluable for keeping up to date with the latest technology and practices, exchanging knowledge and developing new collaborations.

USA

At SC19, the world's biggest supercomputing conference held in the US in November, we welcomed a huge number of people to the UKRI-STFC booth. SCD has been an active participant and exhibitor at this conference for more than 20 years and SC19 was no exception. Together with our Hartree Centre and other STFC colleagues we had a very strong presence, both on the exhibition floor and in the technical programme. At the STFC booth we were pleased to present two interactive models - an impressive Ceph cluster, purpose-built by SCD staff to demonstrate how we use erasure coding to protect the vast quantities of data coming from the Large Hadron Collider at CERN; and a Raspberry Pi cluster from the Hartree Centre to demonstrate how a supercomputer communicates across networks to execute parallel programs. The mini-Ceph cluster, affectionately named 'Wolfi' after the world's smallest cephalophile octopus, was particularly popular as its 12 nodes, each a Raspberry Pi mini- computer with colourful flashing lights and solid state drive, encouraged people to come and ask about it.

Author: Marion O'Sullivan, SCD

Collaborative computational projects

A success story spanning more than four decades

A reflective account from Professor Paul Durham, former director of STFC's Computational Science and Engineering Department – now known as the Scientific Computing Department (SCD) – and former President of the CECAM Council (Centre Européen de Calcul Atomique et Moléculaire). The UK node of CECAM is currently hosted by SCD at the Daresbury Laboratory. Paul continues to serve as a member of CECAM's Council.

Over 45 years ago, one of the key programmes in SCD was born – the Collaborative Computational Projects (CCPs). This note gives a brief account of how this happened and who brought it about.

By 1982 there were nine CCPs covering the wide range of fields indicated in the following table.

Project	Title
CCP1	Electron Correlation in Molecular Wavefunctions (1974)
CCP2	Continuum States of Atoms and Molecules (1978)
CCP3	Surface Science (1979)
CCP4	Protein Crystallography (1979)
CCP5	Molecular Dynamics and Monte Carlo Simulation of Macroscopic Systems (1979)
CCP6	Heavy Particle Dynamics (1979)
CCP7	Analysis of Astronomical Spectra (1980)
CCP8	Nuclear Structure Physics (1980)
CCP9	Electronic Structure of Solids (1981)

The programme of CCPs was initially established by the Science Research Council (SRC), the UK government's main funding agency for scientific research at the time. As well as funding university researchers by means of grants, the SRC also operated two large central laboratories: The Rutherford Appleton Laboratory (RAL) and Daresbury Laboratory (DL). The SRC described the aims of the programme as follows.

The major aim of these projects is to bring together scientists from several different universities and research groups to:-

- provide for the rapid interchange of information on theory, algorithms and computer codes;
- collect, maintain and develop relevant items of software;
- encourage basic research in the given areas by providing facilities for rapid computer implementation of new methods and techniques;
- assess and advise on associated computational needs;
- disseminate information among University and other research groups by organising "symposia" or "workshops".

In hindsight, it seems to me that the SRC showed vision, ambition and generosity in accepting and implementing its academic advice to proceed with the project. That academic advice was strongly driven by the quantum chemistry and atomic and molecular physics communities. Many of the essential elements later to form the basis of SRC's portfolio of CCPs, as we describe below, were put in place from the outset, including the central role of Daresbury Laboratory and its staff and management. Indeed, over time, the aims of the CCP programme have remained in essence unchanged, and although the practical ways in which the projects address these aims have certainly evolved to match circumstances, they can be crystallised into a "standard model" reading something like this:

The CCPs bring together all the major UK research groups in a given field to pool their ideas and resources to tackle the development, maintenance, distribution and user-support of large scale scientific software. This is done by implementing flagship code development projects and by means of networking activities: curating libraries of code; organising training in the use of codes; holding meetings, workshops etc; inviting overseas researchers for lecture tours and collaborative visits; and issuing regular newsletters.

Underlying all this was the realisation that the writing of a world-competitive scientific code was becoming a task too large for any research group to handle on its own. After all, the post-docs who are, in practice, the workhorses of every research group have a limited working lifetime (usually three years in one role). In computational research, this often meant that it was too much for even the most brilliant and effective post-doc to write a serious code, do some recognised research with it, find their next job and leave the code in a fit state to be taken up by a successor. The intellectual investment required to produce useful codes simply had to be protected beyond the canonical 3-year period. The collaborative, collective ethos exemplified by the CCP model was thus driven to a great degree by the increasing demands of remaining competitive in computational research. A cottage industry had to be replaced by a more coordinated collective effort. Of course, it is quite pleasant to work on one's own in a quiet cottage, and it is true that the CCP model, with its slight whiff of centralised planning, did not appeal to everyone. Nevertheless, the CCP model seems to have had, from its outset, a striking robustness, able to accommodate, for the most part, the internal tensions among participants that must inevitably occur from time to time.

According to Phil Burke[1], the idea of the first project, CCP1 – a proposed quantum chemistry project, was born in 1973 during a Working Party at RAL. The Science Board of SRC considered the proposal in October 1973 and gave it a green light. This first (pilot) project was a great success and in 1976 the steering panel approved its continuation. The following year it was decided to co-locate CCP1 with the next project within the newly-formed Theory and Computational Science Division at Daresbury, with Phil Burke as Division Head, and John Pendry[2] as head of the theory group. John went on to establish a number of the early CCPs, especially those with a solid state physics theme.

As Phil Burke writes, this move "set the scene for a rapid increase in the CCP programme". It also established the key role of Daresbury as the cornerstone of the CCP programme.

1 Professor P G Burke FRS was a leading theoretician of atomic and molecular physics, with a chair at Queen's University Belfast. He made seminal contributions to the R-matrix theory of atomic collision processes. He was the founding director of the Theory and Computational Science Division at Daresbury Laboratory in 1977, and was the essential visionary of the CCP concept and its early implementations in several fields. With his quiet, thoughtful manner, he was extremely effective in establishing the basis for UK computational science and supercomputing strategy. Sadly, Phil died in the summer of 2019. He is greatly missed.

2 Professor Sir John Pendry FRS joined the new TCS Division in 1977 as Head of the Theory Group at Daresbury. At Cambridge and then Bell Labs, he had pioneered the theory of LEED and EXAFS. When he came to Daresbury he drove theoretical support for the SRS experimental programme, making important contributions to the interpretation of ARPES and XANES spectra. He also established a number of the early CCPs, especially those with a solid state physics theme. I can testify to the stimulating atmosphere he brought to the Division in those days. In 1981 he left to take up a chair at Imperial College where he has had a glittering career in physics, notably in developing conformal optics using photonic materials.

What was the scientific context into which the CCPs were born? Perhaps we might identify the 1970s as the beginning of "computational science", at least in the UK, as a respectable scientific activity, with its own intellectual and technical status. Nowadays, of course, it is obvious that computational methods are utterly indispensable to all fields of scientific research, both theoretical and experimental. Nevertheless, in the 1970s there was a feeling in some theory departments (again, at least in the UK) that there was something not quite respectable about computational work. "Brute force" numerical calculations was a phrase that one heard fairly regularly in physics departments, with the implication that gentlemen theorists did analytical work. But this kind of attitude was by no means universal and things were changing. In chemistry departments, quantum chemists were establishing a strong presence and doing hitherto impossible research. In the USA, the value of computational methods was already appreciated. By 1974, the Gordon Conference on the Liquid State had 5 or 6 computationally based papers. Thus, while there was some resistance, it was clear that computational science was moving centre stage.

It seems reasonable to suggest that the foundation of the CCP programme was probably motivated by a happy combination of scientific altruism (the collaborative spirit) and self-interest (showing value for money, in order to get more money). Perhaps this is true of most programmes that turn out to be successful over a long period. Over the lifetime of the Programme as a whole, computing technologies of all sorts have been completely revolutionised, funding organisations - both national and international – have come and gone, and individual CCPs have come and gone too. And yet, the essential CCP concept has proved to have a remarkable and perhaps surprising longevity.

What does the overall CCP Programme look like now? The overall scale of the CCP programme has remained essentially the same for 40 years, with defunct projects being replaced by new ones. Often new CCPs evolve old ones as their key ideas and techniques become used more widely. Perhaps the most enduring and defining characteristic of the CCPs is the notion of a community. What is a research community, as distinct from a disconnected collection of individual researchers or groups in a given field? In my opinion, a research community is one in which the members know each other personally, collaborate with each other when appropriate on specific research projects, jointly nurture the careers of young researchers within the community and self-organise to act together to promote and advance their field. No doubt there are other important characteristics, but, broadly speaking, this kind of community exhibits a "family feeling", a sense of belonging and mutual support. Most CCPs have put quite a lot of effort into developing and sustaining this community ethos, and this must go a long way to explaining the longevity of the concept.

