Public Transcript - SWH Symposium and Summit 2023
(English)

UNESCO event page – English:
https://www.unesco.org/en/articles/software-source-code-documentary-heritage-and-enabler-sustainable-development

SWH 2023 Webcast
https://webcast.unesco.org/events/2023-02-07-software-heritage/

https://www.youtube.com/watch?v=GyNrYmXZe1Q

Welcome address and opening

Mr. Fackson Banda, Chief, Documentary Heritage Unit, UNESCO

Welcome to the second annual Symposium and Summit on Software Heritage under the theme *Software source code as documentary Heritage and an enabler for sustainable development*.

It is being co-hosted by UNESCO's Unit for Documentary Heritage and Inria. The Unit for Documentary Heritage is the Secretariat of the Member of the World (MoW) programme, whose threefold aim is to identify, preserve, and access documentary heritage, including in its digital form.

As such, this symposium and summit is pivotal to our efforts to place software source code at the heart of digital documentary heritage. Key issues relating to this will be unpacked in a panel that I will be moderating after the welcome addresses. I am Fackson Banda, Head of the Unit for Documentary Heritage, and I will be a moderator for the first part of this morning's proceedings.

Mr. Tawfik Jelassi, Assistant Director-General for Communication and Information, UNESCO

Mr. Fackson Banda
Turning to the welcome addresses, I am immensely pleased to invite the Assistant Director General for Communication and Information Mr. Tawfik Jelassi to deliver his welcome remarks.

Mr. Tawfik Jelassi
I'm very pleased to welcome you all this morning to the UNESCO Headquarters for this annual Symposium on *Software source code as documentary heritage and as an enabler for sustainable development*. We are gathered here today to
demonstrate the multiple dimensions and interconnections of the domain of software and the role it very much plays in
safeguarding preserving, as Fackson said, documentary heritage and software heritage and share it across the world.

In today's digital world we know that software is everywhere. It empowers our personal computers, it empowers our mobile phones, it is the foundation for scientific research, and it enables access to public services. Madame Shaeder is with us here to represent the public sector in France.

Of course, there are many other applications of software as an enabler and as a driver for value creation and digital transformation. However, the value of software is quite often overlooked. The failure to recognize software as a

distinct form of a human knowledge, as a creative expression of problem solving, and as part of humanity's heritage. The fact that we overlook this distinct part aggravates the problem of software preservation.

UNESCO works to preserve many forms of heritage: national heritage cultural heritage, intangible heritage, as well as documentary heritage. This is part of the sector I'm leading at UNESCO, which is called the Communication and Information sector with documentary heritage being the unit in charge of software heritage.

Mr. Banda mentioned our Memory of the World (MoW) programme. This is a 30-year running UNESCO programme that identifies, preserves, and improves universal access to the world's documentary heritage and safeguards it for future generations.

For UNESCO, digital heritage is part of humanity's documentary heritage that obviously needs to be protected. With a clear alignment of our missions UNESCO and Inria, partnered back in 2017 to support and encourage international effort to preserve and make available the knowledge embedded in software source code centered on the ongoing development of software heritage archives.

Recent key moments of the collaboration between UNESCO and India include the publication in 2019 of the Paris Call - Software Source Code as Heritage. This Call identified several challenges facing software. That same year we launched The Software Heritage Acquisition Process with the University of Pisa in Italy to rescue and archive landmark legacy software. Last year we supported Software Heritage Stories visually highlighting software materials and promoting computer museums and other cultural entities.

Today software contains a large portion of scientific knowledge. In 2021 UNESCO Recommendation on Open Science which was adopted by our 193 Member States. This recommendation was developed to identify Open Access and Open Data measures to facilitate the production and dissemination of scientific knowledge worldwide. This was made possible through the use of software source code.

We are very pleased to see France include software in its National Plan for Open Science, Second French Plan for Open Science, alongside publications and data leveraging software heritage to ensure that research software is archived and properly referenced.

Software source code also powers public services and industry applications that are vital to any large organization. Artificial intelligence (AI) and digital transformation (DT) competencies are much needed for civil servants if governments want to succeed in their digital transformation initiatives.

Last September UNESCO developed a Digital Transformation and Artificial Intelligence Competency Framework for Civil Servants worldwide. We believe this is the ability and competency that civil servants and governments need to have in order to succeed in their digital transformation journey. Obviously, software source code is part of this work.

Software is essential to promote access to information and building a knowledge-based society that goes beyond the well-known information-based society. UNESCO will continue to support the Software Heritage initiative ensuring its long-term viability by raising international awareness of the value of software and its source code.

Finally, I take this opportunity to thank Inria for its commitment to this mission as well as all the supporters of Software Heritage who are gathered here this morning.

Software source code represents unique knowledge of humanity's recent history. It is crucial to work collectively so that the knowledge embedded in software source code is properly preserved, valued and shared with all.

Ms. Stéphanie Schaer, Directrice interministérielle du numérique, France

Mr. Fackson Banda
It is now my singular honor to invite Ms. Stephanie Share who is the Director responsible for Interministerial Digital Affairs, interministérielle du numérique (DINUM), in France to give her welcome address she will speak in French.

Ms. Stéphanie Schaer
First of all, I would like to thank UNESCO and the Director of the Software Heritage Foundation for this invitation, significant in my view of the role that public administrations can play in the field of free software.

One of the founders of the free software movement, Richard Stallman, used to say that the motto of free software is "Liberty, Equality, Fraternity:

Freedom, because the users are free.
Equality, because they all have the same freedoms.
Fraternity, because everyone is encouraged to cooperate in the community. »

This reference to our national motto resonates particularly in our ears, here in France, and I know how much France is expected to play a significant role in the free and Open Source ecosystem, an ecosystem in which communities and businesses, large and small.

The French public administration has a long history with free software:

1. **From 2006**: the first groups were created at the General Directorate of Public Finance around free office automation;
2. **In 2012 publication of the so-called Ayrault circular** setting guidelines for the use of free software in the administration, in particular to develop the culture of use of free licenses in the development of public information systems;
3. **In 2016: the Law for a Digital Republic** which makes source codes communicable and reusable administrative documents,
4. **in 2021: the so-called Castex** circular reinforces the rules on the openness of source codes and public algorithms and each ministry appoints a referent on these issues: the ministerial administrator of data, algorithms and source codes. I assure myself as general administrator of the data, algorithm and source codes the animation.

This rise in power concerns both the more massive use of free software within administrations and the opening of public source codes and we are working more and more as a contributor for external projects.

I am indeed convinced that free software has its place to offer public officials an attractive, ergonomic and increasingly collaborative working environment.

And this is beginning to be a reality for many agents:

The state videoconferencing tools Webconf and Webinar are instantiations of Jitsi and Blue Button respectively the instant messaging tool is an instantiation of the Matrix and Element codes and I know that public officials have recently been able to contribute to the development of a still little known free collaborative spreadsheet, Grist.

This is how I envision the public sector's investment and contribution to free software in the first place: by using the products and as much as possible by contributing to improving them.

Public source codes are also strategic assets of the State, in the same way as data.

- We therefore support any initiative that encourages their openness and sustainability, such as the Software Heritage project.

- The State must ensure that the source codes it has developed are published for the purposes of transparency and pooling.

- Support for Software Heritage is part of this desire to build open and shared infrastructures at the service of all. Free solutions have the potential to offer efficient, ergonomic products when a critical mass of contributors is gathered, whether from civil society, companies or States. It is through initiatives like Software Heritage that we can find a space for dialogue and exchange conducive to more sharing.

