Inria and EOSC: some insights from the French national research institute in digital science and technology

Jean-Frédéric Gerbeau*†[1]

[1]Centre de Recherche INRIA de Paris Laboratoire Jacques-Louis Lions, France. – Institut National de Recherche en Informatique et en Automatique – France

Abstract

Good morning,
Thank you for giving me the opportunity to share a few thoughts with you from Inria.
As probably not everyone in the audience is familiar with Inria, I'd like to start by saying a few words about my institute. I will then briefly present some initiatives to which we have contributed in the field of open science, as well as our participation in EOSC. Finally, I will share some thoughts on AI for science.

Inria is the French national research institute for digital science and technology. In the French academic landscape, we have the particularity of being under the dual authority of the Ministry of Higher Education and Research and the Ministry of the Economy. Inria's historical motto is "scientific excellence and transfer". We have more than 200 research teams, most of them being joint with the universities and gathering about 4,000 scientists, which cover most of the fields of computer science and applied mathematics.
Inria has a very strong software culture. At Inria, software development is closely linked to research. Inria has supported a number of pioneering projects at various time in its history, for example, by supporting the European node of the W3C consortium in the '90s for the open standards of web technologies, or more recently, by supporting the Software Heritage Project that has been presented yesterday, and I will come back to this later.

Since December 2023, and a decision by President Emmanuel Macron, Inria has had a new responsibility in France: in addition to being a research institute, Inria is now also a National Program Agency for Digital Science and Technology.

Inria has a long-standing activity in the Open Science ecosystem. Open publication archives have been recognised for over 20 years as an essential tool for open science. Inria has been heavily involved in the development of the HAL open archive, alongside CNRS and INRAE, within the CCSD, as we have just seen in Nathalie Fargier's talk. Today, about 92% of Inria's publications are available in full text on HAL. Besides publications, the sharing of research data is also a major pillar for Open Science. However, it is now generally agreed that an article describing a method and a set of data is often insufficient for reproducibility if it is not supplemented by the software used to obtain the results. These three pillars –

*Speaker
†Corresponding author:

publications, data and software – are now fully recognised in the French National Plan for Open Science of the Government, and of course, within EOSC. With this in mind, we see Software Heritage as an important tool for Open Science. Software Heritage, as we saw yesterday, is a unique international infrastructure for archiving software source code. Its development began in 2015, under the leadership of Roberto Di Cosmo, and has been continuously supported by Inria ever since. To take another example supported by Inria in a completely different field, I can also mention the France Life Imaging, national action for the management and processing of in vivo imaging data, which was presented yesterday by Camille Maumet in a poster. This software environment is used in particular to carry out challenges for comparison of image processing tools, the results of which are then shared via the Shanoir portal with the scientific community. Through these few examples, we see that, to a large extent, it is digital technology that makes open science possible. So we believe that computer scientists have therefore a special role to play in this. That's why we decided to get involved in the EOSC initiative right from the start, in particular with Jean-François Abramatic, an Inria emeritus researcher, Jean-Yves Berthou, Laurent Romary and others.
Within the EOSC Association, Inria is the French-Mandated Organisation. As such, we work closely with the other French Members. Last year, we decided to hire a full-time Project Manager, Victoria Dominguez Del Angel, entirely dedicated to EOSC. I would like to congratulate her for co-organizing this event today, hand in hand with Volker Beckmann.

Digital technology is making a major contribution to the way in which science is practised today. Let me just mention the Virtual Research Environments or 'collaboratories'; the remote control of research workflows, the digital twin paradigm which establishes a tight loop between a model and reality, etc. All this is based on digital technology. I refer you to the excellent 'Opinion paper on advanced digitalisation of research' published in May 2024 by the EOSC Steering Board expert group. In what follows, I will concentrate on a very specific digital technology: Artificial Intelligence, which is deeply changing the way we do science.