As regards the future, Professor Dominic Tildesley who, among his many achievements has chaired CCP5, authored "A Strategic Vision for UK e-Infrastructure" for the UK government and been President of the Royal Society of Chemistry, comments: "I think the CCPs are in a strong position to influence the industrial uptake of simulation and modelling. This requires support in the provision and use of leading-edge machines (as exemplified in the Hartree Centre at Daresbury). Give industry open and free access to the software library and expect them to make contributions of code and examples to the community. Most importantly, make sure that the CCPs continue to offer consultancy and direction on the use of modelling in industry through special events, invitations to CCP conferences, and confidential evaluations of company modelling strategies. I believe that an important future role for the CCPs will be as the custodian of a number of honest and inspiring case studies of the use of molecular simulation in pharmaceuticals, fast moving consumer goods, chemicals and the aero-space and automotive industries."

Author: Paul Durham

DAFNI : enabling and supporting decision-making

In this second year DAFNI (Data and Analytics Facility for National Infrastructure) has made good progress with the release of a pre-production Bronze version platform. Developed using Agile processes and following an extensive period of pilot implementation, June 2019 saw the launch of the first version of DAFNI at an event at The Royal Society in London. Over 200 representatives across Academia, Government and Industry attended to see, first-hand, the platform at this early stage of development, witness a live demonstration of each of the components on DAFNI and understand the current system developments and how its key capabilities can support greater research collaboration across infrastructure sectors.

DAFNI aims to transform the use of data, modelling and simulation in infrastructure research and decision-making by providing a new centralised hub for infrastructure data and unique computational capabilities. The focus is upon enabling collaboration by sharing data, enabling the coupling of system simulation models and informing decision making through advanced visualisation. It provides a central point for data and compute to carry out large scale analysis, specialised software tooling to support single and multi-component models, novel ways of visualising outputs and aims to support wider factors e.g. resilience planning, air quality impacts, and environmental well-being.

"Today, unprecedented amounts of data are at our fingertips in an instant. New technologies such as artificial intelligence and machine learning offer the potential for the UK's existing infrastructure to become smarter and work as an optimised system. DAFNI provides a strong platform to help us to do this. It gives us the opportunity to ensure that the recommendations we make in the next UK National Infrastructure Assessment will be based on the best data and robust modelling." Sir John Armitt, Chair of the National Infrastructure Commission.

A key example of how DAFNI is supporting decision making is seen in the pilot implementation of the University of Southampton's 'Railway Station Demand Model'[1].

According to Network Rail, 1.7 billion people per year travel by rail in the UK, and the number of passengers is rising by 6% each year. With ever-increasing demand for the rail network to be as efficient and wide-reaching as possible, it is important that any improvements or additions to the network are thoroughly thought out before spending the limited budget available.

One of the key questions facing Network Rail and local authorities alike is where new stations should be located in order to best serve both business and community needs. As the impact on the local environment, budget and connected infrastructure during and following the redevelopment process

1 Full-technical document available at: DAFNI Pilot 3: DAFNI on Track with Railway Station Demand Planning.

is enormous, the answers to this question in turn need to be supported by evidence-based predictions formed from in-depth analysis of projected future station use.

To try to answer this question, the Transportation Research Group at The University of Southampton has developed the Station Demand Model. This model generates a demand forecast (predicted trips per year) for one or more proposed local railway stations. It can also produce an analysis of the potential number of passengers who would change to a new station(s) and what net impact a new station would have on rail use. The model is flexible in that it can perform forecasts for multiple stations at once. These can be treated either independently (involving the assessment of alternative station locations) or concurrently (whereby the proposed new station will coexist with ones currently in operation).

The model, developed by Dr Marcus Young and Dr Simon Blainey, goes beyond the limits of existing models to better represent real-life travel behaviour, thus making the resulting predictions more accurate. One of the key benefits of this model being on DAFNI is a web interface has been developed which allows users to interact with the model for their specific requirements. It also allows for map-based visualisation of the results for the user to gain a better understanding of the impact to the surrounding area.

Running the model on DAFNI means that no specialist technical knowledge is needed to be able to use it, so it can be accessed by researchers and professionals in the transport industry alike. In a fast-paced world, one of the most sought-after advantages of any technology is of course how quickly it can help you achieve your goals. DAFNI has been able to reduce the model run-time considerably, enabling more efficient decision making for some of this country's most important infrastructure systems. Hosting this model on the DAFNI platform will enable it to be linked with other models as the platform grows, thereby improving the efficiency of nationwide infrastructure research on a more holistic level.
"We have now managed to automate our model in DAFNI so it can run based on a set of initial inputs with no further user intervention. These inputs are entered using an easy to use visual interface in DAFNI, meaning that non-expert users can run the model. Whereas previously it would take around one day to model one station; we can now run a scenario in less than an hour for multiple stations," said Dr Blainey.

2019 has also seen other key highlights for DAFNI as follows:

- Pilot integration of the following models on the facility:
  - Digital Communications Models Mobile 5G and Fixed Broadband Network model, Dr Ed Oughton – Environmental Change Institute, University of Oxford;
  - Housing market model developed for the Bank of England, Dr Adrián Carro – Postdoctoral Research Assistant, University of Oxford;
  - Station Demand Model, Marcus Young, Senior Research Assistant in GIS and Transport Engineering Data and Dr Simon Blainey, Associate Professor in Transportation, both at the University of Southampton;
  - Synthetic Population Estimation and Scenario Projection model, Dr Nik Lomax, Associate Professor in Data Analytics for Population Research, and Andrew Smith, Research Fellow, both at the University of Leeds
- A series of DAFNI roadshow events where DAFNI is demonstrated in Bristol, Cambridge, Cranfield, UCL, Heriot-Watt, Edinburgh, and at the Geovation Centre in London.
- DAFNI is included in Chancellor's Philip Hammond's letter to the National Infrastructure Commission as the facility of choice for this year's study of the resilience of UK national infrastructure assessment.
- A MoU was signed with Oxfordshire County Council to provide DAFNI for data storage and modelling advice.
- DAFNI has started to support a number of Centres for Doctoral Training in the area of water modelling and cyber security with researchers from Bristol, Cranfield, Newcastle, and Sheffield.
- January 2019 saw the first step of working with UKCRIC's Newcastle Urban Observatory in Real Time Flood Modelling to provide requirements of working with real-time data.
- DAFNI forms part of the winning team in the Infrastructure Projects Authority's Hackathon event to investigate how modelling can help with key decision making.
- DAFNI is presented to government officials visiting from Singapore Housing development as a facility which can support modelling, data curation and visualisation to aid decision making.

Author: Marion Samler, DAFNI Group

Developing advanced technologies

AI-Driven cloud masking

Weather is an integral part of human life, to an extent that almost every aspect of our life, from holidays to agricultural planning to navigation are underpinned by weather. Sea and land surface temperatures (or SST/LST) are two of the many parameters that can significantly influence the Earth's weather. For instance, large variations of the SST in the Pacific can cause varying weather conditions, such as severe drought, heavy rainfall, and tropical cyclones. As such, accurate measurements of SST/LST across the entire Earth's surface is a basic requirement for numerical weather prediction and other applications, such as predicting climate change.

There are many methods for measuring the SST/LST, such as ships and surface buoys. However, their coverage is often limited, albeit being accurate. Satellites, on the other hand, can offer much better coverage from space through specialised sensors. For instance, the Sentinel-3 satellite, a mission operated jointly by the European Space Agency (ESA) and by the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), houses the Sea and Land Surface Temperature Radiometer (SLSTR) for measuring SST/LST. It is possible to make direct measurements of surface temperature from these satellites everywhere, except in the presence of cloud. The presence of clouds can significantly affect the signals measured by these satellites, and thus can contaminate the retrieval of SST/LST.

Therefore, the presence of cloud must be established, so that no surface temperature retrieval is performed wherever there is cloud. This operation is known as cloud masking. The cloud masking operation, in its simplest form, involves processing thousands of satellite images and marking each pixel in these images, as cloud or as clear sky. To avoid expensive human interaction, this operation must be automated, and computational methods are developed for this purpose. The cloud masks currently used are very accurate, but no cloud mask is perfect and there are a number of scenarios that make cloud masking a challenging task. For example, differentiating clouds from sea ice, identifying clouds in the presence of sun-glint, identifying low warm clouds, and differentiating clouds from dust storms in the desert or fires/smoke. Below shows an example, where the atmospheric dust in the Arabian Peninsula can challenge the masking process.

