Making Research Data FAIR
A CCC Town Hall on LinkedIn Live

Recorded May 10, 2023

With

- **Debora Drucker, Embrapa**
- **Christine Kirkpatrick, GO FAIR US**
- **Barend Mons, CODATA**
- **Francisca Oladipo, Thomas Adewumi University**
- **Erik Schultes, Go FAIR Foundation**

For podcast release
Monday, June 5, 2023

KENNEALLY: Welcome to a CCC Town Hall – Making Research Data FAIR. I'm Christopher Kenneally, host of CCC's Velocity of Content podcast series.

Last week, the Spanish national government approved a strategy on promoting open science. Spain's ministry of science and innovation said the effort would be implemented over a four-year period, starting with a budget of $26 million US for 2023. One objective is to promote the management of research data using the principles that data should be findable, accessible, interoperable, and reusable. Mandated by funders and governments and implemented at universities and research-intensive organizations worldwide, FAIR data principles ensuring that data is findable, accessible, interoperable, and reusable are expected to drive innovation in science in the years ahead. Proponents say FAIR will accelerate machine readability of research and thereby lift discovery to greater heights.

My CCC Town Hall panel will share best practices for developing research data that is FAIR through culture, training, and technology. We will learn how FAIR data saves lives, saves money, and drives confidence in science on four continents. And our very special guest, George Strawn, a scholar at the US National Academies of Sciences, Engineering, and Medicine who served as CIO with the National Science Foundation from 2003 to 2009, joins our roundtable discussion with reflections on a half-century of internet evolution and the transformative role of FAIR data in the future.

Barend Mons in the Netherlands, you are president of the executive committee of CODATA, the Committee on Data of the International Council of Science. As a molecular biologist, you are responsible for groundbreaking research on malaria parasites. In 2014, Barend, you became involved with creating the FAIR data principles, and today, you are a

leading advocate for their adoption and for data stewardship, ensuring the quality of an organization's data assets. So why have you committed yourself to this ambitious effort?

MONS: Thank you for your question. It's actually enlightened self-interest in a way. If I look back a little bit if I'm allowed, in 1996, when George was still very active in the early internet, I started something called Scientists for Health and Research for Development, SHARED, because I saw that we needed a lot of better connections across languages in Africa. But then I really started to apply it for my own genetic research, and in 2005, I published "Which Gene Did You Mean?," which contains my most hated statement, I think – why bury it first in narrative and then mine it again? And I was already starting to see the value of machine-readable data, seeing that the amount of data we generate is so large nowadays, which is a lot later – we actually double our amount of information in genetics every six to 10 months. So there's no way to work without machines, and that's why I need FAIR data for my own research. But I also became an advocate because I think it can change the face of science and the whole way we do open science fundamentally, which is needed.

KENNEALLY: Christine Kirkpatrick, you lead the San Diego Supercomputer Center's Research Data Services division, and you are head of GO FAIR US. Are data scientists the only ones in the research community who care about FAIR?

KIRKPATRICK: Definitely not. And a whole bunch of people are data scientists who wouldn't call themselves that, so data scientist is already a fraught term. But let me start by saying that I think the answers to many of our societal problems, from climate to human health, you name it, are in data that we already have, and the answers are waiting to be unlocked. This is the reason that I have dedicated so much of my brain space and time to this.

A 2016 Bloomberg survey said that 80% of the time that a data scientist spends with data is finding and cleaning data. That wasn't the first time. Before they called it data science and data mining, we had same things in the academic literature – that people who had spent lots of time in school were spending time basically doing a lot of clerical tasks trying to clean up data. This is why I spend so much time trying to figure out the connection between the FAIR principles and time to science – making sure the data scientists or anyone working with data can get started quicker and spend more of their brain power on actual analysis. So I really think that anyone who analyzes data or relies on the data analysis, which again is anyone on the planet, should care about this topic.

KENNEALLY: Erik Schultes with the GO FAIR Foundation in Leiden, the Netherlands, what kind of data do we want to be FAIR? Are there priorities for certain types of data over others?

SCHULTES: Sure. There's a broad spectrum of different kinds of data that we all care about. I think one thing that the FAIR principles has done a very good job at – when you look at how these 15 one-liners, these FAIR principles, were constructed, it really elevates the idea of metadata to be a first-class citizen along with the data themselves. So I think one answer to your question, Chris, is metadata is sort of a non-negotiable for FAIR. Without FAIR metadata – that is, machine-actionable metadata – even your data assets, no matter what they may be, will not be F, A, I, and R in any automated way. So metadata is a given.