For many years now, machine learning – in other words, data-based models – has been combined with physical models in the field of scientific computing. The goal can be, for example, to speed up calculations by several orders of magnitude, or to literally discover new equations by "learning" differential operators sensors, or to find closure laws to capture sub-grid phenomena that we would otherwise be unable to simulate. A striking example is provided by the field of weather forecasting, where this type of approach, combining a physics-based and AI data-based models, represents a real breakthrough. You certainly have seen the spectacular results obtained by Google or Huawei a few months ago. I can testify that public research is not out of a game. At Inria for example, Claire Monteleoni's group has recently obtained some very promising results in weather forecasting using AI, comparable to those of big tech companies.
Let me give you another example to illustrate that AI is really changing science. In Lyon, an Inria researcher, Omar Fawzi, has collaborated with a DeepMind team to develop an AI system, called FunSearch, which enables a large language model (LLM) to produce new mathematical knowledge. The approach is iterative: an LLM generates multiple algorithm source codes, submits them to a program that automatically evaluates them, rejects incorrect answers and returns correct answers to the LLM for improvement. The system has been tested on difficult combinatorial problems and, after millions of iterations, has uncovered original solutions, some of which are better than previously known results. I don't think it's an exaggeration to say that this is a new paradigm for scientific discovery.
I have taken the time to detail these examples to share with you my conviction that, if AI is changing the way we do science, it is also changing the way we do open science. Meeting the 'FAIR' criteria for data is often a tedious task, which can limit the sharing. When it comes to cleaning up data, filling in missing data, etc., AI can be extremely useful. Data wrangling is an important part of AI, and AI can help with data wrangling.

This is one of the dimensions of the ambitious P16 project on sovereign commons for artificial intelligence, which was launched yesterday. P16 is built upon various software packages, including the famous scikit-learn open-source library, which is a world reference in machine

learning, initiated at Inria 15 years ago. Today, in Europe, we need to be able to offer our economic players the possibility of storing their data in a sovereign manner, but also operating them in a sovereign manner with an effective AI software infrastructure. What is true for economic players is also true for research. The aim of the P16 project is precisely to provide this software infrastructure.
We believe that we, Europeans, have the skills and tools, with scikit-learn and many other software packages, to build an AI software infrastructure that will offer European open science an alternative to the tools provided by the big tech companies. The European Commission's and the DG RTD are, of course, well aware and engaged on the transformative potential of AI and generative AI in scientific research, and I'm convinced that EOSC has a role to play in this. These examples were about AI for Open Science. Now, the other way round, Open Science can also be an extraordinary fuel for AI, as we all know. AI is based on massive data. Applying the FAIR principle to research data, whether coming from measurements or simulation, increase the use of this data for training an AI model. A number of communities are already well-organised in this direction, notably astronomers and physicists.
Let me take again the example of Software Heritage, but now from an AI perspective. Beyond the open science and preservation dimensions, which are recognised by UNESCO, Software Heritage is also a wonderful tool for software engineering. After several years spent carefully harvesting source code from all over the world, it suddenly became clear in 2022 that this huge archive was a formidable tool for training generative AI models for coding. This application now seems obvious. It was not until three years ago, and it illustrates, in my opinion, the unexpected outcome that can be obtained from the long term support of pioneering projects. One final point before I conclude. One AI issue that becomes increasingly important, in particular in the era of Generative AI, is the evaluation of AI systems. As a Program Agency, Inria is in the process of building a centre for AI evaluation with the LNE, the French National Laboratory of Metrology and Testing. This goes far beyond AI for science, in particular in the context of the AI Act. But in the context of EOSC, I think it's very important to keep this dimension in mind: the evaluation of AI models, in particular LLMs, must be at the heart of our concerns. At a time when more and more scientists are using them to produce bibliographies, article summaries, state-of-the-art reviews, and so on and so forth. We are exposing more and more tasks to AIs, which risks causing us to lose skills. The least we can do is ensure that these tasks meet quality criteria. Our ability to evaluate AI systems used for research activities will be decisive.

My concluding remark is that, in the context of the profound changes brought by AI, it is more necessary than ever in the EU to develop storage, computing and data processing infrastructures for science. Inria, as a research institute, as a Program Agency, and as a mandated member of EOSC, is ready to invest in this, with all its French and European partners.

Thank you very much for your attention.

Keywords: EOSC, Inria