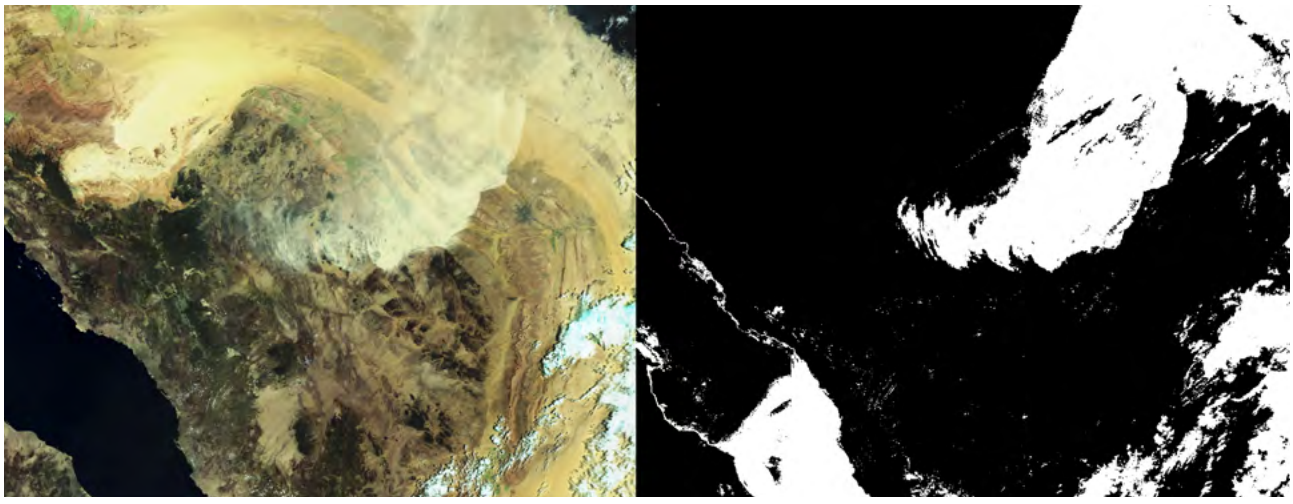
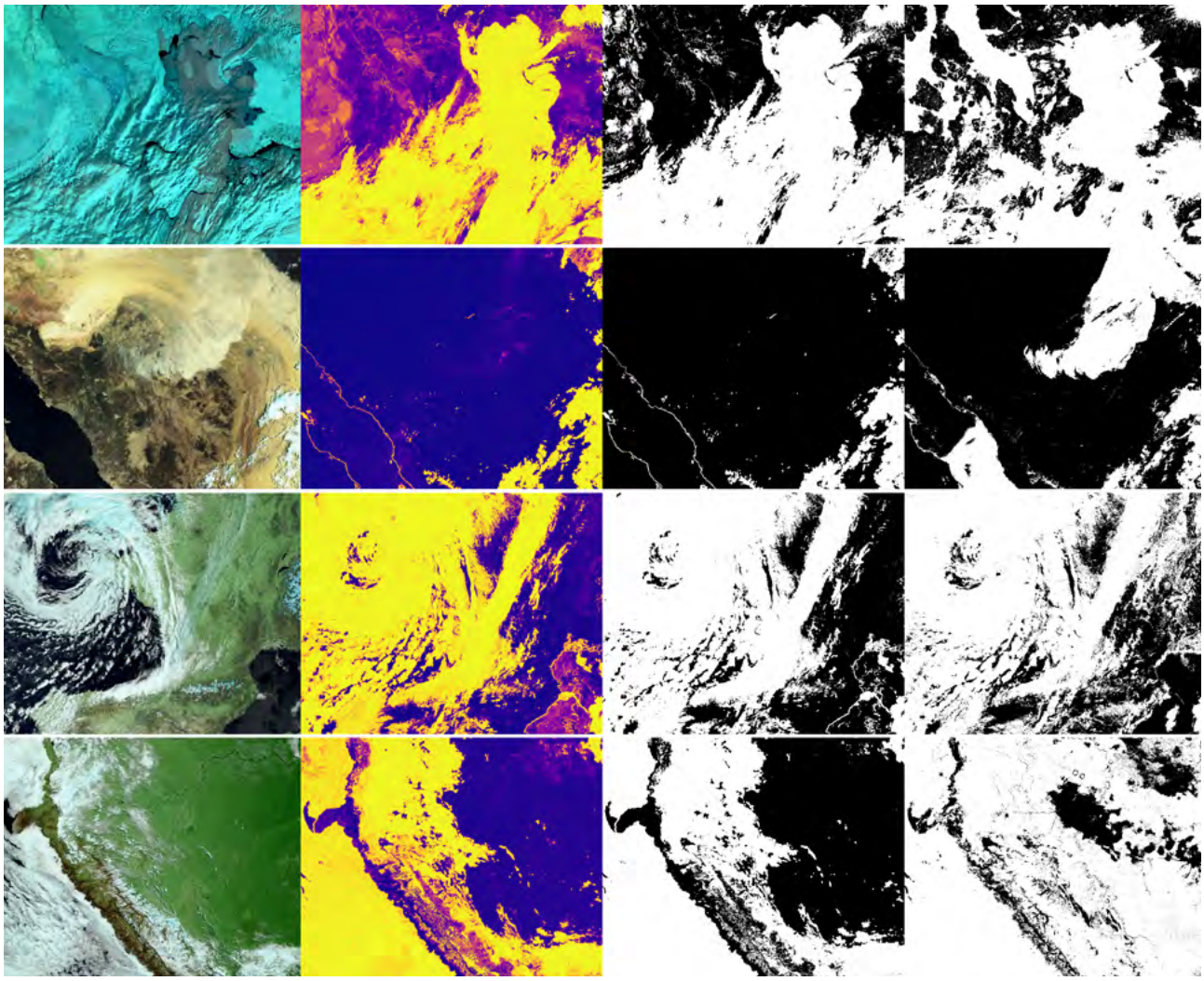


Illustration of atmospheric dust over the Arabian Peninsula being mistaken for cloud. Left: Cloud over the Arabian Peninsula. Right: Incorrect cloud mask.

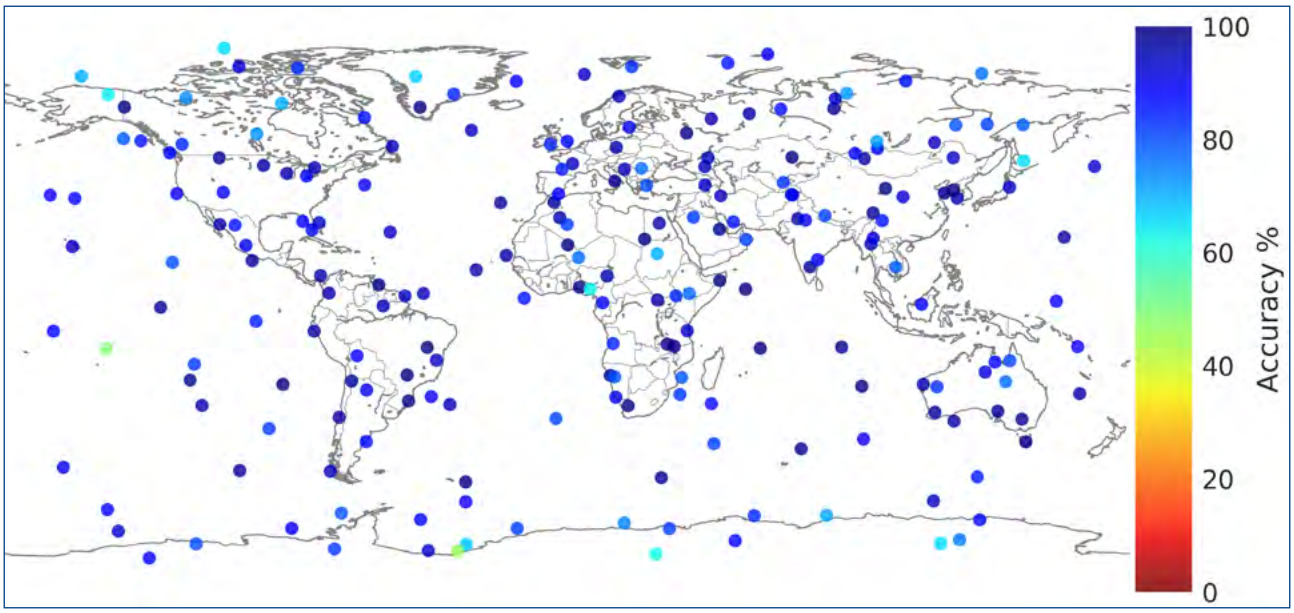
A project undertaken by the Scientific Machine Learning (SciML) group at the Scientific Computing Department, in collaboration with RALSpace, investigated the use of artificial intelligence (AI) for cloud masking, resulting in an intelligent cloud masking.

The research group trained different AI solutions to do this under different conditions, such as snow, sea ice and sun-glint, using several hundreds of images. Some of these AI solutions were trained on a powerful supercomputer, called PEARL, that contains state-of-the-art AI-specific computer hardware. The usage of the PEARL supercomputer accelerated the training process, from several weeks, to a few hours.

The results from one of those AI systems are shown below for different conditions.



Different examples of cloud masking. We show the masking outputs across different regions (one region per row), comparing the original false colour image (column one), AI output with each pixel marked with probability values (column two) and final AI-based mask (column three), and current masks used by the satellites (column 4). The regions are the Arctic, Arabian Peninsula, Europe and Amazon forests, from row one through row four, respectively.



Accuracy of the AI method at different locations when compared against hand-labelled pixels. The image shows lower colours (red) and higher colours (blue) indicating regions where our method performed significantly worse and better when compared against the hand-labelled pixels, respectively.

One of difficulties in developing cloud masking methods is in validating the results. We have validated our findings against an independent dataset, which contained tens of thousands of pixels from cloud images, scattered across the globe, that were manually classified as cloud or as non-cloud by subject experts. The figure above shows the accuracy of our AI method against this manually screened dataset at different locations of the Earth.