But then after that, we can say there's certainly a lot of legacy data out there. We've already created huge amounts of digital information. We could say, well, yeah, all of that needs to be FAIR. But I think more strategically, we would say maybe that data can be made FAIR at the time where it needs to be accessed again or reused. So that might be one way to kind of prioritize the FAIRification of legacy data.

And then looking forward, given that the doubling time for data may be as short as six months in some domains, there's an exponential increase in data. The FAIRification in those domains – let's say in high-throughput data environments – could have a huge payoff, because even within a couple years, you could have the majority of your data in that FAIR format already. So it's really a case-by-case basis. I think one should be very practical in the approach on that. But we have, I think, emerging strategies and guidelines for the prioritization on that.

KENNEALLY: Francisca Oladipo, joining us from Abuja, Nigeria, welcome. You are professor of computer science and vice-chancellor at Thomas Adewumi University and founder of the Africa University Network. The effort to FAIRify data across Africa was sparked by the COVID-19 pandemic and shaped by lessons learned from Ebola outbreaks. Tell us how those two dramatic healthcare calamities converged in Africa around research data.

OLADIPO: Thank you, Chris, for having me. In the case of Ebola, the data collected from the countries that were affected in west Africa, east, and central Africa – they were taken away from Africa, they never returned, so researchers from Africa and even from the affected countries and health facilities no longer have access to those data. Even if we're able to find them, we're able to have access to them, they're not reusable. They're not interoperable.

In the case of COVID, prior to the founding of our network in Africa, the data from Africa was poorly reported. But now with FAIR, our data has become global. We have a set of principles for researchers in Africa to share their data, for the data to be reusable, to be discoverable by man and machine. We're now able to generate, we're able to curate, and

hold data in residence. And of course, with FAIR, we've been able to establish data ownership in African countries. Many African countries are no longer anonymous, in the sense that we can always tell where the data is, where it was collected, the address of the health facility, for example, or the research institution, and then of course, where the repository is and how it can be found, the principles for the reusability, for the access, and everything around the principles of the management of the data.

So our data now through FAIR are generally discoverable by both man and machines. (inaudible) from researchers in China can actually discover data reposited in a small village in Nigeria or Uganda or Tanzania. Thank you.

KENNEALLY: Debora Drucker in São Paulo, Brazil, at Embrapa Digital Agriculture, one of the Brazilian Agricultural Research Corporation research centers, you hold a degree in forestry and work closely with GO FAIR's agriculture and biodiversity networks in Brazil. Welcome. Tell us – how will FAIR data help you advance the Embrapa mission?

DRUCKER: Embrapa's mission, in short, is to provide knowledge to help sustainability of agriculture, and this is, of course, an interdisciplinary challenge. And we do need data from different disciplines in agricultural sciences and environmental sciences. Very often, we need long-term data to understand the complex processes that are also being affected by climate change. All of that is needed to provide healthy food for everyone.

So we do need data to be FAIR, and we have been working with that even before the concept was generated. It's very important to have the data well organized and reusable, with metadata and everything that the previous speakers already touched on. It's also sometimes really time-consuming and expensive to collect data in certain places, so it's important to make the most use of it in benefit of society. Thank you.

KENNEALLY: Well, thank you, Debora Drucker, in São Paulo, Brazil. *Obrigado*. We will return to the roundtable with the audience questions for the panel later in today's program. You can use the chat function on the LinkedIn Live page and let us know what's on your mind. If you have a question specifically for one of the members of the panel, let us know who you would like it addressed to.

Right now, I want to move around the virtual room and learn more about my guests' work on making research data FAIR – bring back Barend Mons in the Netherlands. Barend Mons, you devised the FAIR acronym. I guess you have a hand with acronyms. You have SHARED, and then you brought in FAIR. Explain what makes the four points of that compass – findable, accessible, interoperable, and reusable – so essential for research data.

MONS: Yeah, thank you. In fact, when we had the meeting – and George was there, Erik, a few of us – we didn't even have the acronym in mind. We were just listing all the problems that we encountered with data and specifically making them machine-actionable. That was the focus of the meeting. And there were all these kinds of principles coming up, and they were shuffled and reshuffled.

At some point, I tried to synchronize them in the process, basically saying most of the data are not even findable, to be honest. They're hidden behind articles as supplementary data. If you're lucky, the link works. Machines still find that a nightmare, even if the article is open access, because maybe on page three, there is a link to supplementary data, and if you're lucky, it's PDF. So findability – and Christine already said 80% is wasted of data science on finding data, making them interoperable, data managing. So findable is of course the first thing. If you don't find data, it all stops.

