Entrevista com Barend Mons

An interview with Barend Mons

Patricia Henning*

Luana Sales**

AN INTERVIEW WITH BAREND MONS

Barend Mons is a molecular biologist and, since 2012, he has been a professor of BioSemantics in the Department of Human Genetics at the Leiden University Medical Center (LUMC) in the Netherlands. In 2015, he was chair of the High-Level Expert Group on the European Open Science Cloud (EOSC). Since 2017, he has been heading the GO FAIR initiative's Dutch International Support and Coordination office. In 2018, he was elected president of the Committee on Data for Science and Technology (CODATA) of the International Science Council for a four-year term through 2022. He is also a member of the Netherlands Academy of Technology and Innovation (AcTI), and he is a representative of the Board on Research Data and Information (BRDI) of the National Academies of Science, Engineering, and Medicine in the USA.

Interviewer: Professor Barend, at first, we would like to say that it is a pleasure to be here at the GO FAIR Initiative in Leiden, and we want to thank you for the opportunity to conduct this interview.

* Doutora em Informação e Comunicação em Saúde pelo Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT / FIOCRUZ). Professora Associada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO). Endereço: Avenida Pasteur, 296, Urca - Brasil. E-mail: henningpatricia@gmail.com

** Doutora em Ciência da Informação pelo Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT). Professora do Programa de Pós-Graduação em Ciência da Informação do Ibict. Endereço: Rua Lauro Muller, 455, 4° andar, sala 408, CEP: 22.290-160. Telefone: (21) 97112-7411 / (21) 3873-9450. E-mail: luanasales@ibict.br

How do you see the recommendations of the High-Level Expert Group (HLEG) regarding the direction that is being taken by the European Open Science Cloud (EOSC)?

Barend Mons: Thank you very much, as well, for taking the time to interview me. The recommendations we made in the first HLEG report were largely taken on board by the European Commission. Several follow-up groups were installed to address specific subtopics, such as next-generation reward systems and turning FAIR into practice. Probably one of the most 'controversial' recommendations we made is that we need to train 500,000 data stewards in Europe for the decades to come. We came to that rough estimate by assuming that in modern, data driven research, for every twenty data producers, we need one [full-time, professional] data steward.

Also, there is still considerable discussion on our recommendation that the EOSC should be conceived—and governed and funded—as the European Union's contribution to a global internet of FAIR data and services. Many people seem to think that we can solve the current bottlenecks in science with incremental improvements to the current e-infrastructures, but we strongly believed at the time—and still do—that we have to take some drastic measures. After all, as Einstein already said, 'We cannot use the same thinking to solve our problems that we used to create them'…

When did you begin to worry about data stewardship?

Barend Mons: That is difficult to pinpoint precisely, but it is probably fair to say that in 2005[1], my article 'Which gene did you mean', which starts with the claim that computational biology needs computer-readable information records, was the first formal 'outcry' about the horrible situation we have created for machines [that] are expected to assist us.

All the information I needed as a biologist (when trying to discover complex patterns) was in text. My most hated quote is also in that article. 'Text mining? … why bury it first and then mine it again' in tables or figures—or in relatively obscure (and widely variable and exotic) databases and formats? It was no coincidence that the article was written at the occasion of the 20th anniversary of SwissProt (UniProt now), which was one of the few well-curated databases at the time. Obviously, the period between that article and the emergence of the FAIR principles (almost a decade) was a long and rocky path.

It is important to emphasise here that I have always made a strict distinction between the terms 'data management' and 'data stewardship', where the latter very explicitly includes the challenge to keep the resulting data (and other research outputs, such as workflows and software) available for others to reuse, for prolonged periods of time (way beyond the end of the project that created the artefacts).

[1] https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-142

And what led you to get involved with this subject?

Barend Mons: In my fifteen years of malaria research after my PhD, I became more and more aware of the intrinsic and long-term value of data; I was especially triggered by the lack of access to the data, software, and information from my colleagues in developing countries, with whom I intensively collaborated at that time. That made me acutely aware of a lack of data stewardship skills—and a lack of a sense of responsibility from most researchers for their data. Once they had published their 'high impact' paper, they largely forgot about the data, and they certainly didn't make any effort to make the data reusable for others or publish them in unambiguous, machine-actionable formats; that was—and in many cases still is—completely out of scope. Now that we've moved so rapidly from a data-sparse science paradigm to science (and a society) that is totally overrun by data-driven science, decision-making, and machine learning, even the most senior scientists must wake up to this new era—and that means FAIR (machine-actionable) data and data stewardship suddenly take centre stage.

So, in a way, we could say that 'experiencing first-hand how it felt to be cut off from mainstream science output' led to my switch from active biomedical research to my current focus on data stewardship—in order to make all science more equitable, effective, and reproducible. I believe putting my experience and energy toward better data in science could potentially save even more lives than finally discovering the so-far-elusive malaria vaccine.