This effort is a first step towards a better and intelligent cloud masking method. AI-based methods can offer excellent outputs, and this is an ongoing, active research between SciML and RALSpace.

Acknowledgments: *This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "AI for Science" theme within that grant & The Alan Turing Institute.*

People involved: Sam Jackson (SciML , SCD), Caroline Cox (RALSpace), Jeyan Thiyagalingam (SciML , SCD) and Tony Hey (SciML , SCD).

Author: Jeyan Thiyagalingam, SciML Group

Molecular digital design for the pharmaceutical industry

The pharmaceutical industry is an important leading activity in the UK, worth £42 billion and employing 73,000 people[1]. However, global industrial competitiveness poses new challenges in terms of the research and manufacturing sustainability to the industry, which are also inherently costly to maintain.

Furthermore, changes in patient expectations to provide improved quality health care stipulates a new and innovative technological approach, namely personalised medicines and nanomedicines for site-directed therapeutic treatments. This requires a major shift from the traditional "one-size-fits-all" drug formulation and manufacturing processes, which remained unchanged for the past 40 years, to smart, knowledge based and quality-by-design approach that can response quickly to demands and needs.

Computer modelling is an innovative way to design and study drug molecule behaviour digitally, rather than using the traditional experimental approaches which can be expensive, and sometimes bring little or no underlying scientific knowledge to a problem.

To bridge the challenging gap between chemistry and engineering designs at molecular levels, the digital approaches by means of computer simulations and data analytics have become increasingly important.

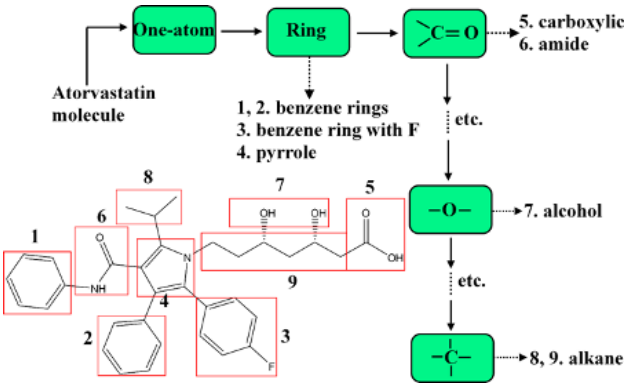
1. 2017 figures.

Modelling techniques such as molecular dynamics can be used to simulate the motion of molecules in bulk solution or solid/liquid interfaces. These are essential computational techniques to provide information on the physical and chemical behaviour of the compound materials at atomistic details that are not readily assessable by experimental means.

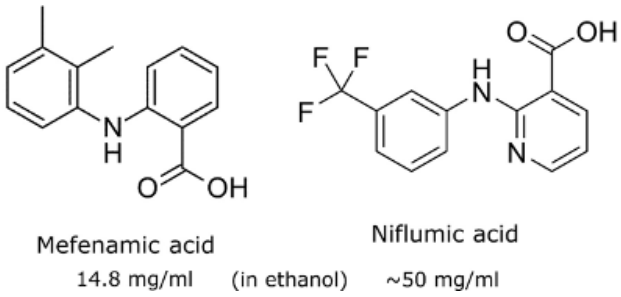
Three separately developed software packages at Daresbury Laboratory have been integrated to form an efficient computational infrastructure to search for clues as to how some underlying effects dictate pharmaceutical processes. These packages are called DL_FIELD, DL_POLY and DL_ANALYSER. They enable scientists to easily set up complex molecular models, running calculations or doing computer modelling and carry out results analysis. Quite often, computer modelling is the only feasible way to investigate some important fundamental phenomena. For example, how and why atoms in different drug molecules interact with solvents differently, that give rise to different crystal shapes. This kind of knowledge is important for the pharmaceutical process industry, for instance, to make medicines.

To be able to study this sort of behaviour systematically, two inter-related unique features have been invented and included in the software. The first one is the DL_F Notation, through which the software can identify the chemical behaviour of every atom in a drug molecule. Secondly, DANAI, a symbolic syntax that can annotate precisely how atoms interact with one another. These features enable scientists to carry out calculations in computers and provide a better understanding of how drug molecules behave at molecular levels. For instance, by comparing two rather similar anti-inflammatory drugs, namely, mefenamic acid and niflumic acid, the computer modelling software can identify and annotate the atomic interactions involved. From such, it provides clues as to why niflumic acid is more soluble than mefenamic acid in ethanol.

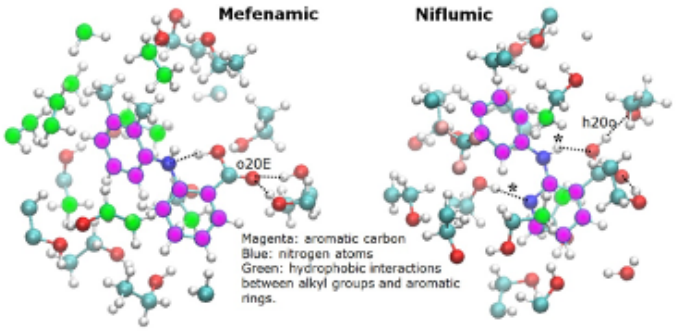
The work demonstrates the use of molecular simulation software infrastructure developed at Daresbury Laboratory, to carry out digital experiments on drug molecules to extract detailed atomistic information that is otherwise not possible from experiments. Two unique features implemented in DL_FIELD and DL_ANALYSER have the potential to carry out data analytics, and from such, to determine relationships between the molecular chemical structure and pharmaceutical formulation and engineering processes.



The atom type detection flow diagram, using atorvastatin medicine as an example, of which the molecular sketch is shown in the Figure. The molecular structure can be expressed in several popular file formats such as PDB and xyz and DL_FIELD will parse the information (boxes in red outlines) through a series of functional analysis subroutines (green boxes) to determine the chemical functional groups of each atom in the molecule.



Molecular structure of mefenamic acid and niflumic acid, showing the compound solubilities in ethanol.



The molecular configurations reveal the reason for the difference in the interactions. For mefenamic acid, the carboxylic group tends to orient in such a way so that the h20O can interact with the aniline N within the molecule, forming a stable cyclic conformation. The dotted lines represent HB interactions.

Authors:
Chin W. Yong and Ilian T. Todorov, Computational Chemistry Group

Ontologies for materials modelling marketplaces

Semantic technologies are a branch of information science which aims to encode knowledge in a machine-readable way, to enable automatic reasoning and a consistent exchange of information between heterogeneous sources. A notable example is the Semantic Web concept, an evolution of the World Wide Web proposed in the 1990s, that is based on semantics rather than ad-hoc links between resources (e.g., web pages). Many tools and search engines that we use every day on the web involve methods of this type.

Semantic interoperability refers to an agreement between multiple data infrastructures on the terms by which potential scenarios and their context can be described. In the presence of such an agreement, different platforms can annotate data with metadata consistently, employing the same concepts and relations, so that the meaning of any communicated content is clarified and misinterpretations are avoided. This is a good practice in data stewardship, by which data are supported and enriched to become FAIR: Findable, accessible, interoperable, and reusable.
Ontologies – specifications of what can exist (hence the philosophical sounding expression) – are a formalism for establishing the required interoperability standard within a domain of knowledge.