And it is in this spirit of sharing and openness, beyond our borders as the place in which we find ourselves invites us to do, that we register."

Mr. Bruno Sportisse, Inria CEO

Mr. Fackson Banda
It is now my privilege to call upon Mr. Bruno Sportisse, the CEO of Inria, our co-host, to give some opening remarks he'll be joining us by video.

Mr. Bruno Sportisse
Hello, I am Bruno Sportisse. The CEO of Inria, the French National Institute for Research in Digital Sciences and Technologies. I cannot be with you because I am in Brussels today for an event near the European Union but wanted to say how much Inria has supported the Software Heritage project since the launch of the project.

Inria has always been for the construction of open digital infrastructures. That's why for almost 30 years we have been alongside other players, including the circle which put it at the start of the World Wide Web Consortium (W3C), which supports the open standards of the web and the legacy software initiative. Which is well in line with this lineage. Since the challenge is to have an infrastructure for Open Source software in the world and to keep it's history, to keep its evolution.

Since the beginning Inria, alongside the dynamic team with Roberto Di Cosmo, in particular, and more than 20 partners, has supported this initiative through the Inria Foundation. I wanted to thank UNESCO which has now been supporting the project for many years, legacy software. I wanted to say how important it was to support this initiative and thank all the partners who will join it.

Mr. Roberto Di Cosmo, Director, Software Heritage

Mr. Fackson Banda
Finally for this part of the program it is my pleasure to invite my colleague and Inria interlocutor Roberto Di Cosmo to give a few introductory remarks.

Mr. Roberto Di Cosmo
I will take a few minutes to drive you through a little bit of what software writers actually is and then present you with the rest of the day's agenda.

We have heard from the introductory remarks that software is really important. Software is all around us. It is the engine of the digital transformation (DT). It is a fuel of innovation. It is a major pioneer of academic research. It is an essential tool for public administration. It is the digital fabric that brings together our digital lives in all their aspects. But what people usually tend to forget is that software doesn't come out of the blue. It is written by human beings developers to other human beings to actually read and understand it.

As you see this remark by a computer science professor.

"Programs must be written for people to read, and only incidentally for machines to execute."
Harold Abelson, Structure and Interpretation of Computer Programs (1st ed.) 1985

Like Donald Knuth used to say, one of the founding fathers of computer science, actually programming is the art of explaining to another human being what we want the computer to do. This may seem a little bit abstract, so let's make it a little bit more concrete.

This small excerpt (slide) comes from the 60,000 lines that were on board the lunar landing module that allowed us to bring a man on the moon, actually NASA. Let me also say that if you manage to put a man on the moon it is because of a woman, Maria Hamilton, who directed a huge amount of people working on that software.

This fragment of source code that you will see here (slide), there is an assembly code that was actually going to work on the machine and it was running on the landing module. On the right side, after the number side, you see comments in English. It says what is going on. I will not go through that today because it would take too much time. As you see, it's very important. This is a message from humans to humans, not just to the machine.

Fast forward many years later, in France, for example, there is a very important piece of source code, Parcoursup source code. Which is a tool that is used in the French Administration to decide where kids will go to university. This code helps understand in everyday life why your kid is sent to one university instead of another. This code is actually public. It is archived today.

As the board director of the computer research museum used to say, access to the source code of a program provides us human beings with the view into the mind of the designer. This is absolutely precious. Of course, I want software to run it and to use it but even more than that I want to understand what it does. To be able to modify, adapt it, or reuse it. For that, we need the source code.

Here at UNESCO several years ago, for two days, forty experts from all the different disciplines from different countries met together to come out with the Paris Call: Software Source Code as Heritage for Sustainable Development and digital sustainability. This Paris Call is published, it has been signed by many people, you can reuse it, it contains the results of many people. There are many interesting parts, but let me just take a next step.

There is a Call to recognize software source code as a fundamental enabler in all human activities. Let me give you an idea of some of these activities. For example, open source software is instrumental for technology and science.

"The real antidote [to epidemic] Is scientific knowledge and global cooperation."
Yuval Noah Harai (on COVID 19)

As we have seen during the terrible past years with COVID, etc., there is a relevant remark by Yuval Noah Harai, who's the author of the book Sapiens which was a bestseller many years ago. He said the real antidote to epidemic, I would say to any big human global challenge, is not closing down on yourself but it's actually scientific knowledge in global cooperation. Some challenges are too big for a single entity to actually address.

The top 100 paper
"[…] the vast majority describe experimental methods or software that have become essential in their fields."
Nature, October 2016

Software today powers most of modern research. Not just in computer science. Actually, in computer science we are producing the least amount of software research. In many other areas: physics, biology, social sciences, everywhere, you find software development.

“Telling historical stories is the best way to teach. It’s much easier to understand something if you know the threads it is connected to.”
Donald E. Knuth
Len Shustek
CACM, January 2021

When you want to teach people about software and bring people to work on software, which is so important, we are all human beings and nothing works as good as a story. In the stories you learn how software is developed, by whom and why. In this is a picture (slide) of Knuth, he was calling for us to rebuild, collect, and share the stories of landmark software source code around the world. If you want to do this, we need a dedicated infrastructure to collect, reserve, and share all of these casuals through school studies. All these special stories.

If you move on and look at other sides of our society, software is extremely complex today. If you look around thanks to free and open-source software (FOSS) you can reuse components that come from the many different places. For example, GitHub, which is the most popular platform today.

So many others, feedback at GitHub instances, local instances, we have here and there distribution like Database Manager (DBM) package like Papaya or NPM, etc. Sorry if I sound like a geek. I actually am, so it's normal that it goes like this. Since this complexity is growing, it is important to ensure the software works properly.

There is a United Nations Regulations on Cybersecurity and Software Updates, which was published in June 2020, that pushes the automotive sector to make sure that all version of software installed on cars are tracked properly.

Politique publique de don d'algorithmes et de codes sources
"... animer les écosystèmes de... réutilisateurs de code source."
Circulaire du Premier ministre, 27 avril 2021, France

Then there was a directive from the Prime Minister here in France about the importance of helping the developers or open source to work together in the different administrations.

Dec. 4. Enhancing Software Supply Chain Security
“…ensuring and attesting, to the extent practicable, to the integrity and provenance of open source software.”
May 2021 POTUS Executive Order

Then, for example, we have seen in the United States last year in a presidential or an executive order by President Biden on cyber security. In there are many items. One of the items is to make sure that we can actually know where all components of open source software come from and how we are developing it.

If you really want to do this properly then we need a trusted knowledgebase. Not just a place where we have information, but a trusted place where we can track the origin of all these software components.

Finally, when we say we need a trusted and shared infrastructure, what kind of instruction infrastructure can we use (slide)? We cannot continue working like we were doing before, not having a common and sharing and dedicated infrastructure because, otherwise, we lose software. There is link rot on a web page but then your web page but then the URL goes away or the web page goes away. Then there is a data rot. You just put it on a floppy disk but then 20 years later you cannot read it anymore. Or you can put it on a good hosting and development platform. Everybody does it. You put in on GitHub, GitLab, etc. That's fantastic, but good hosting platforms are not archives.

We have seen already, so many examples (slide). In 2015, Google Code shut down, Gitorious.org was acquired by GitHub and shut down, leaving one million projects in danger. In 2019, Bitbucket phased out mercurial control system. A quarter of a million projects in danger.