Then, of course, when you find them, many of the data are not accessible. They are simply behind firewalls, or you don't know the licenses on them. There is no metadata, as Erik said. In pharma, people won't use data, for example, unless they know what the license is. Am I allowed to use the data, and for what purpose?

Then if you are accessing them, many times you spend a lot of time to understand them. They are not interoperable by nature, let alone for machines, even not for people.

And of course, the final letter, R – that's what it's all about. It's all about reusability. And even if they are findable, accessible, and interoperable, the reusability is also a matter of are they fit for my purpose? It's not only like, oh, they're there, so I can reuse them. You need enough metadata to know how big the data are. You don't want to download a petabyte of data. What is the license of the data? Are they fit for my purpose? As Erik said, metadata became more and more important during the first implementation years of FAIR.

But when I just reshuffling everything we came up with after the meeting – because the paper was only published in 2016, and the meeting was in 2014 – this thing popped up to me, because it was simply putting all the problems in synchrony and in order. And then I came up with the acronym, and I knew the moment it came up that it would be a powerful acronym, of course, because you can hardly be against FAIR. It also has downsides. We can talk about that later.

KENNEALLY: Thank you, Barend. In your work with the GO FAIR Foundation, who have you persuaded to join the FAIR cause? Can you maybe drop a few names of researchers and public figures who have endorsed the FAIR principles?

MONS: Well, I remember George was at the meeting, and he said don't make it a new standard. And if you want this to be accepted very quickly, and even people that swim in the net and then want to wriggle out, it's too late. That was very wise advice, because strangely enough – well, maybe in hindsight, not so strangely – the funders and the politicians were quicker to accept this idea than the researchers. So the EC, the European Commission, but also the Dutch funders, the G20, G7 – it went very, very fast, and people started to require FAIR data, because they were convinced that they were wasting their public research money and other money by supporting research that then yielded data that were at best usable and reusable for the people who generated them. But most of our data can be reusable for many years and by many other people, and that puts a much higher quality requirement on them.

Then one of my frustrations has been – and I won't mention big names, because I'm against the silverback culture anyway – the young people are adopting this much faster than the silverbacks. First of all, they are least affected by the problem. They are stuck to their Excel sheets and their fountain pens and so on. Don't even waste time. The next generation will take this very seriously. And I think it's much more telling that this month, we will pass 10,000 citations on the paper. So it spreads over all disciplines in all regions of the world. We just heard a beautiful example from Africa. We have a job, though, internally that at least half of the people that cite the paper clearly didn't read it, because they – I don't think they really make their data FAIR. They just do some hand-waving to satisfy the funders. So there's still a long way to go, but it is spreading all over the world very, very quickly now (multiple conversations; inaudible).

KENNEALLY: Briefly, Barend Mons, I think you give us a sense that there's a real cost to science and to the public for not doing this job. But have you estimated the cost of FAIRifying data? Can you tell us what that might be?

MONS: Well, not myself, but we all know probably that when the European Open Science Cloud was started, part of the exercise in the European Commission was to ask PricewaterhouseCoopers to do an estimate or a study on what is the cost of not having FAIR data. And I think they came up with pretty much an underestimation, but OK. They came up with about between €10 and €11 billion per year only for publicly funded research in Europe. So I think if you multiply that by four for Europe if you include industry and policymaking that is also needing a lot of data to be evidence-based, and then you do that again for the whole world, I think it's fair to say that we will lose about €100 billion at least per year. And we have even talked to pharmaceutical industries that say we alone lose a billion per year on bad data from all the companies that we have acquired. Then for 5% of the research costs that generates the data, we can make them FAIR and do proper data stewardship. That's a pretty good return on investment, I would say. Part of the reproducibility crisis – that is partly caused by non-reusable data.

KENNEALLY: We will get back to that crisis in reproducibility later, but Barend Mons, thank you very much for giving us the background on the origins of FAIR. I want to turn now to Christine Kirkpatrick with GO FAIR US. Christine, welcome again to our discussion. In your efforts to get the science community to adopt FAIR, does it come down to a question of making the FAIR data a matter of usability, user-friendly principles? And how do we do that? What can be done, do you think, to improve the process, improve the technology?

KIRKPATRICK: First, let me say the FAIR principles are aspirational. It's not a specification, as Barend said. It's not necessarily tied to specific technology. So part of the work we do at GO FAIR International and GO FAIR US is to try and take those principles and translate them to practical actions and implementations. But at the moment, not enough is built into tools that are easy to use for researchers and people in industry. So they're left to guess, and they're left to know a lot that's probably more technical than it should be.