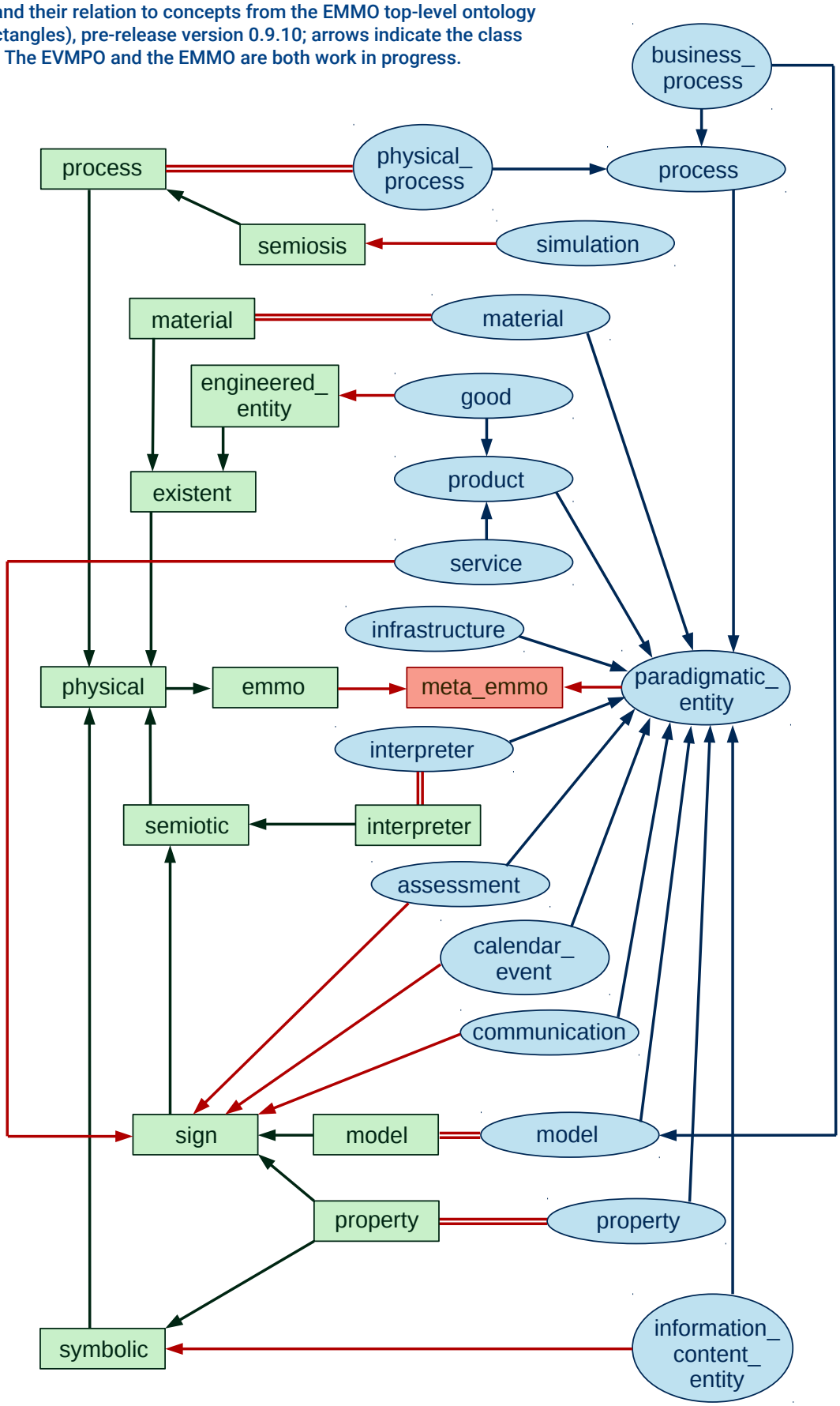
Knowledge bases that employ semantic technology can be split into two components: The ontology, which is also called TBox (terminological box), is one of them; it deals with classes (or "universals") and possible relations between objects (or "individuals"). The second component is the ABox (assertional box) which provides information on the objects and relations that exist in a concrete scenario. The same formats, which are typically based on the resource description framework (RDF) and the web ontology language OWL, can be applied to both components of the knowledge base; in the present work, the terse triple language (TTL) is employed for this purpose. Non-relational databases (triple stores) can be used to store such information directly. Additionally, FAIR digital objects from an ABox, representing parts of an application scenario, can be communicated in a variety of formats with a reduced expressivity such as JSON and XML, and large data sets can be stored together with the required metadata using HDF5 based formats such as the Allotrope Data Format (ADF).

On the Virtual Materials Marketplace (VIMMP), a multitude of providers and users of services, data, and software will come together to interact, contributing to the uptake of multiscale materials modelling in industrial engineering practice. The functionalities of the VIMMP marketplace platform are reflected by a system of domain ontologies which is developed by the Computational Chemistry Group in coordination with project partners and the wider community, including related projects and efforts associated with the European Materials Modelling Council (EMMC), the European Open Science Cloud (EOSC), the Allotrope Foundation, and the Research Data Alliance (RDA). These ontologies aim at connecting platforms, models, and tools in computational molecular engineering, integrated computational materials engineering, and related applications of materials modelling and characterization. For this purpose, a multi-tier modular system is established as follows:

- Top-level ontology: At the most abstract level, the European Materials and Modelling Ontology (EMMO) is used. This is an ongoing development from the EMMC to which our group does not contribute directly; it is used by VIMMP jointly with many further projects and infrastructures.
- Marketplace level fundamentals: High-level concepts for services in materials modelling, referred to as fundamental categories, are specified within the European Virtual Marketplace Ontology (EVMPO). This ontology is co-developed and shared by VIMMP and MarketPlace, a second project that develops a complementary platform.
- Marketplace level ontologies: Expanding on the EVMPO fundamentals, this system of ontologies supports the data ingest, processing, and retrieval functionalities of the VIMMP marketplace platform. It characterizes data and metadata required to achieve service and model interoperability.
- Subdomain-specific metadata schemas, following a subdivision of simulation granularity levels into electronic, atomistic, mesoscopic, and continuum methods.

Further information on VIMMP, which is funded from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 760907, can be found on https://vimmp.eu/

This diagram visualizes the fundamental categories from the EVMPO (blue ellipses) and their relation to concepts from the EMMO top-level ontology (green rectangles), pre-release version 0.9.10; arrows indicate the class hierarchy. The EVMPO and the EMMO are both work in progress.



Authors: Martin Thomas Horsch and Silvia Chiacchiera, Computational Chemistry Group.

Delivering solutions for data-intensive science

Data and software for 21st century large-scale facilities science

From proteins or from protons, experimental data provides the basis for scientific analyses and discoveries. As the experimental data complexity and volume increase, software becomes a crucial component in the analysis process. Thus, both data and software are crucial for 21st century research.

Large scale facilities, such as synchrotrons and neutron and muon sources, generate vast amounts of complex data that need to be managed in an efficient way. This involves supporting all the processes from data generation to long-term storage and archival, encompassing data retrieval, processing, analysis, and data publication workflows. These processes require us to produce and maintain rich associated metadata (or data about the data), making the data **FAIR**[1] (findable, accessible, interoperable and reusable). It also means providing access to an efficient and responsive digital infrastructure.

The SCD Software Engineering Group (SEG) has, for over 10 years, developed and maintained software and services supporting good data management practices across the large-scale facilities at the Rutherford Appleton Laboratory, such as Diamond Light Source (DLS) synchrotron and ISIS neutron and muon source.

Since its inception, the data cataloguing system -- mainly providing storage and data retrieving functionality -- has evolved to support updates on data policies and requirements. For example, in recent years, the ISIS catalogue has been upgraded to support open data, where each proposal is assigned a Digital Object Identifier (DOI), i.e. a persistent identifier that can be used to access information about the proposal and to access the data. Other recent updates are related to Diamond's Data Store and new web interfaces to provide access to the data, as described in the following sections.

1 https://doi.org/10.1038/sdata.2016.18

Diamond Data Store

The current Diamond archive has been collecting data since 2007 and holds over 24.1 PB of data (figure from early March 2020). The data volume is expected to keep growing, user behaviour is adapting to the new situation, and there is also a new data policy implying the data may be made publicly available after a three-year embargo period.

The **Diamond Data Store** (DDS) project, funded by the Ada Lovelace Centre, involves SCD and DLS working to enhance the current system, to provide a better user experience and continued provision of a reliable service, by investigating:

- A more modular architecture.
- More differentiation of data workflows and storage capabilities based on data characteristics and use/reuse scenarios.
- Standard Application Programming Interfaces (APIs).
- Richer scientific metadata to support FAIR and open data.

The project was structured in three phases: a) a scoping survey was performed by a contractor (Sep 2018- May 2019), documenting the current status and providing some recommendations and technical pointers, b) we are currently (June 2019-Dec 2020) in an 'evolution' phase, evaluating existing technologies and products, and working on incremental improvements to the architecture, c) the 'revolution' phase (Jan 2021 and onwards) will implement system changes based on the Open Data vision being developed by DLS.

The highlights of the current work are: a) new functionality was added to restore data to computing clusters at Diamond and STFC Scientific Computing (the SCARF cluster), b) an evaluation of the ICAT metadata catalogue's resilience to the metadata growth was performed with reassuring results, c) other metadata catalogues, i.e. SciCat and ICAT+, are being evaluated to understand their functional and non-functional offerings. We are also working on defining pilot projects to guide proof-of-concept improvements.

SciGateway and DataGateway

As part of the data catalogue system, we are now designing and developing **SciGateway**, a web portal to large-scale facilities science, and one of its components, **DataGateway**, which is a web portal for accessing the facilities data. DataGateway will replace Topcat, which is the current web interface.

SciGateway is the parent web application and provides common functionality such as authentication (ie. logins), notifications, and cookies management. Our ultimate vision is that SciGateway will provide 'exactly one interface' or point of access to large-scale facilities science. At a minimum, its common interfaces will enable better integration between systems. Specific plug-ins will deal with the different processes of the research data lifecycle by providing specialised interfaces.

In turn, DataGateway focuses on providing data discovery and data retrieval functionality, independent of the data storage system it is connected to. This is achieved by building appropriate service interfaces that provide transparent access to the underlying metadata catalogue and data storage solutions. DataGateway enables data browsing, searching, and data visualisation.

Our objective is to make this development user-driven, and we are performing different requirements-gathering exercises to understand the way that users currently manage their data during the research lifecycle. Based on these, we are creating user stories and requirements that guide the software development process.

Last but definitely not least, SciGateway[2] and DataGateway[3] are open-source ReactJs[4] components that follow a micro-frontend architecture, and we are following best software engineering practices during development.

SEG's mission is to design and develop high-quality software that enables large scientific facilities to manage their data and software tools. In this article, we mentioned the recent developments around the Diamond Data Store project and the DataGateway web interface. If you are interested in these projects, feel free to contact us[5].