Interviewer: We know that you have been present for every moment of the FAIR principles' creation, from the 2014 workshop at the Lorentz Center in Leiden (when several stakeholders met to discuss infrastructure improvements to support scientific-data reuse), until 2016 (the moment of their official publication) as the corresponding author of the article 'The FAIR Guiding Principles for scientific data management and stewardship', which was published by *Nature - Scientific Data*. We would like to hear a bit about your experience of having participated in these two moments, which were so important to the creation of the FAIR principles.

Barend Mons: That's a funny story. When we organised the Lorentz workshop, we chose the title 'Jointly designing a data FAIRport'. At that time, it was just a wordplay on 'AIRport', because we proposed a sort-of data infrastructure that is currently closer to the concept of the [European] Open Science Cloud. Gradually (after days of intensive discussions), the workshop converged on 'machine actionability' and a sort of 'internet for machines' [paradigm], with the 'hourglass model' of the current internet in mind.

After the meeting, we ended up with a whole list of guiding principles to drive such a 'machine-friendly' internet of data and services. Only after several reshufflings—and after Mark Wilkinson, Mark Thompson, and Michel Dumontier re-addressed the principles once more at a hackathon in Japan (and most likely with the wordplay still resounding in my head)—did I sort the basic principles along the letters: F for Findable, A for accessible, and I for Interoperable—all clearly with the ultimate goal of making them R (Reusable).

One of the participants—whom I had best not mention by name here—even stated that 'sorting them this way and making up this catchy acronym made up for all the nonsense I said during the meeting itself'... That person was at least right in that the acronym took the world 'by storm'—although it is also widely abused for anything that is even vaguely findable, open, accessible, and 'thus' reusable, which, as you

know, is not exactly what we meant or mean. I currently summarise the first line goal of FAIR as 'The machine knows what we mean'.....

What benefits do you believe the FAIR principles can bring to science as a whole?

Barend Mons: Another funny story: In the USA, recently there was (interim) advice to the NIH that all data should be made 'AI-ready'. Those who know how sceptical I am about the hype-term AI can immediately imagine the wordplay:[2] If we want machine learning (and, maybe in the future, 'real' AI) to work effectively, we had better make all data 'machine readable' = 'Fully AI-Ready' = 'FAIR'. There is no escaping the acronym…

But seriously, for any machine learning and analytics algorithm, the substrate is a form of data; the more 'machine-ready' these data are—with clear accessibility criteria, provenance, licensing, etc., and with all of it machine readable—the more efficiently computers will discover complex patterns for us—patterns that are way beyond human processing capacity—and the more discoveries we will be able to make for the betterment of society. So, as George Strawn (one of the pioneers of the current Internet) states, 'This will cause a real paradigm shift in science'.

In your book *Data Stewardship for Open Science: Implementing FAIR Principles*, you give a general overview of the use, best practices, and applications of the FAIR principles. Could you tell us a little bit about your book?

Barend Mons: When I was asked to write the book on data stewardship, I first laughed, as this field is moving so fast that any book would be outdated at print. However, the publisher, Taylor & Francis, was 'in' on what I consider a very interesting and innovative way to look at a 'book'. First of all, the book is only printed in black and white (which keeps the printed copy as cheap as possible), but it contains a code to access the full-colour e-Book. More importantly, we agreed with the publisher that the 'practical' pages of the book would be made available in the 'Data Stewardship Wizard' (then under development by Robert Pergl's team in Prague) in the scope of ELIXIR.

This Wizard is a pretty amazing tool; it is fully open source, and it allows the user to upload a 'knowledge model', which will then generate a Wizard instance (for instance, in the form of a 'questionnaire') that in turn (after being filling out) leads to machine-readable output. The original knowledge model (on which the first Wizard instance was created) was developed by Rob Hooft in the Netherlands (DTL/ELIXR-NL)[3], and I used the same knowledge model to structure my book. This resulted in a 'practical' page of the book being freely available in the Wizard to explain any of the created questions using the same knowledge model.

This is where the dynamics become interesting: new nodes (such as questions, issues, or challenges) will be added to knowledge models all the time when the field of data

[2] Actually, it was my brother Albert who came up with this immediately.

[3] https://www.dtls.nl/elixir-nl/

stewardship starts to thrive. This can lead to comments in the Wizard on existing pages, but it also prompts the creation of new pages, and we can regularly update the e-book as a collection (for those who want to read or use it for educational purposes) or as new editions of the printed copy (over longer intervals). Meanwhile, the book has been translated into Chinese, and now it can also be made available in the Chinese version of the Wizard. … [How many] other languages may follow? …

I hope that in this way, the 'book'—or as I see it, the growing collection of advice pages on good data stewardship—can become a growing and dynamic resource for data stewards, and they do not *need* to buy the book itself, if they feel the Wizard is sufficient for their purposes. This is also a nice compromise between the traditional 'monograph', which cannot be provided in open access unless a generous sponsor is found, and a fully open-access and open-source environment in which the substance of the book is accessible to all.