This summer we heard some of the executives at Orchid.com saying maybe we should remove projects that has been inactive for one year. You see what is going on. We really need to make sure our precious knowledge is not in danger and properly preserved. The bottom line is that we need a universal archiving platform to take care of this precious heritage. This is where Software Heritage comes in. This is a mission at the service of humankind in some sense.

The Software Heritage project was unveiled in 2016. Actually, we started with Stefano and Sean Francois working on this before unveiling. It was a lot of work to get to the unveiling. The mission is very clear: go out collect the source code of every single piece of source code publicly available, make sure it is preserved, not lost, and make it easy to find and reuse. In a sense we are building a reference catalog. When software is spread all around in all places, now we are building a unified global catalog. In archives make sure the software doesn't go away. Remember, Google Code Gitorious.org, and Bitbucket shutdown a quarter of a million projects. Rest assured that we have collected all that and archived all that so you can still find it.

Last but not least, we are building the first layer of such an important global shared infrastructure to explore the galaxy of software development (slide). The idea is to be ONE shared infrastructure, open and shared among everybody. It is not building a monopoly. It is one shared infrastructure but open. This is also in terms of ecology today. We need to be mindful of how much energy we use instead of having tons of people building an archive here and there, let's build one and make sure the data is not lost and that it caters to the needs of everybody, be it cultural heritage, industry, academia, or public administration.

Today this is the largest archive ever built. Its cost is about 200 million for a project, almost 14 billion unique source code files. These come from many different places, GitHub, GitLab, Bitbucket, etc.

This is an operational infrastructure, an evolving infrastructure (slide). We actually go out proactively, collect and archive everything, and put it in a gigantic graft that allows us to trace the origin of what we've found. We also provide mechanisms for people to go and deposit the software source code that triggers the archive manually. We provide intrinsic, cryptographic identifiers, which are fundamental for security, integrity, and reproducibility, for over 25 billion software artifacts, which are in the archive.

This comes out as a revolutionary infrastructure (slide). On one side, you have the full graph of software developing in a single place. You can trace the origin of all the components developed in the world. Not any graph, actually it is a Merkle graph using cryptographic identifiers inside. You have a half of the blockchain (not a good word), the same kind of technology put to good use. It's a pillar of open science. We finally have a place where we can store, reference, and reuse a source code that is used in research. This incredible playing ground for doing machine learning, big code analysis, vulnerability detection, etc. in an automated way.

This was a long talk, let me give you a short overview. Remember this Apollo 11 except here (slide), you just click to this link, we connect to the Software Heritage archive, here we are in the archive with the cryptography identifier. When we go to the archive, we find exactly the fragment of code that I have shown it to you. It is inside all the context of the archive. It has been found in a particular project on GitHub, it has been authored by a guy who is working on rebuilding the history of this software, etc.

You can find software which is archived using the National Platform for Open Access, whom we have been collaborating with for over 6 years (slide). The HAL platform for which we've been collaborating with for over 6 years. For example, this is a linear algebra toolbox from researchers where you have all the information from the authors of the affiliation, etc. If you click there you get the archival version of Software Heritage and you can visit the source code nicely. It is much better there, than just a zip file on top of it.

Since Ms. Schaer is here, let's remember that the French government, and thanks also to the big work by Sebastien Guerry who is here in the room with us today, has been working for a while on building a catalog of the source code of a software application which is developed by public administration in France. You'll see for each of these, you have this tiny, little logo of Software Heritage that corresponds to the archival version of the source code (computer

screen). You have public administration, we have research, traceability, and security, all coming together in a single infrastructure.

This single infrastructure is not an easy undertaking. We need to build it together. That's the reason why I'm so grateful to UNESCO for working together on this initiative because it is worldwide so we have an agreement with them. We had the honor and the pleasure to renew in 2021 November in this very room. It also requires a lot of effort and money to build the machine to pay the teams, etc.

We are very grateful to Inria for launching the initiative and to the many partners over time, we have many sponsors in the room, who decided to contribute and mutualize the cost of building this infrastructure.

Today Software Heritage has a team of 18 who are passionate and dedicated. Let me thank them all. They have been working day and night for years to get where we are today. They are dedicated people who could get an incredible salary in the private sector. They prefer to stay here for a mission we are all dedicated to so if I may ask you can we just thank them.

Two years ago, we started another program, a Software Heritage Ambassadors programme. Calling on people who want to share the knowledge and the world in different areas and many of them are in this room today and they are growing by the minute. Then we have to thank also the representative from ENEA, who came From Italy here. We are building the first mirror of the infrastructure.

We celebrated 5 years a little bit more than a year ago, in 2021, with the largest community who is growing every day. I hope this moment will grow and this infrastructure will become jointly shared and built by every country in the world.

Let me finish my lengthy presentation. I'm sorry for taking so much time. You see, there is passion behind this and there are many years of work behind all this.

What is going to happen today, we brought all you together for many reasons in the morning for this public event. We have 4 sections. We will start with the panel that will be led by Mr. Banda on the aspect of software and source code from the point of view of cultural heritage and education. Then we will move to a sequence of presentations that will present groundbreaking, bleeding edge ways of preserving over the long term all this gigantic amount of knowledge which is in this source code: from DNA to compression, to mirrors, etc. That's the geeky part, I would say. Then we have a coffee break. After the coffee break, we will have the pleasure to discuss the issues related to software in Open Science. Then we will finish with a panel from industry and public administration. We will hear from a large corporation on the Intel Open Source Technology Center, a representative of the European Commission, the Open Source programme office European Commission and a person from the General Secretariat for Investment (SGPI). Here in France about software and Innovation. This is the program we have worked to offer to the community here and our people listening online.

In the afternoon, there will be work in a separate workshop. Thank you again all of you for this incredible opportunity to work for the service of mankind.

Panel on Software Source code as part of Memory of the World

Mr. Fackson Banda Chief, Documentary Heritage Unit, UNESCO

I will now tend to the first panel on software source code as part of Memory of the World (MoW). Whether in in analog or digital form a software source code encodes personal and all collective memories of the world the aim of this panel is to look at the different aspects of software source code including the practical skills relating to its production, or development elements involving its value as digital cultural heritage, as well as issues surrounding the usage of software

source code as an integral part of the larger ecosystem of cultural heritage preservation and accessibility. To help address these and other related issues I have three experts all of where, all of them women I might add.

Ms. Natasa Milic-Frayling CEO, Intact Digital Ltd; UNESCO Preservation Sub-Committee; Professor Emerita, University of Nottingham

Mr. Fackson Banda
I would like to call upon Dr. Natasa Milic-Frayling who is the founder and CEO of Intact Digital Ltd. a digital continuity company that provides services for long-term care of software to enable use of digital artifacts ranging from archived scientific data to digital artwork affected by technology obsolescence. Natasa has more than 25 years of experience in computer science research and innovation, including 17 years at Microsoft Research. She has authored over a 100 research publications and has a dozen of approved patents to her name. Besides her research role Natasha led the MSR research partnership program promoting collaboration on strategic ICT industry challenges, including digital preservation.

It's my honor to invite Natasha to address us. She will start by giving us an overview of her institutional mandate after that I will pose one or two questions to her as she takes us on her journey of experience.

Ms. Natasa Milic-Frayling
In 2016 when Roberto told me about the Software Heritage Foundation, I was actually starting Intact Digital. So, we are sort of the same age. I heard about this fantastic effort to preserve the source code and I said, okay well, I'm going to do something different, but also in the line of preserving software.