Author:
Alejandra Gonzalez-Beltran, Software Engineering Group.

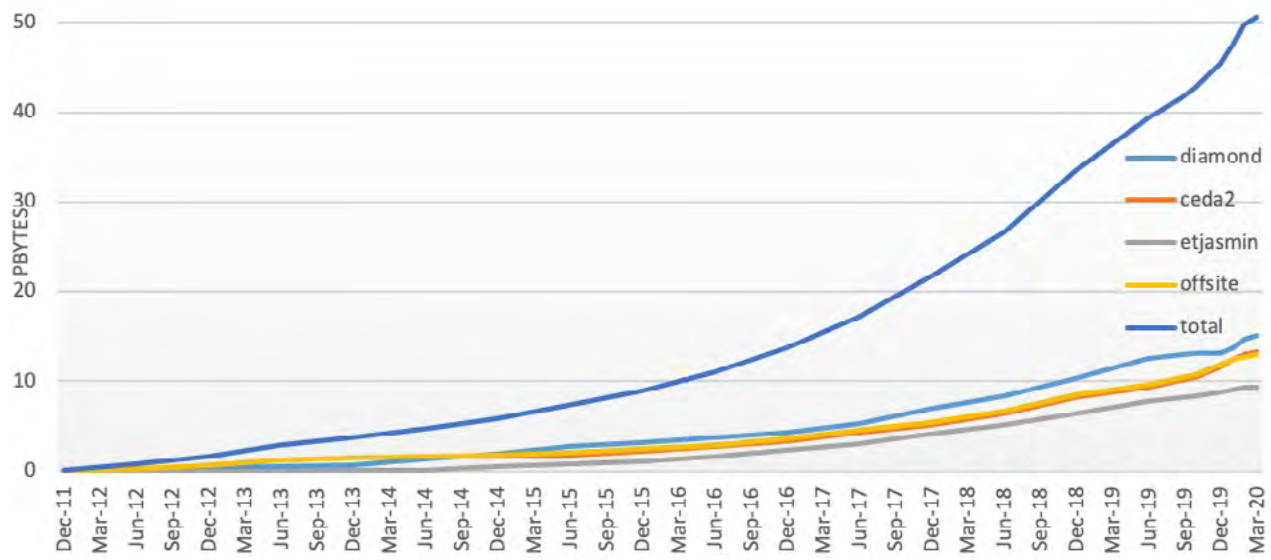
2 https://github.com/ral-facilities/scigateway
3 https://github.com/ral-facilities/datagateway
4 https://reactjs.org/
5 alejandra.gonzalez-beltran@stfc.ac.uk

March of the Tape Robots

The Scientific Computing Department provides data archive services for projects and facilities which have an ever-increasing need to store and manage their data; for much of this data the expectation is that the data will be stored indefinitely. From particle physics data from the Large Hadron Collider, the Centre for Environmental Data Analysis (CEDA) to data from the IRIS science communities: astronomy data, the ISIS Neutron Source, Central Laser Facility and the Diamond Light Source (DLS) the requirements continue to grow.

In addition to disk-based storage systems, SCD need to provide for long-term data storage and for a low cost-per-gigabyte solution to meet the needs of the experiments. High-capacity automated tape libraries provide the best fit for this and to meet these requirements we now have two Spectra TFinity tape libraries, A.K.A. tape robots, installed in the Scientific Computing Department Data Centre at Rutherford Appleton Laboratory (RAL). These will, in time, replace the use of the current Oracle SL8500 libraries which are over a decade old and allow for expansion in tape storage over the next decade.

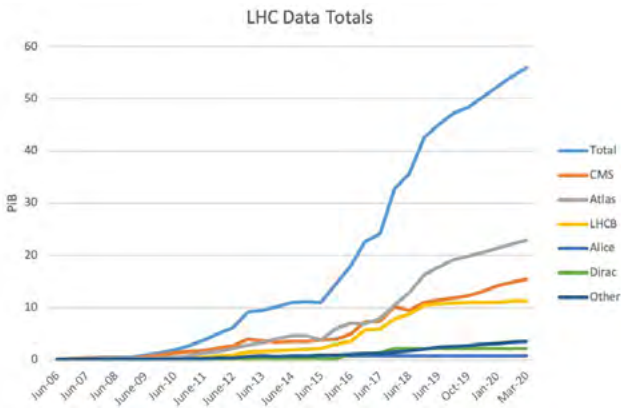
The first of the Spectra TFinity tape libraries was installed last year in order to provide 65PB capacity for CEDA and IRIS. This initial installation comprised 7 frames with a mix of the latest tape drive technology, the industry open standard LTO-8M (Linear Tape Open) Drives and TS1160 Drives. The Spectra TFinity provides the ability to expand on this initial capacity and the Diamond Light Source Ltd. have recently taken advantage of this to expand the library with an additional two frames including 12 TS1160 enterprise drives for their Diamond Archive data. This brings an increase in capacity for this Spectra library to just under 100PB.



Data taking from CEDA and the Facilities has increased to over 50PB.

The GridPP Tier-1 service which is run by STFC at RAL also provides for tape storage for the LHC experiment data, their data volume growth is shown here. Note that the plot on the right shows the last year in quarters, so the rate that data is being written to the tape library is continually growing.

In order to meet the needs of the experiments as they continue to expand their data taking, a 9-frame Spectra tape robot with a capacity of around 130PB was delivered in January 2020 and is now



The Spectra TFinity Tape Robot for CEDA and STFC data.

operational. Over the next few months, this tape robot will be integrated with the tape management system to ensure the continued archiving and availability of LHC data.

During LHC Run 3, RAL will archive a little over 10% of the raw data from CERN which is around 40PB of data each year. This is the equivalent data rate of continuously streaming 400 4K (Ultra HD) videos and would saturate a 10Gb/s network link for the entire year. In addition to this approximately 60PB of data will need to be migrated from the current libraries. The library is expected to hold around 190PB of data by 2024, which will require a new generation of higher density tape storage as well as additional frames.

Authors:
Alison Packer and Alastair Dewhurst, Systems Division

The Spectra TFinity Tape Robot for GridPP data.

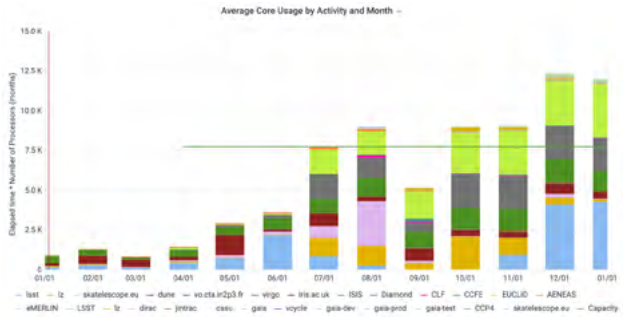
From Europe to IRIS: Key SCD projects in 2019

Driven by the physics communities supported by UKRI-STFC, the eInfrastructure for Research and Innovation for STFC, or IRIS, is a collaboration of STFC's science activities and its national computing centres at universities and STFC's own sites. IRIS's vision is to develop a single federated national computing Infrastructure for STFC science. For many years, SCD has played a leading role in the development and operation of global scientific computing infrastructures, such as the Worldwide LCG Computing Grid (WLCG) and the European Open Science Cloud (EOSC) and its forerunners. SCD is now exploiting tools developed in the international context and tailoring them to meet its national requirements. Three specific areas – resource accounting, identity management, and trust and security frameworks – are of immediate and essential benefit to IRIS.

Identity and Access Management (IAM)

Having a coherent identity and access management framework defines and characterises the experience of using the IRIS infrastructure. SCD's experience with INDIGO IAM from previous European projects, along with its ease of deployment, made this an obvious choice for the IRIS IAM service. Now in operation, this service allows users to log in with credentials from their own institutions, allowing streamlined access to services as they are added to IRIS IAM. The technologies and workflows used by this service are widely used by research institutions and universities as well as social media platforms like Facebook, Google and Twitter. This leads to an authentication process which is both familiar and sustainable given its wide adoption. If you run a service that you think would benefit from this capability, please contact Tom Dack (thomas.dack@stfc.ac.uk).

Accounting (APEL)



Accounting Dashboard view of IRIS resource usage for calendar year 2019.

APEL has been providing the resource accounting for WLCG for many years, currently recording over 600 million compute jobs per year. The development, support and operation of APEL have been primarily supported by a series of European projects, starting with EGEE in the early 2000s. A dedicated APEL Accounting Repository has been deployed for IRIS, alongside the creation of a new web-based Accounting Dashboard to view the data stored in the Repository. It is built on a MySQL database with a lightweight customisable Grafana dashboard on top, deployed by the APEL team at STFC, which provides access to the accounting data. The Dashboard uses the IRIS IAM for authentication and authorisation. Additional views are being developed for the GridPP Tier1 and the STFC Openstack Cloud, making use of the digital asset that was created for IRIS.