I've been in information technology for 30 years, and for those of us who've been on that journey, we remember early word processors. If you wanted to bold text, you had to know the function key. There was no way to look it up. There was no menu. You had to memorize these things. Well, now we all use modern word processors, and you can just click something that looks like a bold symbol for B, and it's much easier.

Same with electronic accessibility. It used to be pretty difficult to make PDFs, web pages, Word documents accessible for people with screen readers, for example. Most people had never heard of the W3C standards, the WCAG standards that help you do some of this. But now, some of these tools are built into Word. You can right-mouse click on a picture, and it'll say what's the alternative text? You put it in there, it's done to standards. Screen readers know how to make that accessible for people that are sight-impaired.

So we're developing tools like this all throughout GO FAIR, but also the whole ecosystem. I myself have been working on a tool called FAIRIST, which helps people step through what they're doing with their data and tries to ask them in sort of a more TurboTax, for those of us in the US who know what that is for filling out our taxes – something that's more user-focused. There's a whole bunch of other wonderful activities like that that are going to make it much easier.

KENNEALLY: In January, Christine, the National Institutes of Health issued a data management and sharing policy to promote the sharing of scientific data, something that was in development for many years. Moving forward, all data associated with NIH grant-supported research must be made publicly available in the NIH's own intramural database. I understand you recently applied for an NIH grant. What did you think of the data-related questions in the application? Have they got it right?

KIRKPATRICK: Well, it's one thing to be an evangelist and to talk about these things, and then it's another to put on your researcher hat and have to fill these things out. I have to say they're asking the right questions. I think it's going to be a game-changer. Because instead of allowing people to opine, which all of us academics are good at doing, and sometimes it lets you dodge the specifics, it asks you what formats and standards will your data be in? Where are you going to deposit your data? Again, I think it's a big step forward for anyone who's NIH-funded, and many of those researchers are funded by different agencies, so they're going to bring those practices across to other places and lift the bar, as you will. And it's driving really important conversations across research teams, and those conversations are happening at the start of the project so that there's a chance for data to be born FAIR.

KENNEALLY: And those conversations you're talking about – I guess we're part of that conversation. You have pointed to me – there's a project at the National Science Foundation, the EarthCube, which is looking to create a kind of integrated data management infrastructure for geosciences. Briefly, would you say that a new data culture is emerging and being driven by FAIR?

KIRKPATRICK: Yes. I know it's an often cited quote, but it's so true from Peter Drucker that culture eats strategy for breakfast. So you do have to drive culture change. It was one of the honors of my career to be the PI of the EarthCube coordination office, which closes at the end of this month as NSF finishes their 10-year project that was an $85 million portfolio of projects to help Earth scientists and what we call cyber-infrastructure providers, people like me, to work together.

Two things I'll highlight – one is a flagship project called DeCODER, which was built on something that was called GeoCODES. It gives you an interface to search across dozens and dozens of academic repositories that are focusing on Earth and the biological sciences now, too. But the interface isn't even the important thing. It's working with the different repositories to adopt standards, like schema.org and Science on Schema, so that their datasets can be found not just by DeCODER, but by the Google Dataset Search.

The second one is that we really wanted to move the needle on reuse of data, and one of the ways that you can drive that is by the exchange of electronic notebooks – Jupyter Notebooks that many of you might have used before. But when we went to create a directory of these, we couldn't find them, and the ones we found, we weren't sure if they were good exemplars or not. So we created, along with the help of the founder of Jupyter, Fernando Pérez, and other important technologists, like Dan Katz at the National Center for Supercomputing Applications, and Kenton McHenry, an annual competition. It's peer-reviewed. Academics get credit for this. And they put up their notebooks. Now, we have

at earthcube.org a directory of geoscience notebooks. So the more that we can put this into practice, have other types of scholarship, we're going to see data that's more FAIR and data that's more reusable.

KENNEALLY: Thank you, Christine Kirkpatrick with GO FAIR US.

I want to turn now to Erik Schultes, also in the Netherlands, with the GO FAIR Foundation. We've just been hearing from Christine, Erik, about the advance and adoption in the US around this EarthCube project. But can you give us a reading on global levels of FAIR data awareness? I'm fond of scales of one to 10. So in this case, one might be no signal detected, and 10 is been there, done that, got the grant to show for it. So where are we in adoption – in awareness, I should say, really, even – of FAIR data?

SCHULTES: Yeah, so I could try to take my own experience and come down to a single number, but I don't think it would be that meaningful. What I can say, Chris, is that the landscape of FAIRness levels is very lumpy and very complicated, and there are vast plains of very low level FAIR awareness, and there are very high peaks and alpine regions where people are doing work that I think technologically speaking could be around for a long time, because it's really cracking a lot of intellectual nuts. So it's a lumpy landscape.