Preserving software source code is absolutely critical because as you're moving to more and more to Artificial Intelligence (AI) and machine learning, understanding what they're doing is becoming absolutely critical. For Intact Digital we position ourselves as a digital continuity company. Which means that the software is preserved but we take it one step further. It must be usable. In computer science we computer scientists and developers were keen to really get the logic right, but then we also depend on our colleagues in IT, who make your printer work and make your security network work. These are the people who actually make software run. Intact Digital as an organization is increasing the awareness of the importance of software in general, and using the legacy software within the ecosystem in particular. Without software you can't really access digital artifacts.

I should also mention that I'm a member of the Preservation Sub-Committee (PSC) of the Memory of the World (MoW) programme. As a member of the Preservation-Sub-Committee we think about preservation in general. Within that Sub-Committee we also have a PERSIST programme, which is focused on digital preservation. We know with digital preservation when anything is digital it's absolutely critically dependent on software. That preservational activity then becomes by extra magnitude a harder problem to solve. With the book I have in my hands, I can have for centuries, but with software I'm lucky if it runs after 3 or 4 years. It becomes old very quickly.

Mr. Fackson Banda
Thank you Natasa for those introductory comments. Software source code itself is a cultural heritage. Why, in your view, is software also critical for other digital documentary heritage?

Ms. Natasa Milic-Frayling
The software is practically the light that we need in order to see information that's the fire. If you don't have light, you can't read the book. If you don't have software, you can't really make anything out of the files that you've stored.

It has been really hard to explain this to people. If you ask somebody, where are your documents? Where are your files? They point you to 1, 2, 3, 5, 7 copies of their files. You ask them where is the software? Well, it was on some old laptop or it was some in the hands of IT but nobody really knows where the executables are.

Even if they had the executables they wouldn't run on the laptop. They won't be able to install. On that journey of explaining that just storage of your files is not sufficient, we had to engage on a number of aspects. With UNESCO we're trying to educate that with digital documentary heritage the digital itself is a very sophisticated artifact and it lasts only while the software runs.

This was not quite understood. You can see it on the screen only when the software runs. When software doesn't run you can't see anything. This dependence between the storage of artifacts and the digital instantiation of it is not very clear. This journey is really about people's awareness. Then after awareness, of course, there needs to be action. If you want to have a digital artifact that's available to our children in the future, we need to think about how legacy software can be usable within the contemporary ecosystem.

In short, we as Intact Digital work with life sciences. I always say we've got good news. It's a hard problem but we've got good news. We've actually implemented the platform and we can show you that in virtual machines you can run legacy operating systems. You can install your stuff and, more importantly, you can have secure access to it. You can bring your data into the secure access.

We always say there is a problem but we don't just talk about the problem, because in the ecosystem at the moment we're really lucky. With the qualification of data centers and digitalization, everybody in Google Cloud, virtualization is becoming a commonly used practice in IT. Therefore, what they need to just think about is how to make sure that we can provide secure access to the fundamentally non-secure software. For cultural heritage this is key. We really need to think about it.

Just like Roberto mentioned, we have a platform that is going to save all our source code, because we need to know what's in the source. We need to know what is coded, and then we will be thinking about the platforms to run this software and make it usable are going to be.

Mr. Fackson Banda
Arising from what you've just said, what are the key factors in protecting software use?

Ms. Natasa Milic-Frayling
It's not very simple. People might think oh this is just a couple of virtual machines, then remote access and this is it. In fact, there are, if you look at the core framework, first of all a technology, of course. We need to know where the source code executables are, what it compiles and what we need and then put the technology in place.

That's fine but then you have to look at the licenses. The world as we know is not just run on open source and open source licenses. You can negotiate, but we have a whole commercial factor as a sector that has shaped the digital economy for the last 40, 50, 60, 70 years, you can argue. They don't have licenses for legacy software. That's a legal aspect.

Then you have operations. Suppose you've got a platform, then you need to make it secure. The software ages gradually. Whatever was okay today from the licensing perspective or installation perspective it's just fine, but in three years from now it won't be. It becomes legacy. It's aging a software is a continuous process. Operation is another thing.

Finally, the human factor is the biggest thing. We ourselves, we need to think about how to transfer the knowledge about source code, creating executables, and also installations. Even we should just simplify at the level of installation. When we are working with the light, with the life sciences and Pharma sector, they have to preserve things for 30 years. That means going back say forty years, going back to XP, and remembering what are the bits and pieces of what people worried about when they were installed right then and see how they can be installed in a virtual machine, for example. That memory of the practice needs to be there. It's going to be education as well as technical aspects.

Mr. Fackson Banda
On this specific issue of Digital Data, which is mounting every minute including data from valuable scientific experimentation, do you think that it is economically viable to keep all these?

Ms. Natasa Milic-Frayling
The time is short because we know people say okay storage is cheap, so we should be able to store everything. Then it's not actually cheap when you start thinking about things like paying your bill to Amazon, or Microsoft, or whomever. Wherever you store it, or if you have your own data center, it's not cheap. However, there are lots of efforts now to create new storage capabilities, very exciting ones. Like storing things in the DNA form is fantastic. I want to comment on other things.

Say if we do keep it, we might be able to keep it but then if you don't have software what are we going to do with it? This is another aspect. The good thing about a ratio of benefit costs when it comes to a software and data is that one single copy of software would enable you to start to open any file that's compatible with that software. The cost benefit ratio of maintaining a copy of a working software is huge. With legacy applications and tools, we don't go back that often. It is rarely used but it's critically used.

This is important to the understanding the sustainability model for providing access to the software that's rarely used.
That's why you don't really have legacy software anywhere yet on Amazon or Microsoft, because it doesn't fit the Cloud utilization model. The Cloud utilization model is all about computation, computation and then pushing data up and down the networks.

In short, I think when it comes to data, yes we may or may not be able to store, but the most important thing is we need to think about how we're going to do the software associated with the data. I believe that we are in a good position there, because with software, as I said, one copy was used to create millions of valuable documents and so we can read those millions of documents.

Mr. Fackson Banda
Finally, what is absolutely essential to ensure that the coming generations will be able to access and use digital heritage?

Ms. Natasa Milic-Frayling
It could be education. We need to teach them. My kids don't remember anything past 2000. As we are teaching them programming teach them the value of the source code and the value of cultural heritage that is in the digital form. We really need to attract them to think about and have appreciation for its legacy.

Also, when it comes to our responsibility for our generation, we truly need to go and talk to vendors and make sure that they having in their mind continuity as a part of the design process.

It also means from the legal perspective. They need to know that when they stop producing software, when they stop maintaining software, and when they stop selling software, they should give people rights to use the software. So that the legacy and the value that was created doesn't go to waste.

Ms. Claire Gillissen-Duval, Senior Director, EMEA & MEE Corporate Social Responsibility, Global Corporate Affairs, SAP

Mr. Fackson Banda
I would like to invite Claire Gillissen-Duval to take the floor. A few words about her. In role as a Senior Director of Corporate Social Responsibility at SAP, Claire builds and nurtures strategic partnerships that liberate talent, technology, and capital to help the world run better and improve people's lives. She has supported hundreds of non-profits over the past 20 years, with a sharp focus on youth empowerment, accountable grant making, and

philanthropy diversity. Claire is also the Founder and Global Lead of SAP's Africa Code Week, the largest digital literacy program in Africa that has introduced coding skills to millions of young people across 54 African countries since 2015.