Trust and Security Framework

Any collaboration requires clear rules of engagement. The IRIS Trust and Security Framework provides a basis for service providers to operate – and for users to perform their work – in a safe and secure manner. The European AARC (Authentication and Authorisation for Research Communities) projects built on considerable experience in creating policy for European and global projects, such as EGI and WLCG, to develop a Policy Development Kit. This work is being used by IRIS to develop a set of security policies – one of the first communities to do so. This policy set enables IRIS to work with others on an international scale that use similar frameworks. This process both helps secure the infrastructure and gives useful insights into its structure.

Service and topology information (GOCDB)

Another area of vital benefit to IRIS will be added in 2020: management of the services used by IRIS and their contact information. This is essential to make the most effective use of the resources available to IRIS. GOCDB, which also began under the EGEE project in the early 2000s, is a proven scalable solution for managing the service registry and topology information of a large-scale eInfrastructure, with a current database of 369 sites and 3925 services. The GOCDB and APEL teams work closely together, with APEL being integrated with GOCDB during their parallel development.

Author:
David Crooks, Distributed Computing Infrastructure Group

Sharing knowledge

Enabling Open Science

Teams within the Scientific Computing Department are working on a number of research and development projects investigating means to support, enable and advance open science practices.

> *Open science aims to ensure the free availability and usability of scholarly publications, the data that result from scholarly research, and the methodologies, including code or algorithms, that were used to generate those data*[1].

In recent years there has been a global move towards open science that involves not only Open Access to articles but also opening up of research data and other outputs of research. Whilst the ideas and principles of open science are well established in policy, adoption of open science in practice is not without barriers and the implementation of organisational practices and services to support open science are still evolving.

We are participating in a number of projects under the Horizon 2020 funded European Open Science Cloud (EOSC) Programme. Each of the projects draws on specific strengths within the Scientific Computing Department and positions us to develop national and international influence in the areas of open science. The three projects: FAIRsFAIR, Freya and ExPaNDS, whilst separate are complimentary and progress solutions and approaches for opening up scientific research:

FAIRsFAIR – fostering FAIR data practices in Europe:

FAIRsFAIR project started in March 2019 and aims to supply practical solutions for the use of the FAIR data principles throughout the research data lifecycle. FAIR refers to the concept that data should be Findable, Accessible, Interoperable and Re-Usable. STFC are one of the core partners in the project leading a package of work looking at the development of a FAIR competence centre to provide information and guidelines on FAIR data practices. Whilst FAIR as a concept is not restricted to open data as the FAIR principles can apply equally to closed data our involvement fits within our portfolio of work and research into supporting open science at STFC National Laboratories and the broader goals to enable open access to data resulting from publically funded research.

FREYA - connected open identifiers for discovery, access and use of research resources:

Freya is a three year project running until the end of 2020. The project aims to extend the infrastructure for persistent identifiers as a core component of open research. Working with project partners, SCD is developing the concept of the PID Graph which draws the connections between different persistent identifiers systems to link together the relationships between people, datasets and publication to improve discovery of and understanding of how different research outputs relate to one another.

ExPaNDS - the European Open Science Cloud (EOSC) Photon and Neutron Data Service:

ExPaNDS is the last of the projects to have got underway and is still in the early stages. ExPaNDS relates directly to the photon and neutron community drawing on the experience of SCD to provide services and support to STFC's ISIS Neutron and Muon Source and Diamond Light Source who are also partners in the project. Like FAIRsFAIR, ExPaNDS is also drawing on the FAIR data principles and will be looking at the adoption of these principles in the Photon and Neutron community with the aim of improving access to the data behind the scientific papers generated by the Facilities.

The three projects give a flavour of the work that we are involved in to improve support for and access to research outputs, enabling open science and progressing FAIR culture and practices.

Further information:

FAIRsFAIR: www.fairsfair.eu
FREYA: https://project-freya.eu/
ExPaNDS: https://expands.eu/

Author: Elizabeth Newbold, Open Science Services

1 National Academies of Sciences, Engineering, and Medicine 2018. Open Science by Design: Realizing a Vision for 21st Century Research. Washington, DC: The National Academies Press. https://doi.org/10.17226/25116.

Towards an Open Science Portal for STFC

In 2012 the Royal Society published an influential report 'Science as an Open Enterprise', which examined technological drivers for change in the practice of science, recognised the opportunities for accelerating progress and enhancing quality of science through openness, and defined Open Science as 'open data (available, intelligible, assessable and useable data) combined with open access to scientific publications and effective communication of their contents.' The idea of openness goes beyond simply making the outputs of science openly available in ways that they were not before, to encompass new opportunities from linking together these outputs to reveal the context of the scientific work and thus make it more meaningful and reusable: linking publications to the data underlying them, connecting datasets with similar scope, and indeed including connections to researchers, their collaborators, their projects and publications.

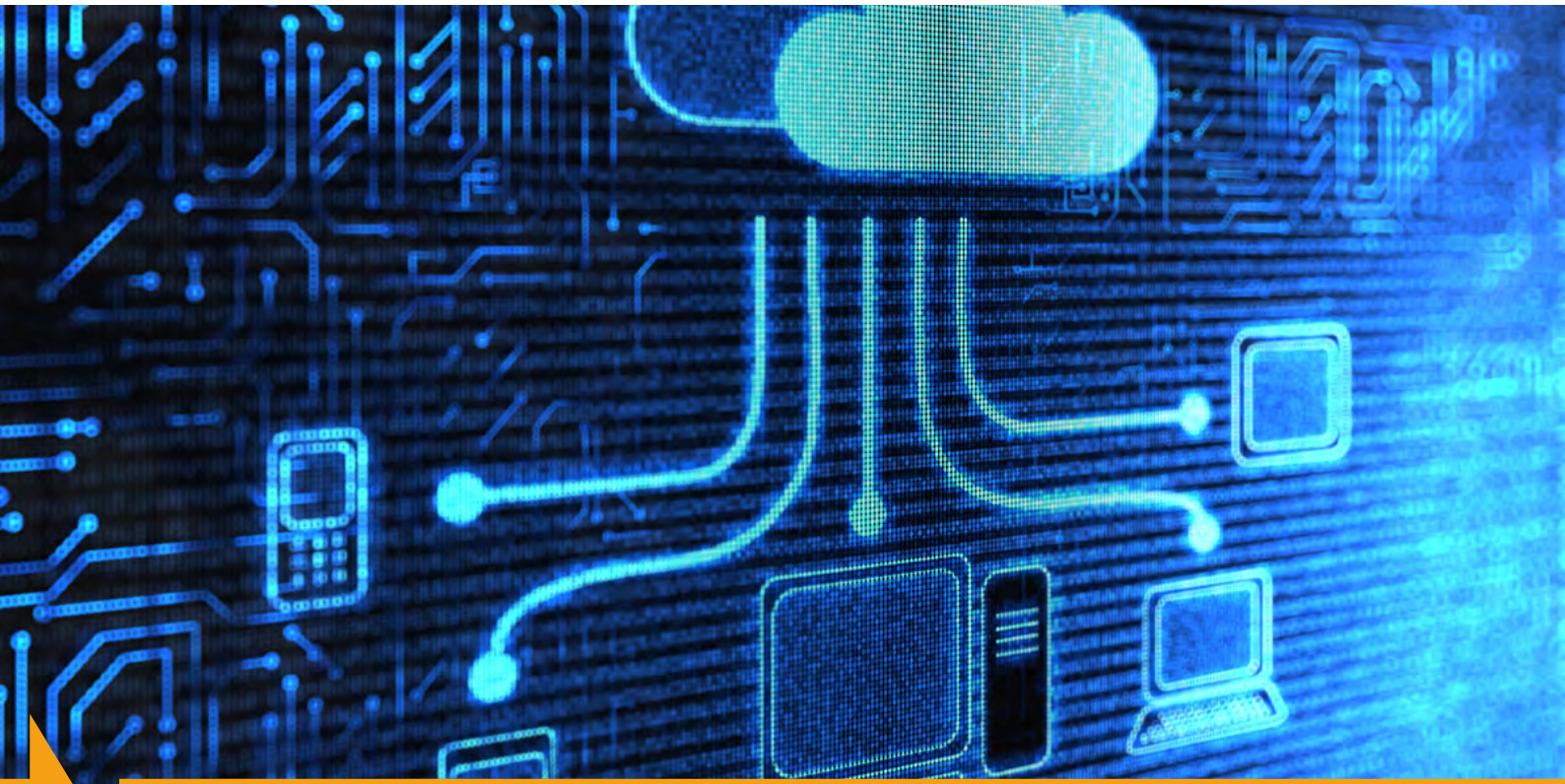
In order to achieve this kind of linking, it is necessary to have precise identification of the things that are to be linked—and this is where Persistent Identifiers (PIDs) are essential. PIDs are labels that identify digital and real world objects. 'Persistent' means that there is a guarantee that the identifiers will continue to be available and correctly identify their objects in the future. Examples of PIDs are Digital Object Identifiers (DOIs), commonly used for identifying research publications and datasets, and ORCIDs used for unambiguously identifying researchers, but there are many other types.

