Data Fairport:

enabling global exchange of research data

On the 13-16 of January 2014 a varied group of stakeholders from the global life sciences data field met in Leiden, The Netherlands, to discuss how they could create an environment that enables effective sharing and re-use of biological research data. *International Innovation* spoke with the initiators and some of the attendees to discuss how public and private stakeholders and international communities can work together to enable secure and effective data 'interoperability'

Professor Dr Barend Mons: Professor of Biosemantics at Leiden University Medical Centre, Head of the Dutch ELIXIR Node and an Integrator for Life Sciences at the Netherlands eScience Center

"I'm one of the co-organisers of the workshop and my main interest is modern big data-driven science, knowledge discovery with the computer based on experimental data and narrative information."

Dr Niklas Blomberg: Director of ELIXIR, the European Infrastructure for Life Science Data

"My role is to participate both as a representative of the ELIXIR infrastructure to help to picture this in the overall landscape as well as understand these community efforts and develop standards going forward."

Dr René van Schaik: CEO of the Netherlands eScience Center

"We have co-funded the Data Fairport conference as this is in our view a great contribution to stimulating eScience in The Netherlands, especially when it comes to data stewardship and software sustainability."

Dr George Strawn: Director of the Federal Networking and Information Technology Research and Development

"I am an observer from the US federal government and especially interested in this conference given the recent requirement to provide open access to scientific results funded by the US federal government covering both scientific articles as well as the supporting data."

Dr Ruben Kok: Director of the Dutch Techcente for Lifesciences, Director of the Netherlands Bioinformatics Centre

"Today's data-intensive life science research poses common challenges to biologists across sectors to integrate, analyse, securely share and manage their research data. We need a global environment that enables biologists and solution providers to make data sharing a common practice."

"Once we have this system in place, it opens up the possibility for all sorts of services, but we need to start with a common vision, a common ground and a common set of rules to work from."

- Dr René van Schaik

Could you briefly explain the main topic of the Data Fairport conference?

BM: The backdrop of the conference is the immensity and complexity of contemporary biological data, and the requirement for data to be understandable to both humans and computers. This poses a daunting challenge where close collaboration between various stakeholders in the life sciences is essential.

RS: The amount of data that are being generated now and the new techniques in biology that are being used are generating a tsunami of very diverse data. More and more funders are demanding sound data stewardship plans for every grant awarded through their funding system, because they realise that they create a lot of value with these data.

GS: At the highest level what we're looking to establish is the interoperability of heterogeneous data sets as we can't expect the data collected by thousands of investigators to be in a similar format.

Why do you think this issue of 'big data' in the life sciences has become so prominent of late?

GS: It's only relatively recently that the disk storage has been large and cheap enough; that computers have been fast enough; and that the networks have had wide enough bandwidth that we could seriously think about storing most things. Now that we can do all this, we see that there are great advantages if we can develop the software to support the hardware and improve data mining into this tremendous source of scientific data.

NB: It is also expensive and a long-term commitment, so you do think twice before you embark on this journey.

BM: Our ability to generate enormous amounts of data has grown much faster than the capacity to store, link and analyse them. Only now are we catching up and creating the data-handling capacity that should have been developed at the same rate as the capacity to generate this data.

What do you see as the main challenges ahead? Is it the technology or are other aspects just as important?

BM: One of the conclusions that we can take from the conference is that the technology needed to make this happen is essentially 99 per cent there. However, to start an initiative that is also endorsed by major funders, the social hurdles that must be overcome are equally important. This includes the challenge to align people – now that everyone realises the importance of data, the danger is that we will get 500 different initiatives to solve the same problem.

RK: With this conference we have taken an important step forward as we have quite a representative group here from many different disciplines and stakeholders. If this group gets behind this initiative, people will take it as a very serious attempt.

NB: I think that at an individual level, most biologists take great care in preserving data both for their own purposes and to make sure publications are well-founded and reproducible. What they require is guidance and support in how to do this with larger datasets and how to follow the new data stewardship requirements from funding organisations.

Making data accessible in itself isn't hard: you just take a hard drive and hook it up to the internet. Making data accessible so that other people can find and use it is difficult. There are a lot of ongoing community initiatives, but there are no widely accepted guidelines for how to do this on a European, let alone global, scale. Researchers are simply looking for support in how to do the right thing.

RK: One of the things we concluded at the conference is that we need a better rewarding mechanism for entering data into public data systems – this is one of the social elements that we need to try and address.