Ms. Claire Gillissen-Duval
I work at a multinational, European, German software company called SAP that was founded 51 years ago. It was set in 1972. I have the pleasure of leading the CSR department for the EMEA & the MEE regions. When we were preparing with Fackson for this panel and meeting, Fackson said "Claire there are some acronyms that you would need to explain." Actually, SAP is an acronym in itself, we love using acronyms. CSR is an acronym for the corporate social responsibility department. In summary, it's the philanthropy arm of a company. Usually all major, international companies have a CSR department. I'm the CSR Senior Director. Also, the EMEA region stands for Europe. Middle East and Africa. Also, the ME region which is the Middle East and Europe region. I've been working with SAP for the past 25 years.

Mr.Fackson Banda
That's a long time. I'm particularly interested in the Africa Code Week. Can you tell us more about that and also, perhaps, point out the legacy that is intended by this important initiative?

Ms. Claire Gillissen-Duval
Absolutely. First as a way of introduction, I was also thinking with my team about the digital world where de materialized innovation emerges on a daily basis. Eventually, billions of new ideas that flow throughout our planet every second are taken for granted. Yet, Innovation is the results of years of research, tryouts failures, and hope. Considering immaterial goods as heritage is actually an acknowledgment of the fact that we are a continuity of the past and that our work today is meant to leave a trail for the generations to come. Heritage is absolutely at the center of the memory and a cornerstone for our development as a global community.

A few words on Africa Code Week. Africa Code Week and its counterpart Meet and Code were funded back in 2015. They were inspired by Europe Code Week that was funded, I think, in 2013 or 2012 by a very young and energetic woman Neelie Kroes who was, I think, 71 years old at that time. She was such an inspiration to us.

These times, I'm sure you will remember we had the desperate need to make coding accessible and available to the largest number of children, teachers, and communities. SAP imagined those digital literacy programs first as a way to spark an interest into technology, but also to use digital literacy programs to foster further inclusion, growth, and equity. We wanted to do that through the soft skills that are now mandatory to thrive in today's workplace. We're speaking about: teamwork, problem solving, collaborative learning, and tools.

Over the years we had education experts from the Camden Education Trust, which is a non-profit organization that is based in in Dublin, Ireland. They developed hours of teaching material on MIT Lab Open-source coding learning software. The name of this software is Scratch. Scratch was gender proofed by UNESCO. I want to take a minute here to thank immensely Mr. Jelassi who has introduced and launched this event and also Davide Storti. Davide Storti was critical when we worked hand in hand together and launched Africa Code Week.

Mr. Fackson Banda
Fantastic. Why the focus on open-source software to achieve your goal?

Ms. Claire Gillissen-Duval
As a starting point, knowledge is one of the things that grows when it is shared. Open-source gave us the tools to walk this talk. For various reasons, open-source software served SAP corporate social responsibility's goals through Africa Code Week and for the following reasons. Accessibility - as they are free and openly available to anyone making it accessible to obviously a wider range of people and including, most importantly, those in underprivileged areas. They are very collaborative. The collaboration reason was critical. They rely on collaboration, on community driven development, and they promote teamwork and social responsibility. We thought that it was critical for us to

use Scratch coding, the MIT lab media Labs Scratch coding, for transparency in the development process. We thought it was promoting accountability and trust. Last, but definitely not least, education and skill building. It can absolutely provide opportunities for learning, skill building, promoting, as well professional development, and technology education.

Africa Code Week had three major components. One was around the coding workshops. This is where we had several million children participating in Africa Code Week and developing their Scratch little program, and we had the African Code Challenge. It's more than a hackathon because it doesn't only last a couple of days. It's a friendly competition amongst the children of the various participating countries in Africa. The kids, absolutely, and the teachers as well like the African Code Challenge.

The last program was set up in 2019. It was launched prior to the pandemic in person in Rabat, Morocco. We had the pleasure of inviting women from the district for the occasion. With this women empowerment program we dug deeper into our commitment in breaking the technology gender gap by empowering African, female teachers and encouraging them to share their newly acquired skills in the classrooms and with their communities. All the dimensions of our signature digital literacy programs in Europe, Africa and in the Middle East are meant to be sustainable empowering and setting goals for the long run, in countries, and also economical settings where hardware is a real challenge. Where the internet is far from being reachable or affordable.

Our team has developed computational, thinking material that carry on our goal to raise a tech savvy generation. Also, because we're speaking about the Internet, I also wanted to recognize the fact that today I think 7th February is a Safer Internet Day (SID), so I thought it was nice to mention it here.

Mr. Fackson Banda
Finally, based on your experience of building partnerships at SAP, in what ways are collaboration and partnerships essential to the future of software?

Ms. Claire Gillissen-Duval
Collaboration and partnerships are essential for various reasons. I will mention three innovations as joint efforts between different companies and organizations, which can lead to breakthroughs and innovations that wouldn't occur otherwise. The cost savings is also an important point for any company to mention. Sharing resources and expertise can result in drastic cost saving measures and also faster development time. Which is good for us. The access to diverse skills and expertise that partnerships bring together results in better products. Which is good.

To summarize the idea, collaboration and partnership simply made Africa Code Week possible. Also, open-source coding is more than just a characteristic of a software, but rather a philosophy. It embodies the will to break down barriers, and to use science as a vector for inclusion, discussion, and interaction. It gives people the opportunity to improve one another's work and, at some point, one success is also the success of others. It propagates a fresh feeling of achievement across nations.

Last year Africa Code Week impacted the lives of nearly 3 million children, actually it was 2.6 million. Over the course of 8 years, we had, I think, more than 20 million children engaged. We were quite proud. It was not only SAP, it was the work of SAP but also all the other partners; UNESCO, Irish Aid, Google also supported us. It was only due to the fact that we had a very strong solid network of partners across an entire continent.

During the pandemic, instead of putting the programme on hold, as we did for other SAP CSR department's projects, we decided that we would innovate with our partners. This was the year in 2020 when we created the African Code Challenge and it was also the year where we developed the Africa Code Week app. We had launched the women empowerment programme just a year before, but to me this is the heritage of Africa Code Week. This is what this project is living for - the future, the possibility to impact millions of people, mostly children. Provided that we work all

together, that we feed from one another's ideas, and that we give back what we've achieved for it to grow sustainably.

The reason why I'm mentioning give back is also because we are coming to the final stage of the development of our programme and we're working hand in hand with UNESCO and with the Ministries of Education to hand over Africa Code Week for the Ministries of Education themselves to run the programme from 2024.

Ms. Brigitte Vézina, Director of Policy and Open Culture, Creative Commons

Mr. Fackson Banda
Brigitte is passionate about all things spawning culture, arts, handicraft, traditions, fashion and, of course, copyright law and policy. That passion comes through as Director of Policy, Open Culture and GLAM (galleries, libraries, archives, and museums) at Creative Commons. Furthermore, Brigitte gets a kick out of tackling the fuzzy legal and policy issues that stand in the way of access, use, reuse, and remix of culture, information, and knowledge. Before joining Creative Commons, she worked for a decade as a legal officer at WIPO and then ran her own consultancy advising Europeana, Spark Europe, and others on copyright matters.

Ms. Brigitte Vézina
It's a true honor to be here. As a little anecdote, I was an intern at UNESCO in the early 2000s, working in the copyright unit, of course. So, it's kind of a homecoming for me to be here today.

I work for Creative Commons. I've been there for three years and I lead the policy and the open culture programs. Creative Commons is fairly known among this group. You might recognize this logo, it's known for its set of licenses that were created over the last 20 years, in various iterations that really enable freeing up the knowledge and the culture that is otherwise locked behind copyright walls.