Barend mentioned earlier that the original FAIR paper that was peer-reviewed and enunciated the FAIR principles will reach very shortly 10,000 citations, and one thing we'd like to do, I think in the next year, is begin to evaluate – kind of do a literature search on that to see exactly what are the disciplines and domains that are using FAIR and how they are trying to implement that for their own purposes.

The other way of looking at this is that there's a lot of stakeholders out there. Barend also mentioned these high-level stakeholders, like funding agencies and government ministries, that were really the early adopters of this or actually the early interested parties. I think this is because those groups feel the heat when it comes to big budgets being spent on research, and then later on there's a need for some kind of accountability – where is the data that was generated by that? So those high-level groups are very responsive to FAIR.

And it turns out that the researchers – in my experience, anyways – are often the least well informed about FAIR. In a way, that makes a lot of sense to me, because researchers are already highly trained and highly specialized in their particular area of research. To then go out and to begin thinking deeply about data and machine actionability – yeah, that's a whole other career path that we would really be asking a lot for researchers to respond to that.

So you mentioned – or you asked Christine some questions about the technology. I think some of the technological advancements that make FAIR more easy and more automatic – that's how you will pick up the awareness levels or the participation of researchers. But certainly from the high levels, there's a great deal of awareness and interest in implementing FAIR.

KENNEALLY: Something I need to help you to ask me understand, Erik, is about this notion of data getting lost, or even the danger of data disappearing, which sounds oxymoronical, because there's so much data already, and so much of it is produced every single day. So how can we keep producing more and more data and then losing it or seeing it disappear?

SCHULTES: I think it's probably safe to say – I mentioned earlier that the doubling time for data could be in some domains as short as six months. That means that of all the data we have today, six months from now, it's going to be doubled. A year from now, that's going to be quadrupled. So it's an enormous increase in data production. If that data is not made accessible online – let's say on the internet – and if there isn't sufficient metadata that can allow search engines or machine agents to locate it in some meaningful way, then indeed that data can be lost in plain sight, right? It'll be right under your nose, but you'll never see it. And the problem gets compounded every six months or every year – every doubling factor – because the rate of data creation is so enormous.

By the way, the data is complex, as well. There's lots of different domains, lots of different data types. So if we're asking a human to manage the F, A, I, R, it's just simply getting beyond human comprehension and human capability. The machine has to help. Otherwise, in fact, the data will not be – or will no longer be F, A, I, and R for anyone.

KENNEALLY: Erik Schultes in the Netherlands, thank you very much with the GO FAIR Foundation.

I want to turn now to Francisca Oladipo, who is in Abuja, Nigeria, with the Thomas Adewumi University and the founder of the Africa University Network. We were chatting earlier in preparation for our program, Francisca, and you were telling me about the effort to collect all of this data and to make it FAIR, and you were comparing and contrasting the work you did in Nigeria and in Ethiopia. So tell us about the special challenges that you faced in those two very disparate countries.

OLADIPO: Thank you. For Ethiopia, it was kind of easier, because there was more like a central authority. That was prior to the conflict in Tigray. So we needed just approval or the buy-in from the ministry of health and then the universities.

But in the case of Nigeria, there are 36 states and then there is the federal capital, making 37. So we had to convince 37 different entities with different laws. And then there are also different levels of health facilities and institutions. Some are owned by the states. Some are owned by the federal government.

So it was kind of easier in Ethiopia, and that was why we were able to move faster in Ethiopia than in Nigeria, because even after getting the buy-in from a state authority, there is also the need to get a buy-in from the federal authority and then a primary authority. When we're talking about getting ethical approval for the work, we had to go to different parts of the country and different ministries of health, for example, different institutions, even though there is one body overseeing tertiary education in Nigeria. But each university still has some level of autonomy.

So that is one big challenge working in the two countries. Ethiopia was kind of straightforward prior to the conflict, but Nigeria was quite hectic, because we had to replicate things. It was circular. It was repetitive. But that is the law and that is the way the structure – the way the research community, the research approval in the country – that's the way it's set up.

KENNEALLY: Francisca, you managed to make it work, so I want to hear your pitch, whether it was to those national governments or to the various state governments in Nigeria and so forth. What's the pitch? What do you have to say to convince them that sharing data, collecting data according to the FAIR principles, is so important?

OLADIPO: Already we had to remind them or to make them understand that they all know the difficulty in collecting data. What they don't know about is that there is a principle that could make things easier. So we had to go African to remind them of the problem that we face, the problem of data collection.