BM: One important 'non-technical' hurdle is that a new profession – the data scientist – is emerging as a key requirement. There is no way that all biologists or any other researchers can become experts in handling data. The Economist predicts a shortage of about 190,000 people globally in 2017-18 that know how to deal with data. However, trusting your data to an outsider is not easy, and there is no structure in universities at the moment to train data specialists and give them permanent positions.

Can you each give your personal vision for the Data Fairport in the long run?

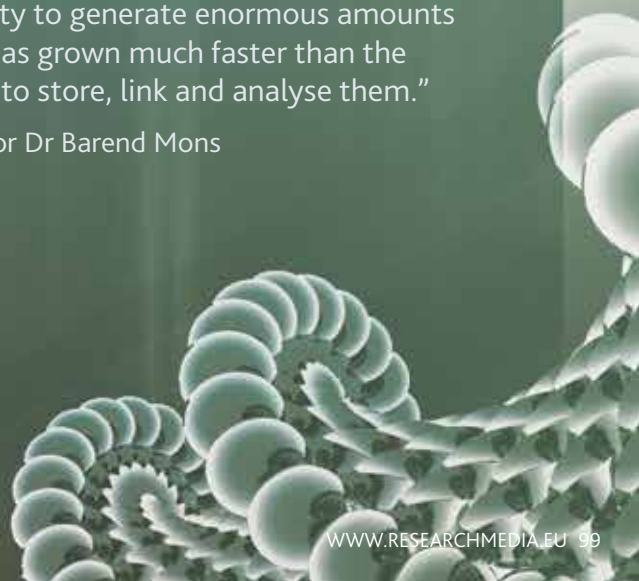
GS: At the highest level, I am hoping that we will develop the technology and the social willingness to work on interoperability of heterogeneous datasets so that we can combine them in novel ways. If we can truly structure scientific data, we will be able to conduct new science.

NB: Interoperability, that's really the key. The other one is longevity – how do we sustain the data going forward? To achieve this, there needs to be agreement on key standards – I would like to see how all of these community

Life science, the study of living organisms, is built on a tradition of cataloguing biological facts. As such, biology is inherently data-intensive, but the digitalisation of information and increased compute and storage capacity of computers and the speed of data transport across networks has created a new age of data-driven and computational opportunities for the life sciences. From the molecular level through to the organism and population as a whole, data capture covers every complex interaction and builds a picture of mechanisms of disease and drivers of behaviour to a resolution never previously imagined. This increases data variety and complexity as much as it drives up data volumes, and this presents many challenges: how can data be successfully integrated, analysed, securely shared and managed among scientists across many different institutions and sectors? These challenges are social as much as they are technological. What is ultimately required is a global and sustainable data sharing environment that makes it easy to publish research data, discover data sets and re-use them. Above all, this requires the global adoption of a series of standards that make data and software talk to each other: they must be 'interoperable'.

Our ability to generate enormous amounts of data has grown much faster than the capacity to store, link and analyse them."

– Professor Dr Barend Mons



efforts are coming together and how we can define a process and a strategy for interoperability.

RS: Once we have all that in place, it opens up the possibility for all sorts of services, but we need to start with a common vision, a common ground and a common set of rules to work from.

BM: With Niklas and George we have representatives from European and US authorities attending. Also at the conference is Professor Abel Packer, who is running Scientific Electronic Library Online (SciELO) for 16 countries in South America. Currently, South Africa and China are also in the process of building up SciELO instances. I think linking SciELO to the Data Fairport backbone would create a great opportunity for millions of bright minds in developing countries to make a career and get into the scientific mainstream.

GS: I would just add optimistically that science already has a community of sharing via research articles, so all we have to do is extend that concept from just articles to articles and datasets. It will be very important for universities and other funders to expand the concept of faculty rewards to include rewards for publishing data, just as now faculties are rewarded for publishing their research articles. This could also be extended to include software.

> "Science already has a community of sharing via research articles, so all we have to do is extend that concept from just articles to articles and datasets."
>
> – Dr George Strawn

What do you hope to have achieved at the end of this week, coming out of this conference?

BM: We are quite modest in our short-term objectives. We have another meeting planned in The Netherlands in September, coinciding with a plenary session of the so-called Research Data Alliance. The first step is to form a steering committee to reach consensus about the minimal requirements and to develop a really solid plan to be presented at that meeting. In parallel we want to raise some initial funds and build some prototypic implementations; so by the end of this year we are ready to start building this thing globally.

NB: There are many initiatives already ongoing and there is a need to find a way to bring the community together and represent the needs of ordinary researchers as well as the longer-term aims of the funding organisations.