It was built after some events happening in the United States where the term of protection was increased by 20 years. This effectively reduced the public domain by protecting works for an extra period of time, thereby making the works that would have fallen into the public domain no longer available for free use. This led to a very powerful realization that, actually, some creators want their creations to be free of copyright barriers. They want to share. They want to have the freedom to decide on which terms and conditions their creations can be shared.

That's how Creative Commons licenses came into existence. They give creators the freedom to decide on which conditions they want to share. It's extremely powerful to build a commons of creative content that anyone can then use, reuse, and build upon to grow the universal sum of knowledge. So, 6 licenses built on 4 different conditions over 20 years has led to more than 2.5 billion works released openly.

For software, we don't recommend Creative Commons licenses, as a policy matter. We prefer to encourage people to use free software licenses. But generally, CC licenses can be used for all sorts of creative works. We're built on the values that I've heard repeated over and over throughout the morning sessions: openness, collaboration, and reciprocity, and that we all have a stake in building these commons, based on mutual benefit and equity. By making content more openly accessible we encourage more equitable sharing of knowledge. We also encourage transparency, interoperability, and a sense of community, where together we all realize that we can contribute to solving the world's most pressing problems.

I just want to say that preservation of software is only part of the equation. It's much more than storage. Software can be stored but if we don't have the tools to be able to shine a light on them and understand them, their worth is very limited. That's why preservation needs to come hand in hand with increased access, as the Memory of the World (MoW) programme promotes. But, to even go further to encourage use and reuse of that creative content.

Roberto, you said that the software code is kind of a way to see into the mind of the programmer, but what happens if people can't reuse them, then the software or the heritage in general cannot live through the minds of others that

will continue those traditions, and will be able to transform it, and reimagine it, and reinterpret it, so that we can continue to build our thriving software commons for the benefit of society. Software is heritage, a precious product of human creativity and it must be valued as such.

Mr. Fackson Banda
So, based on what you said with respect to use and reuse, what would you say is better sharing and why is it fundamental to building a common future?

Ms. Brigitte Vézina
Software is the infrastructure that enables everyone to be able to share their cultural heritage, but it's also heritage in and of itself. That's why it needs to be preserved, made available, and shared as widely as possible. It's really a necessary condition for us to address the world's biggest problems. For us to reach the Sustainable Development Goals we need all the knowledge, the data, the learnings, the culture to be open so that everyone has, at least, an opportunity to come and contribute with their own interpretations to help bring solutions to problems that concern us all.

The problem when the knowledge or the heritage is closed is that it becomes a privilege to be able to access it and work with it. When we open it up, it's more inclusive, fairer, and more equitable. So, like I said, based on these values of reciprocity we can together as a community help each other to try to find solutions to address the world's most pressing problems.

I just want to share a few statistics with you. I already mentioned the 2+ billion works that were openly released and that's just over 20 years, with the help of 6 licenses based on 4 different conditions that we can mix and match and two CC public domain tools.

Because I work in the field of open culture, I want to share a few numbers related to cultural heritage institutions. In 2018, a few years ago, the New York Metropolitan Museum of Art decided to open up its heritage collection and what it saw is, over a course of a year, a 385 percent increase in its visibility. People could access, view and understand those collections because they were placed online freely accessible thanks to Creative Commons CC tool cc0. That means that 10 million people per month could view the collections from the comfort of their living room, if they wanted.

A bit closer to us in 2021 the Welcome Trust in the UK, which is a pioneer of opening access to their collection, had reached 1.5 billion views of items in their collection. That's huge and that would not be possible if those were not openly accessible.

You might know that without Creative Commons licenses there probably wouldn't be Wikipedia. If you go down to every page of that free knowledge resource you will see that every page is licensed CC-BY-SA. Enabling people to not only access it but also reuse it in their school, in their projects, in whatever other endeavors.

Last statistic: 1 percent. Only 1 percent of the world's heritage institutions share their collections openly. That's very little. It's a tiny fraction. And why? Because there are so many barriers that prevent institutions from opening up their collections.

Three types of barriers. We did a survey of what could be those barriers last year and we came up with three categories: people, money, and policy. In terms of people, there is this very conservative mindset. Institutions in general tend to think of themselves as gatekeepers. We have those collections and we must take care of them. If we release them openly, who knows what's going to happen to them. There's a real fear or risk aversion of what might happen if we allow others to access it and reuse it and build new knowledge around it. There's also a lack of skills. Digital transformation is a very long process and it takes a long time for the staff of those institutions to gain the skills to be able to not only start an open project but also maintain it over time. Money. It's very expensive. We all know

that storage costs a lot of money and that's not every institution that has the funds to be able to sustain those projects over time.

There's also a fear of losing revenue. Some institutions like to license out their images and they think that this is a source of income if they can charge for people to download their images. However, many studies have shown that it's actually minimal and when you factor in all the costs of maintaining this licensing structure, it's actually not financially sustainable at all. In terms of policy, I might end on this note, copyright erects enormous amounts of barriers for the preservation of cultural heritage and the preservation of software. It makes it very difficult for institutions to make copies just for preservation purposes. When the challenges of climate change, armed conflicts and even the pandemic showed how precarious the sharing and the preservation of heritages is.

We need to find ways to limit the negative impact of policies that inhibit the possibilities of sharing culture. I must say that there's also a lack of a positive policy framework. We know that UNESCO championed a Recommendation on Open Educational Resources and, more recently, there's UNESCO's Recommendation on Open Science but there is no equivalent for culture. There is no Recommendation on Open Culture. There is no Recommendation on Open Heritage. Without this positive framework where people come together with share values and recognize the importance of having this openness in the heritage sector, it makes it very difficult for individuals for institutions and for communities to reach those goals

Mr. Fackson Banda
Thank you so much. You have answered almost all the questions I had wanted to ask you but perhaps a final one given the time constraints. Culture was just declared a public good. What does this mean for the sharing of Software Heritage?

Ms. Brigitte Vézina
I think that it means that there's a huge overlap in the values and then the ultimate objectives of UNESCO in terms of declaring culture as a public good. Creative Commons provides legal, technical, and, also, social infrastructure that allows people to share culture openly as a public good. Creative Commons infrastructure is public, it's decentralized, and it's built on open source. Without support from the community, it can go away very quickly. So, it's important to recognize this sometimes invisible, sometimes forgotten pillar of our shared culture.

Perspectives on long term preservation:

DNA storage and the future for long term archival
(Mr. Marc Antonini, coordinator of the MoleculArXiV flagship project)

Mr. Roberto Di Cosmo
It will be a series of presentations. We will start with Marc Anthony who is actually leading a groundbreaking project on archival on DNA.

Mr. Marc Antonini
I'm Marc Antonini, I am the Program Director of the Project Research MoleculArXiV flagship project and I will talk today about digital data storage and synthetic DNA.

I wanted to start my presentation with this slide. Which shows the first hard disk drive in history. It was built by the end of the 1950s. It's by IBM and its capacity was 5 megabytes. This means that in this disk you could store only one single image. It's an image that you can store on your smartphone, but only one image. The weight of this disc was 1000 kilograms. Today, of course, things have evolved and this kind of tape that is used for restoring data. For example, we can store 20 terabytes of data. So, millions of hard disk drives of IBM, like this one

The future, and the future that we believe in, is storing data in such kind of capsule. In this capsule synthetic DNA is stored. In this kind of capture we can put a lot of data. We can put thousands and thousands of tapes like that. It is shown that for a capsule of one gram of DNA we can write the equivalent of 200 petabytes of data (slide).