And then as a computer scientist, we had to use case studies. Our machine learning and AI models are trained with foreign datasets. Many data collected by Nigerian researchers are not FAIR. They're not available. They're not findable, they're not accessible, they're not interoperable and reusable. It creates a kind of repetitive process for the researchers. You see an average PhD candidate in Europe will spend between three to five years. In Africa, if you're spending eight years, nobody sees it as anything. But out of the eight years being spent, the first four years or thereabouts will be dedicated to data collection. So we have to present those different kinds of case studies.

And then in the health facilities and the ministries of health, there was a pandemic. So we presented an opportunity for Nigeria to join other African countries in the fight against the COVID-19 pandemic. How? We cannot be at the forefront of vaccines, detection –

there's a lot of misinformation going on, so we cannot even really compete globally in trying to use the communication science to fight against the pandemic. There is a new kind of science. So we preached – we presented to them the new kind of science, the science of FAIR data, that will assist the frontline healthcare workers with data for them to do analytics, understand progression. And of course, we also brought up the idea of globalization – that now, Nigerian data, African data, can be discoverable outside of the shores of Africa.

So we had to use different approaches. Actually, it depends on who we are talking to. We had to kind of look at the different circumstances. When it is the federal government, we looked at globalization. When it is the state, we looked at inclusion within a federal system.

And then, of course, we had help. Globally, we had our partners in Europe. We had the funders. So all those big names also helped. When we hear, oh, a group of Nigerian researchers are proposing to work with researchers in Uganda, in Kenya, in Tanzania – you know, oh, Nigeria is going to be part of a global thing. So those were some of the pitch topics or pitch areas that we used to ensure the buy-in.

Then we had African students through a PhD, and they needed the African data pipeline. Through them, we were able to convince the heads of the institutions, the heads of the health facilities, and then the ministries of health to buy into the VODAN Africa and then FAIR in Africa projects.

KENNEALLY: Francisca Oladipo in Abuja, Nigeria, thank you for that background. It is a very persuasive argument, indeed.

I want to turn finally in our walk around the virtual room to Debora Drucker in São Paulo. Debora, welcome back. We were speaking earlier about Embrapa and the research that you are doing on a wide range of issues facing agriculture in the tropics. An objective is to unite all the disciplines that are involved, which can range from soil conditions and plant diseases to sustainable farming methods and so on, under a single data tent. Tell us – how far along you are in this quest to create that single data tent, and what role are you expecting FAIR data to play?

DRUCKER: Yeah, so it's much more complex than it might sound. Actually, what we are supporting – so Embrapa Digital Agriculture, which is the research center I work for, is one of the 47 research centers in Embrapa, and we're hosting Embrapa's datacenter. We are a 50-year-old organization, so there is a lot of data that has been generated all over the years. In the previous decades, many information systems were built, but then they

became obsolete, and it's very hard to tailor a specific system for a specific project or discipline or sub-discipline.

So we are hosting the institutional repository in order to sustain and preserve this data for a long term. And on that, we can build tools to really mobilize this data and analyze it and help really build in solutions for the agriculture in the field. We also host an API platform that reads the data and provides solutions on time and predictions on the productivity or related to climate or hazards and things like that, as an example.

We have more than 8,000 employees, and it's lots of work. So it's very hard to build that one-size-fits all solution, but with the repository, it's already one step. But there are many other initiatives. We are working on the data that is legacy data to make it FAIR, and also for the new data and future projects, we have a new data policy that's very much based on the FAIR principles. And we have data management plans and things like that to deal with this really huge challenge. We also connect with other institutions. We are connected with the GO FAIR office in Brazil, and we are engaging other organizations to mobilize the data – the data producers, the data consumers. It's a lot of work, but it's also fun.

KENNEALLY: It sounds like it is a lot of work, and your situation kind of echoes the description that Francisca Oladipo gave us regarding Nigeria and the various states there and the states in Brazil. You mentioned Embrapa is 50 years old – 50 years old this year, I believe, is the anniversary. A half-century of data, a half-century of research. Tell us a little bit more briefly about the data culture at Embrapa.

DRUCKER: For sure. In the beginning, you used to have a data – quantitative department where the researchers would send their data in floppy disks for the analysis. And then with internet and everything, the different research centers and different groups have been building their own paths, and that will vary with the different challenges of the disciplines and everything.

It's more of what was already said. People don't like things that will – if they need to write the metadata, that's something people don't like doing. So the most machines can do for us, the better. There is also different approaches on sharing data. In some cases, we cannot share it, because there are companies that are connected, or if there are some personal information involved. But we do need to make it FAIR anyway and to host it and to preserve and describe and everything.