GS: If history is any guide, we've seen some community activities with similar aspirations work in the past. In the 1980s-90s for example, a group called the Internet Engineering Task Force arose out of the original foundations of the internet to make community decisions on internet standards and protocols. Then, in the 90s and the

The Data Fairport Conference

13-16 January 2014

25 experts from the worlds of research infrastructure and policy, publishing, the semantic web and more were brought together for four days to discuss how best to deal with life science data and proceed with the Data Fairport. Here, *International Innovation* outlines their conclusions:

AIM

The Data Fairport aims to provide a minimal (yet comprehensive) framework in which current issues in data discoverability, access, annotation and authoring can be addressed. The Data Fairport will not dictate a single platform or a tightly integrated data infrastructure, but will instead focus on conventions that enable data interoperability, stewardship and compliance against data and metadata standards, policies and practices.

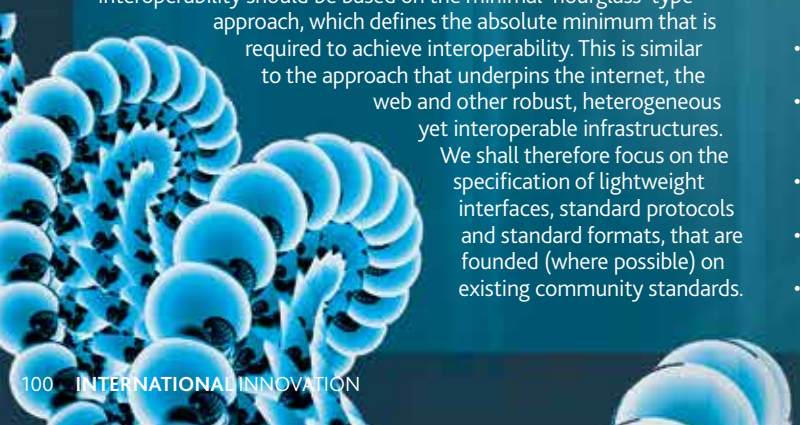
APPROACH

It was proposed that the convention for data and model services interoperability should be based on the minimal 'hourglass' type approach, which defines the absolute minimum that is required to achieve interoperability. This is similar to the approach that underpins the internet, the web and other robust, heterogeneous yet interoperable infrastructures. We shall therefore focus on the specification of lightweight interfaces, standard protocols and standard formats, that are founded (where possible) on existing community standards.

SCOPE

The Data Fairport is not about the development of additional standards, but rather:

- Adoption of standards
- Communication of standards
- Simplification of standard interoperation
- Adoption of cross-cutting standards for provenance, versioning, identity and dependency for data and for metadata covering identifiers, formats, checklists and vocabularies
- Interoperation of data services
- Reconciliation of evolving standards and the datasets organised or annotated by them
- Minimal models of investigation for grouping results
- Metadata required to link data with analytics (notably models)
- Data citation – mechanics, adoption, recognition



2000s, the World Wide Web consortium arose to do the same thing for standards and protocols associated with the Web. Both of these activities are what you would call non-profit community-orientated activities; but they have produced key platforms upon which other entrepreneurs have been able to found very important businesses in service to science and society.

RK: Most of the technological solutions and standards are floating around already. We don't want to re-invent wheels here. So, in the next nine months a major role for our group will be to visit the major players and invite them to participate in a comprehensive and coherent approach to foster data sharing and re-use.

Is this concept unique to life sciences? Do you think it is a model that could be rolled out for other disciplines?

RS: You can definitely apply it to other areas of science. For instance, in astronomy, the datasets are even bigger, but they are also simpler, and the same is true for nuclear physics. There is a lot of noise in the data and they throw a lot away just to get at the interesting parts. Life science is special because of the enormous variety of data that we have to deal with. When we get to the medical domain and start dealing with patient data, there is the added complication of privacy as well.

NB: Social sciences and life sciences have a lot of things in common in the health domain, as they also deal with highly sensitive personal data, so ethical and privacy issues are very similar. Maybe it should be added here that we are talking about life sciences as if they are one discipline, but they are still quite a heterogeneous group, and multidisciplinary their own right.

GS: I would just add that that not only are these technologies ultimately applicable to all science and other scholarly domains, their ultimate value will hopefully be to promote interdisciplinary research. Overlaps between chemistry and biology are well known; and between biology and geology now as climate change is considered – if we can use electronic technology to help us articulate between and among these scientific fields, I think we will create entire new tiers of knowledge.