Why use DNA for storing data? First, as I said, it's ultra-compact. It's one billion times more compact than a hard disk drive. Also, it can last hundreds of thousands of years. If we keep the DNA in a cool and dry place. As an example, we were able to do sequencing of the DNA of mammals one million years old or even 700 thousand years old. It just means that the DNA is very stable for a long time. This is also an eco-friendly solution because once the DNA is created, when you start once you store the data into DNA, there is no carbon emission because you can keep the capsule. You can keep it at room temperature without any restriction.

So today, of course, the price is still high for writing DNA, but still it's less expensive to store one gigabyte of DNA than one megabyte of data hard drives. We are confident that this price will decrease. Also, the synthesis speed is low. We have to do things for accelerating the writing for the DNA synthesis in the future.

What kind of data are we talking about? We are talking about code data. So, data that you never access, or you access rarely and that must be archived for the long term. This kind of data represents around the 60 percent of the data that is stored today in data centers.

How does it work? How are we able to store images into DNA? The workflow is presented here (slide). As you see, I took an example for one image, but it can be converted for any kind of digital data. There are three parts: an encoding part, a decoding part, and in the middle there's something related to the biochemical process.

The main goal is to convert the binary information into quaternary information. The famous nucleotide of DNA, ATC and G so a denied t mean C design and guanine. Once we are able to convert binary into this quaternary code, we must do the synthesis, the molecular synthesis of the DNA. Once the DNA is synthesized, if you want to store it we have to put it in the place where it is preserved from water, oxygen and light. For example, in the light capsule I showed you. Then you can store it for a long time. The day you want to recover the information that you stored in the DNA you have to open your capsule to do the sequencing of the DNA of the strands that you store into the capsule. Then, after sequencing, you can recover the DNA strands. The famous ATCG. Then you have to convert it, ATCG into a binary file.

Obviously, this chemical process is prone to error. When you recover, after sequencing the ATCG strands, you have to

do an error correction, and then to decode your corrected strands. The noise that is introduced in the biochemical process is mainly substitution, like in the standard communication channel. In there, so insertion and deletion which is more complicated to handle and which correspond to insertion or deletion or nucleotide of nucleotides in the work in the process.

It's not so obvious to convert binary into quaternary, because we have to take into consideration of course the noise that can occur in the process, this is the goal of error correction codes, but also some coding restrictions. First of all, we are not able to synthesize today very long nuclear strands of DNA without noise. So, we are restricted to approximately 300 nucleotides.

There are some constraints that are imposed by the sequencing. For example, we cannot generate codes that represent some homopolymers, like AAA, for example, or TTT (slide). We cannot repeat later like that, so this avoids some coding codewords. Also, the percentage of GC should be lower than the percentage of 80 in the strand. We should avoid pattern repetitions, ATC and TC. In some cases, there are also other restrictions like prohibited codewords that cannot be used for the encoding. For example, people are doing amplification, not with PCR but with bacteria. In that case, we should not clear the bacteria with prohibited codes, and could create some RNA.

The challenges today are various in this kind of research topic. We have to deal with the new Next Generation Synthesis (slide). To define, really fast chemistry for adding polymers to accelerate the synthesis. There is also some research in non-DNA synthetic polymers that will allow use of more than 4 bases for encoding the data. We are talking about NRA code, and not for a quaternary code. There are also some challenges in the Next Generation Sequencing. With ultra-fast sequencing we are talking about the third generation sequencing of third generation using your nanopore technologies. We have to consider the noise and the constraint introduced by the new chemical processes and sequencing. Especially, the noise model, to model the noise that appears in the channel, to design operation for this noise, and to also provide robust decoding after sequencing.

One point, which is very important, also is the management of Big Data. Once you have, for example, stored a whole data center in a capsule or one in one gram of DNA, you must be able to recover one file in such a kind of soup of DNA. There are millions and millions of DNA strands. You must find the good strength that’s useful for recovering your files. It's what we call random access. We can call it like a DNA SQL. Also, there are some works in joint sequencing decoding to adapt the sequencing to the encoding process.

In that context, one of the main goals of the MoleculArXiV project is the development of a new paradigm for digital data storage, based on efficient coding decoding solution and Next Generation Synthesis (slide). Next Generation Synthesis will allow you to go 100 times faster than commercial techniques of today. This will allow reach, by the end of the project, 10 gigabytes of data written in 24 hours with off-the-shelf parallelization.

On the other side, we present some examples of possible achievements in terms of data storage for cost of one Euro over the next decade. On the right-hand side (slide), you can see that in 24 hours we hope to reach the read and write of one megabyte of data for one Euro. We start from the storage of value data in the near future, such as cultural heritage data, for example. To a point where a molecular data, multiple data storage could fundamentally change the economics of data storage and distribution. As it becomes a credible alternative to data centers at the cost of one Euro per terabyte. In that case, molecular strands will fundamentally alter the economics of storing and distributing data because it becomes a credible alternative to data centers.

The objective of the MoleculArXiV Project is not just to get a few people to work together, but to create a new research community (slide). To encore it in an efficient internet national variation system, through a series of relevant tools and articulated in articulations, made possible by the involvement of the INR and the research institutions. We want to start a community by allocating funds to a first cycle of laboratories directly involved in the project, as well as, to new, existing platforms. The community will be developed through integrated research (IR) calls for projects, and through the recruitment of young researchers via the Chairs. One of the objectives is to set up recurrent workshops, as well as, recurrent summer school, mainly every two years. Which will allow to encourage discussions

between researchers from different disciplines, which will allow the emergence of new collaborations. The technology transfer will also be privileged with the objective for creating prematuration projects or startups in the field. The project will start at the end of next year and we are starting scientific work today.

To finish my presentation, I just want to highlight that the project will focus on two aspects. The first one, is building tools to manage DNA databases and secure them, and second one, is demonstrating various applications of DNA data storage, ranging from molecular tagging to special recording of information in DNA.

The main objective of MoleculArXiV is to demonstrate the end-to-end feasibility of storage on DNA and non-ed polymers (slide). To do so, the idea is to perform large-scale data storage experiments for several institutions, like UNESCO and Software Heritage, with which we are already in contact, and to simulate aging and physical stress on the data. We expect many innovations, as well as, a good understanding of the need of the end users.

Roberto Di Cosmo
We are really looking forward to seeing this next revolution. It was very nice to see what was a hard disk in the 50s. We see what will be in the next 50s.

Large scale compression of software source code (Mr. Paolo Ferragina, Pisa University)

Mr. Roberto Di Cosmo
In the technology session you have seen the promise of a long-term archival using DNA, which is revolutionary today. Still even with that incredible focus, with the amount of data we have to store, for example, satellite is too big. So, what can we do? One possibility is to try to shrink it a bit. We are looking at what Paolo Ferragina, who is a professor of University of Pisa and a specialist in data compassion, will tell us today.

Mr. Paolo Ferragina
I will talk to you about data compression in the software archive. I will tell you what we've done up to now, and what we plan for the future.

When you talk about data compression you typically think about shrinking, hence space saving. This is very beneficial for transmission and storage; but there are other rich dividends, that from data compression. The first one is energy saving, because you are using less servers to store and access your data. Then, speed because whenever you shrink your data you move more data close to the processor, and therefore you can access them with the speed of nanoseconds instead of the milliseconds of your hard drive. Another advantage is given by the burning out of the sectors in solid state disks, because you are using less space. Finally, something that will come to the end of my talk is the fact that there are new techniques nowadays that allow you to compress the data and access them only for the part you need.

Why we are here is because in recent years we have done something about data compression (slide). For example, we contributed to the design of Brotli, the data compressor of Google. This is the paper (slide - Large-scale compression of software source code) that we published and allowed us to get a Google Faculty Award.