So it is a challenge, because it's different from what most researchers, especially the ones that are here for 50 years, are used to. But many also see the value, and it's good to know that the data is published, and if people will use it, people will cite you, and things like that. So the culture is changing. But I can say that most people in Embrapa and elsewhere

in science want to see science progressing, so it's a matter of communication and engagement. I think it's still a long way, but people usually see the value, so it's good to see things evolving.

KENNEALLY: Debora Drucker with Embrapa in São Paulo, Brazil, thank you so much.

The discussion so far has really sharpened the focus for me on why FAIR data is fundamental to the future of the internet as a welcoming place for innovation and discovery and research. I think we can see paths for improvement, and we can begin to understand how FAIR data will lead to breakthrough interoperability across research domains.

So we'll bring back the panel. We're going to have a discussion with you. But I first want to welcome a very special guest and continue the conversation with him. It's time to say greetings to George Strawn, a scholar at the US National Academy of Sciences, Engineering, and Medicine. George Strawn served as CIO with the National Science Foundation from 2003 to 2009. He joins us today and our roundtable discussion with reflections on a half-century of internet evolution and the transformative role of FAIR data in the future. George Strawn, thank you so much for joining us today on this CCC Town Hall.

STRAWN: Glad to be here.

KENNEALLY: Now, you identify three steps in the evolution of computing. The first step came after World War II with the ENIAC and UNIVAC, the first programmable digital computers. They were enormous standalone machines. Then came the internet working era that began in the 1970s, when you were starting your own career as a computer scientist at the University of Iowa. What new era in computing will FAIR data bring?

STRAWN: Thank you, Chris. I think that's the right head-in-the-clouds question. As you say, the first wonderful step was computers themselves. The second, equally important step has been the internet age of interconnected computers. I think FAIR is ushering in the third phase of information technology, which is interoperable data as well as interconnected computers.

Looking back over the past couple decades, I think I can highlight three examples that already prove the value of this third phase of FAIR data. The Human Genome Project discovered eventually 20,000 genes or so of the human genome, and all of those gene discoveries were placed in one FAIR database. What if we had had 20,000 independent researchers each contributing their own data in a non-compatible format, and then somebody had to try to put that data together afterwards? We didn't call it FAIR back

then, but that's what it was. The human genome database was a FAIR database that accelerated greatly the conclusion of the Human Genome Project.

In a different discipline, although beginning about the same time, astronomers worked to get all telescopes to have standard output metadata – that is, FAIR metadata – that in effect allowed all pictures of a given part of the sky to be overlaid, looking for similarities and differences and whatever. This resulted in astronomy being now called the virtual observatory. According to my understanding, it has revolutionized the science of astronomy being able to have FAIR from the instrument data.

Third, and most recent, is the solving of the 50-year-old protein folding problem, which is an example of machine learning. Of course, we now know that machine learning is in the news as a – should we worry about futures? Well, perhaps we should, and perhaps that's a discussion for another day. But certainly the benefits are at least as large as the dangers.

Barend has pointed out from time to time that FAIR is an overloaded acronym. It can also stand for fully AI-ready data. In order for the training of machine learning algorithms to occur, huge – the huger, the better – databases of data need to be analyzed to do the training to make the algorithms work. Well, that's what happened with the protein folding problem. People had been working for 50 years on finding a computing way to solve that problem rather than the expensive and time-consuming X-ray crystallography approach. A couple years ago, it was designated as the science breakthrough of the year and is revolutionizing the proteomic part of biology.

If we think of the revolution in proteomics, in genomics, and in astronomy, these three breakthroughs have happened because of FAIR data, even though we didn't call it FAIR at that time. I'm convinced there are untold additional breakthroughs of those magnitudes that will be available with the progress that my colleagues have just described as we are moving forward to fully implement this third phase of information technology – interoperable data.

KENNEALLY: George Strawn, thank you for that, and thank you for adding to our definition of FAIR. I love fully AI-ready. That seems perfect and brings us right up to date from the internet birth to today, because AI is so much on people's minds. We've had some questions about the use of data in the training of things like ChatGPT, which are, of course, large language models. As you say, George, it's a topic for another town hall. In fact, we did a town hall on ChatGPT about a month ago.

But I want to ask an open question to the panel here about the relationship of that other piece of AI – George, you raised it – the concerns people have about AI, the power of AI, the potential for it being not all for the good. Does FAIR data play a role in developing

greater trust in science? I see Christine nodding. I'll let you go first, Christine. What about that? What about FAIR as a way to help the public feel more confident in research and to trust the science that they're told about?

KIRKPATRICK: Absolutely this is something that we're working on and that's very needed. Several of us on the panel are part of something called the FAIR Digital Objects Forum, which is working on what we think will be the next architecture of information for the internet.