In your point of view, what are the main motivations that lead researchers to share their research data and to publish it openly?

Barend Mons: Currently, very few—if not zero. Today's researchers are almost exclusively judged and 'ranked' on perverse parameters (such as journal impact factor) and derivative parameters (such as H-factor). These are, of course, archaic measures that drive people back into the 'journal-publishing age', which is a nightmare—and a desert for machines.

That is also why I consider the 'Open Access' (OA) movement (as long as it is only focussing on OA articles) as an already-done and now-rather-trivial exercise. First of all, it does not help machines—although professional text miners love to show how well they can recover stuff that was buried in these articles and put it back into structured and unambiguous formats; then they publish hundreds of articles about the results. … Secondly, OA is 'just' another business model, which now puts the bill at the producer of the science, rather than at the re-user—which is, first of all, not entirely fair (especially when expensive and large data sets come into play), and it is not much better for colleagues in developing countries either: 'What is worse, not being able to read, or not being able to publish'?

We simply have to accept that publishing research outputs in reusable formats and keeping these data available for decades is (very) expensive, and that 'good data stewardship' implies that we budget for that properly (as an expected part of research costs). We produce data, results, and conclusions largely on tax-payers' money, and we have a moral obligation to make the output as accessible and reusable as possible. For that, of course, it needs to be findable and interoperable first.

Unless we apply next-generation evaluation criteria, then (for scientists who respect the first-class citizenship of data) we will inhibit open science for decades to come. Data and other non-textual research outputs should therefore be citable—but also measurable and 'rewarded'. I would personally not hire any scientist [today] who is solely driven by high-impact papers and would not be dealing responsibly with the underlying research outputs. I believe the universities—as well as the funders—have a very important role here: to request—as well as reward—proper, FAIR-compliant data stewardship as an absolute prerequisite to receiving research funding and tenure.

As a researcher and biologist, have you had experience sharing your research data in other surveys?

Barend Mons: Strangely enough, the shift from being a 'data producer' (although I never used very large data) towards being a 'general advocate for FAIR data' (and assisting other people in making their data FAIR) was so fast that I have never shared much of the data I generated during the first phase of my career. However, I have many examples of scientific questions that would either be unanswerable without FAIR and shared data or that would take weeks [to answer] with many people; now, these would only take two minutes. So I think we have ample and rapidly growing evidence in many scientific disciplines (not only life sciences) that reusing data—although I like the term 'data visiting' or 'distributed learning over FAIR data' much better than 'data sharing'—is the only way to leverage the enormous potential of our new assistants: our computers.

Interviewer: In 2017, you became the creator and leader of the Global Open FAIR (GO FAIR) initiative, and you have recently been elected president of CODATA. How have you seen the reception of GO FAIR within the international scientific community, and what are the plans of this initiative for the future?

Barend Mons: GO FAIR apparently answered a 'smouldering need', because it exploded into more than thirty implementation networks, which span large parts of the globe and many disciplines. However, as you know, I do not see GO FAIR as a goal in itself, much less as 'the next self-service kingdom'. The goal is to kick-start the internet of FAIR data and services—and to serve the domain-expert communities that want to create speed with its coordination capacity, meeting options, and convergence tooling.

So far, Brazil is the only country in Latin America that is participating in GO FAIR. What do you think about Brazil's participation?

Barend Mons: I am very proud that, although GO FAIR originated in Europe, there are now activities in many other regions, including, indeed, Latin America (where I hope Brazil will take a leading role in helping other countries benefit from the approach), but also in the USA, Asia, and Africa. It is still very early days, but I hope that the GO FAIR 'function' will get a sustainable place in future research infrastructures, such as the European Open Science Cloud and its sister initiatives in other regions.

Please do not treat GO FAIR Brazil as a goal in itself, but as a means to make Brazilian (and Latin American) policymakers, funders, and researchers aware of the need to converge on FAIR data and services and to embed the 'function' of GO FAIR in future research, innovation policies, and infrastructures in Brazil and the region—also ensuring that they 'seamlessly work' with other regional data and service infrastructures.

As president of CODATA, could you tell us the strategic plans you have for your management term?

Barend Mons: In addition to continuing the important activities of CODATA in complex science systems, data policies, and education, we have launched a strategic plan that involves other international players, such as Research Data Alliance (RDA) and GO FAIR in fulfilling a sustainable role at the international level to collectively

detect, document, recommend, and implement good practices leading to FAIR data and open science. CODATA, as a formal committee of the International Science Council, has a natural role to co-develop a pipeline to allow these good practices to be finally endorsed and formally recommended by its parent organisation, UNESCO, and others.

What advice or recommendations do you have for the Brazilian scientific community?

Barend Mons: Brazil, *vamos* FAIR!