BM: If you create a contingent of data experts with no specific disciplinary bias, those will be the living connectors between areas in a way – by implementing the same approaches throughout disciplines. So I think in the end what we are discussing here has implications not just for the life sciences but for the wider scientific community as well.

"Interoperability, that's really the key. And the other one is longevity – how do we sustain the data going forward?"

– Dr Niklas Blomberg

KEY DRIVERS

Understanding complex biological systems remains a great challenge for the life sciences. Perhaps the most immediate challenge is the human body, which has between 20,000-25,000 genes located on 46 chromosomes, whose expression is modulated by large amounts of additional genetic elements that constitute multiple layers of regulatory control. The emergence of high-throughput 'omics' research to examine these components and correlate them to health and disease has led data production to increase exponentially. The increased potential to sequence many genes and the cost and time of sequencing rapidly decreasing has led to an overwhelming deluge of data production in medical science. In just 10 years the price of human genome sequencing has diminished from €4 billion (the cost of the Human Genome Project) to around €1,000 thanks to advances in sequencing technologies.

Such complex biological systems and the enormous volumes of data being generated impact greatly on the life science community, including biomedical researchers and clinicians, but also those working for scientific communication outlets, such as publishers. The PubMed database, for instance, has received 20 million biomedical research articles to date – which amounts to one new submission every 40 seconds. This surge of data could lead to inaccuracies in research; new and potentially important data slipping through the net; and expensive research projects needing replication. Proper use of these data, however, has the potential to generate an array of exciting new discoveries. If scientists are to utilise all existing and incoming life sciences data, a shift from the 'old (data poor) paradigm' to a 'new (data intensive) paradigm' will be required. This will potentially require a total shift in the way science is performed and scientific success is measured.

"Most of the technological solutions and standards are floating around already. We don't want to re-invent wheels here".

– Dr Ruben Kok

THE DATA FAIRPORT

OBJECTIVES

The Data Fairport initiative focuses on agreeing conventions that enable data interoperability, stewardship and compliance against data and metadata standards, policies and practices. It does not dictate a single platform, nor a tightly integrated data infrastructure, but focuses on the specification of lightweight interfaces, standard protocols and standard formats to define a set of minimal requirements and combining existing community standards as much as possible.

CONFERENCE PARTICIPANTS

Dr Myles Axton, Nature Genetics • **Drs Arie Baak**; **Drs Albert Mons**, Phortos Consultants, Euretos • **Dr Jan Willem Boiten**, Dutch Techcentre for Lifesciences (DTL), CTMM-TRaIT • **Professor Barend Mons**, Leiden University Medical Center, DTL, ELIXIR NL • **Dr Niklas Blomberg**, ELIXIR Hub • **Olivier Dumon**; **Dr IJsbrand Jan Aalbersberg**; **Gaby Appleton**, Elsevier • **Professor Carole Goble**, University of Manchester, ELIXIR UK • **Professor Jaap Heringa**, VU University Amsterdam, DTL, ELIXIR NL • **Dr Bengt Persson**, BILS, ELIXIR Sweden • **Dr Thierry Sengstag**, SIB Swiss Institute of Bioinformatics, ELIXIR-CH • **Dr Maurice Bouwhuis**, SURFsara • **Professor Anthony Brookes**, University of Leicester, Gen2PHEN/GWAScentral • **Professor Tim Clark**, Harvard Medical School, Mass. General Hospital, Force11 • **Dr Michel Dumontier**, Stanford University, NCBO, Bio2RDF • **Professor Frank van Harmelen**; **Dr Paul Groth**, VU University Amsterdam, W3C, Open PHACTS • **Dr Rob Hooft**, DTL, Netherlands eScience Centre • **Professor Joost Kok**, Leiden University • **Dr Ruben Kok**, DTL, Netherlands Bioinformatics Centre (NBIC) • **Professor Johan van der Lei**, Erasmus Medical Center, EMIF • **Dr Rene van Schaik**; **Dr Scott Lusher**, Netherlands eScience Center • **Dr Erik van Mulligen**, Erasmus Medical Centre, S&T • **Professor Abel L Packer**, ScieLO, Brazil • **Dr Ted Slater**, YarcDATA • **Dr George Strawn**, National Coordination Office/ NITRD (USA) • **Dr Morris Swertz**, Groningen University Medical Centre, DTL, BBMRI-NL • **Drs Jan Velterop**, Acknowledge • **Dr Mark Wilkinson**, University of Madrid, SADI

CONTACT

Dr Barend Mons
Dutch Techcentre for Life Sciences (DTL)

E barend.mons@dtls.nl

www.dtls.nl