If you remember back – those of us who were there when we had our first websites, you had no such thing as a secure connection. You didn't see a lock in your browser. There was no such thing as that. Then over time, we had certificate-granting authorities, and people started to look, and they said I'm not going to put my credit card information in unless I see a lock up there on the top left in the browser. We're going to develop the same things, and we're going to need everyone to learn how to look for whatever the equivalent of the lock will be.

Right now, there's a whole bunch of information coming out of AI that's just garbage. I also wanted to highlight another project. It's called FARR. It's FAIR in Machine Learning, AI Reproducibility, and AI Readiness. This is a consortium of people who are looking at the intersection of these two things. If that interests you, it's farr-rcn.org, and we'd love to get your thoughts, especially people who tuned in and have more to say.

KENNEALLY: Thank you, Christine. It's interesting – not surprising to find you involved in that.

MONS: Can I say something, Chris?

KENNEALLY: Absolutely, Barend. Please. Thank you.

MONS: Because I'm just back from a world forum of Frontiers, the publisher, and we had a very good keynote of Yuval Noah Harari on the future of AI and of course the dangers of ChatGPT-type large language models. I actually have a whole schema worked out with a small group to offer to the large language model providers – (inaudible) publications which are FAIR Digital Objects, as Christine said, with full provenance, because it all comes down to trust. Which source do I trust? Who do I trust, as Christine already said? These (inaudible) publications have complete provenance.

Let me give you one example on the notorious thing that hydroxychloroquine would treat COVID. That was published in *The Lancet.* So these language models will just pick it up and tell you it's a good idea. If you are able to feed them with machine-readable

provenance information, they would find out that now about 500 ORCIDs – people, scientists – don't believe this anymore, and there's probably two people in the world that still believe it. We won't mention names.

So the idea is really if these large language models, which are potentially powerful – they're just word calculators, of course – they actually could be very important in the future if they cough up scientifically validated information, and they tell you if you only believe a subset of information, even a subset of scientists, in which bubble you are, so that you can also leave that bubble again and say, what if I look at this thing from a totally different perspective? That is where FAIR will play an enormously important role in the future.

KENNEALLY: Right. Thank you. Barend, you talk about provenance and the reliability. I want to end our discussion. We have come through the hour. It's been fascinating. I want to end with Francisca, though, because you've been talking about the importance of including African-born data – I'll call it that – in the global data network. When we speak about these large language models, questions are raised about whether they are diverse enough in the sources that they probe and whether they reflect biases and so forth. So it sounds to me like you would be also a part of really wanting to see FAIRified data from Africa included in the training that happens for these LLMs. Is that true?

OLADIPO: Yes, that's true. That's correct. It's actually part of the goal. We are looking at the globalization of data collected from Africa. So yes, it's part of the goal, and we are encouraging researchers who are working in the areas of LLM and those who are working in machine learning, in AI, who are working generally in natural language processing, to ensure that they concentrate also on low-resource languages, languages that are spoken by millions, but there are very low resources for them, to ensure that at the end of the research – to ensure that not only the data, but all the research artifacts are FAIRified and then they're discoverable by these language models.

KENNEALLY: Well, thank you, Francisca. British economist and Nobel laureate Ronald Coase once said that if you torture your data long enough, it will tell you anything. Today's CCC Town Hall has taught us the value in treating our data well, because then it will return the favor.

I want to thank my guests – Barend Mons, president of the executive committee of CODATA, the Committee on Data of the International Council of Science and with the GO FAIR Foundation in Leiden, the Netherlands. Thank you so much, Barend. Also, Christine Kirkpatrick, head of GO FAIR US, based in San Diego, Erik Schultes with the GO FAIR Foundation in Leiden as well, Francisca Oladipo joining us from Abuja, Nigeria, professor of computer science and vice-chancellor at the Thomas Adewumi University and

founder of the Africa University Network, and Debora Drucker at Embrapa Digital Agriculture, São Paulo. Thank you all very much indeed.

Rob Simon of Burst Marketing is our director. Thanks as well to my CCC colleagues Joanna Murphy Scott, Amanda Ribeiro, Hayley Sund, and Molly Tainter. Stay informed on the latest developments in publishing and research by subscribing to CCC's Velocity of Content blog and podcast. CCC's interactive visual report, The State of Scholarly Metadata: 2023, is online at stateofmetadata.com. We invite you to have a look, interact with it, and share with us the problems and the challenges that you have around metadata. Again, that's at stateofmetadata.com.

I'm Christopher Kenneally for CCC. Thanks for joining us. Goodbye for now.

END OF FILE