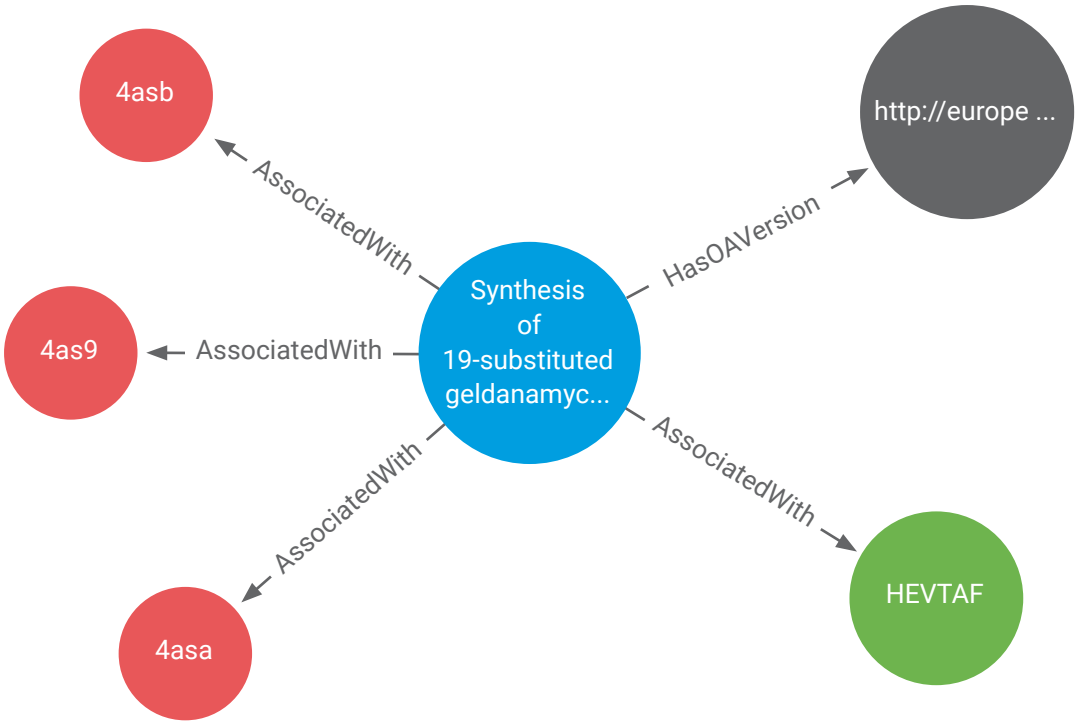
Scientific Computing Department of STFC coordinates FREYA, a Horizon 2020 project[1] developing the infrastructure for PIDs in Europe and globally. It brings together leading PID service providers, publishers and PID users to develop and embed PID-based services and applications. FREYA is one of the projects building the European Open Science Cloud, an important initiative to federate existing scientific data infrastructures to foster open science and open innovation.

One of the drivers for FREYA is the vision of the 'PID Graph', a rich and extensive network of research entities linked through their Persistent Identifiers. The PID Graph can serve as the basis for new services and insights through querying or visualisation. The PID Graph is motivated by such questions as:

> *'As a data centre, I want to see the citations of publications that use my repository for the underlying data, so that I can demonstrate the impact of the repository.'*

1 The FREYA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777523.





'As a researcher, I want to discover data collected by an instrument mentioned in a paper I just read because that data may be useful in my research.'

'As a funder, we want to be able to find all the outputs related to our awarded grants, including block grants such as doctoral training grants, for management information and looking at impact.'

The FREYA project is now in its last phase, and the team in Scientific Computing Department is developing a pilot application of the PIDs-rich knowledge graph for STFC's facilities-based science and other flavours of STFC sponsorship, such as research grants and studentships.

Collecting records of science from multiple information sources and linking them in the common knowledge graph brings new information context that cannot be obtained by looking in the individual sources. The illustration above shows an example: a publication in Diamond bibliographic database (central node) connected to its Open Access counterpart in EuropePMC (top right), to Cambridge Structural Database record (bottom right) and to three Protein Data Bank records on the left. The metadata sources from which these records are harvested are not necessarily directly connected to each other, but their integration in the knowledge graph provides a common context and allows cross-walks between any of the records of science involved.

The aim is to take this work forward into a new research information infrastructure as an 'STFC Open Science Portal'. The vision is a publicly available resource for the discovery of records of science that have been produced with the support of STFC funding or other types of sponsorship (such as facility time awarded to visitor scientists). Such a portal could have many uses by STFC and external stakeholders, contributing to knowledge preservation and knowledge discovery, raising the visibility of STFC-sponsored research, demonstrating STFC adherence to the principles of Open Science, and supporting practical applications of these principles to impact studies, professional engagement with other research organizations and funders, and to public engagement.

Further information:
FREYA: www.project-freya.eu
European Open Science Cloud: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

Author:
Simon Lambert, Data Science and Technology Group

Contacts

Tom Griffin	**Director, Scientific Computing Department**	**tom.griffin@stfc.ac.uk**
Shirley Miller	*Personal Assistant*	*shirley.miller@stfc.ac.uk*
Dr Peter Oliver	**Head of Operations**	**peter.oliver@stfc.ac.uk**
Sally Prydderch	*Personal Assistant*	*sally.prydderch@stfc.ac.uk*
Professor Tony Hey	**Chief Data Scientist**	**tony.hey@stfc.ac.uk**
Sally Prydderch	*Personal Assistant*	*sally.prydderch@stfc.ac.uk*

Division Heads and Group Leaders

Dr Juan Bicarregui	**Head of Data**	**juan.bicarregui@stfc.ac.uk**
Amanda Chapman	*Personal Assistant*	*amanda.chapman@stfc.ac.uk*
Group Leaders		
Dr Brian Matthews	Data Science & Technology/DAFNI	brian.matthews@stfc.ac.uk
Elizabeth Newbold	Open Science Services	elizabeth.newbold@stfc.ac.uk
Mr Gordon Brown	**Head of Software Infrastructure**	**gordon.brown@stfc.ac.uk**
Amanda Chapman	*Personal Assistant*	*amanda.chapman@stfc.ac.uk*
Group Leaders		
Gordon Brown	Dynamic Infrastructure	gordon.brown@stfc.ac.uk
Alex Dibbo	Cloud operations	alexander.dibbo@stfc.ac.uk
Dr Alejandra Gonzalez-Beltran	Software Engineering	alejandra.gonzalez-beltran@stfc.ac.uk
Dr Jeyan Thiyagalingam	Scientific Machine Learning	t.jeyan@stfc.ac.uk
Dr Tyrone Rees	Computational Mathematics	tyrone.rees@stfc.ac.uk
Dr Andrew Sansum	**Head of Systems**	**andrew.sansum@stfc.ac.uk**
Amanda Chapman	*Personal Assistant*	*amanda.chapman@stfc.ac.uk*
Group Leaders		
Nick Hill	Research Infrastructure	nick.hill@stfc.ac.uk
Alison Packer	Data Services	alison.packer@stfc.ac.uk
Ian Collier	Distributed Computing Infrastructure	ian.collier@stfc.ac.uk
Dr Barbara Montanari	**Head of Computational Science and Engineering/ CoSeC**	**barbara.montanari@stfc.ac.uk**
Esme Williams	*Personal Assistant*	*esme.williams@stfc.ac.uk*
Group Leaders		
Professor David Emerson	Engineering and Environment	david.emerson@stfc.ac.uk
Dr Gilberto Teobaldi	Theoretical and Computational Physics	gilberto.teobaldi@stfc.ac.uk
Professor Ilian Todorov	Computational Chemistry	ilian.todorov@stfc.ac.uk
Dr Martyn Winn	Biology and Life Sciences	martyn.winn@stfc.ac.uk

Administrative Team

Karen McIntyre	karen.mcintyre@stfc.ac.uk
Georgia Lomas	georgia.lomas@stfc.ac.uk
India Reeves	india.reeves@stfc.ac.uk
Helen Walker	h.walker@stfc.ac.uk

Further information

Scientific Computing Department	www.scd.stfc.ac.uk
CoSeC (Computational Science Centre for Research Communities)	www.scd.stfc.ac.uk/Pages/CoSeC
DAFNI (Data and Analytics Facility for National Infrastructure)	www.dafni.ac.uk
Science and Technology Facilities Council (STFC)	www.stfc.ukri.org
UK Research and Innovation	www.ukri.org

Notes

Science and Technology Facilities Council

Head Office
Science and Technology Facilities Council,
Polaris House, North Star Avenue, Swindon SN2 1SZ, UK

Science and Technology Facilities Council
Rutherford Appleton Laboratory,
Harwell Campus, Didcot, Oxfordshire OX11 0QX, UK
T: +44 (0)1235 445000 F: +44 (0)1235 445808

Sci-Tech Daresbury
Daresbury, Warrington, Cheshire, WA4 4AD, UK
T: +44 (0) 1925 603000 F: +44 (0) 1925 603100

Establishments at:
Rutherford Appleton Laboratory, Oxfordshire
Daresbury Laboratory, Cheshire
UK Astronomy Centre, Edinburgh
Chilbolton Observatory, Hampshire
Boulby Underground Science Facility, Boulby Mine, Cleveland