In the year 2000, i.e. more than 20 years ago, we designed the first compressed data structure (slide). In the sense that we were able to store data in compressed form, and we were able to search the compressed data without decompressing all of it. It's called the FM Index nowadays, and actually connects to what Marc was saying in the previous talk (slide).

The most circulated killer application of our proposal is related to DNA sequencing, because this technology is able to squeeze whatever sequence of bytes you have and be able to very efficiently count how many times any sequence of bytes occurs in your files. For example, it’s the backbone of the tools which are called BWA and BowTie, which constitute the most important tools for DNA sequencing nowadays.

Starting from these premises, we decided with Roberto to attack the problem of compressing the Software Heritage archive. This is because the raw space of the files (blobs) is very huge. It's more than 90 percent of all the data contained in their card.

How to approach the problem (slide)? If you try to approach the problem in a, let's say, classic way, you go to a benchmark suite and you try to see what happens. The most known is the Squash Compression Benchmark. You can see in the home page that they have more than 46 codecs available. If you consider all possible configurations, you get about 7,000 possible compressors.

Hence, if you try to approach the compression problem in a brute force way, you would be lost. So, we need to approach it in a more principled way (slide). Let's look at what actually data compression offers nowadays. There is a very classic result out which is called the Lempel-Ziv parsing. All of you should surely know it, because it's the backbone of gzip. It very recently became more and more active and more known, because of other implementations which are much more performant, like 7z, brotli,zstd.

Then, there is another family of compressors based on the Burrows-Wheeler Transform. This is actually the core of the FM index. It's a very sophisticated mathematical transform that takes your data, and transforms it in a way that is more compressible. In the year 2000, we showed that these data are not only more compressible but also searchable. The BWT is behind the bzip compressor, and it is also behind the FM-index, the third option in the class of searchable compressors you want to scale to petabytes of data it is very, let's say, dangerous.
By the way, of course, machine learning is coming into the data compressions field in a very important way, but up to now, as far as I know, machine learning compressors that are very slow and don't achieve the best state of the art. For this reason, we have not yet investigated these kinds of compressors.

What is the important feature of Software Heritage, that is not compressing a single file (slide). We are compressing a collection of files, and these files have different kinds. It's not only source code. In order to compress and to attack this problem you have to use what is called the PPC Paradigm. PPC stands for permute, partition, and compress. In the sense that whenever you have a collection of files what you want to do is to permute the files, so that similar files come close to each other. Therefore, whenever you apply the compressor and the compressor operates, let's say, in blocks you can compress blocks of similar files. So, you can get the most from them, from the compression of all of them together.

The questions are, how do you: design the permutation; design the partitioning; and choose the compressor after that? You have three pieces. We designed these three pieces by studying two experimental scenarios. I will comment only on the first one. That scenario which is the BackUp scenario. You have the collection, you want to store the archive, and you want to keep it there and guarantee streaming access to the compressed data.

There is also another scenario, which is possibly the most interesting one. Which is the random access scenario. In which you ask for one specific file and you want to decompress only that file. I have no time to discuss this part.

Finally, whenever you want to permute and partition, you have to establish what the features you use in this permuting and partitioning strategies are. You could use the file name, you could use the path, which is an information available in the Merkel graph. You could use the file type, the content, but on the partition side you have to decide what you want to do. Because if you want only to compress the data and to backup it up, you can use very large blocks.

Whenever you want to be very surgical and access only a very small amount of information your blocks must be shorter, because you want to decompress only these blocks. Of course, changing the size of the block impacts on the compression ratio. This is something that I will address.

We have taken a very tiny snapshot, for now, of the source code collection (slide). We took C and Python codes because, with Roberto, we decided to see what happens, if there is a difference between these files. Actually, there

are some surprises. Also, we took some repositories that got more stars. These are just a few logos about the main repositories we have downloaded. Each data set consisted of 25 gigabytes. It's very small compared to the entire archive, but it gave us very interesting feedbacks for now.

So, this is the main picture I will comment on during the rest of my talk (slide). To tell you what it means, on the X-axis you have the compression speed. On the Y-axis you have the compression ratio: How much is the compressed file with respect to the original file? The lower the better. The arrow that goes toward the bottom, right corner means that the more you gor the better it is the algorithm, because the algorithm is faster and data are highly compressed.

What are these symbols on the picture? I will comment on some of them. The ones on the top left, the orange diamond is the current implementation of software archive. This is just compressing every single file with gzip. If you change gzip with zstd, just a simple change by using a better compressor nowadays, you get the purple diamond on the top left. You see that the compression is about to 36 percent. You are able on this data collection to transform the original archive into one third of the original size.

You can do actually more, because nobody tells you that you have to compress every single file. What about just serializing all of the files, put one after the other, and apply a data compressor so that you try to exploit the similarities between the files? The first idea is to just write an arbitrary permutation, and see what happens.

These are the performances you get. The orange one is gzip. You see that gzip does not take advantage of this approach, because the window of gzip is very small. It's just one megabyte. You cannot take advantage of other files. Whereas the zstd is very powerful, because the window of zstdis even more than 60 megabytes. You see significant improvements.

You can do more. The red line that I have pictured there is what a BWT obtains, the Burrows-Wheeler Transform. I've shown that because in information theory the BWT is able to achieve what is called the kth order entropy of the data. It's a sort of lower bound to the best possible compression. This is in theory. In practice some things have to be done. The red line told us a few months ago what could be achieved, and so we decided that zstd 24% was not enough. We could go down.

How to go down, because the BWT is very costly to be computed? The first idea, according to suggestion by Roberto, was GitPack. Let's see what GitPack is able to do, and so the G symbol shows the performance of GitPack, that actually takes every repository and compresses every repository. It's very good. It is about 20 percent, it is on Python files, and is very fast to compress.
We sorted the files according to the file name. Every main.c comes together, and also took into account other tricks. We plug in other tricks, and we went down the red line. We arrived at 16 percent, so 1 over 8 of the original compression of the original data set. Then, Roberto said no because we don't want to use the file names. What can you do? We started to play with the content. This provided us with a collection of new algorithms. I don't want to enter in the details but for anyone who knows the locality sensitive hashing it's a very powerful technique that is used in near duplicate detection in databases, web crawlers, and so on. We took locality sensitive hashing in two flavors, simhash and minhash and then we experimented with several algorithms to cluster hashing fingerprints.

All these points in cyan show the performance of several variations of this approach that are going slower and slower but better and better in data compression. Up to the point where we designed the green triangle, that is actually very close to the best algorithm to date. They only use internal information about the file, but not file name, and nothing about coming from the graph.

What can we do in decompression? This is the new picture (slide). you see that GitPack decompression speed is 40 - 50 megabytes. Actually, our approach is much faster because we are around the 500 megabytes per second in decompression. This is a good performance.

Just some conclusions. If you take the C repository, it is much more compressible. The C repositories are able to go to 6 percent. Which is about 1 over 20 in the original data, up to the 25 gigabytes collection.

The final observation, the nice part that we discovered is that if we turn to the random access of single files after compressing, we can make blocks of 4 megabytes only. According to this, we are still able to achieve the same performance of the BackUp scenario by just losing only 1%. So instead of compressing in the case of Python up to, let's say, 16%, we arrive to 17% of compression. But we can decompress surgically just one single file.

A conclusion, we have two approaches nowadays (slide). One is called the Data-aware-PPC, in the sense that we use only the content of the file in order to compress them. We also have a Context-aware approach, we use file name, path, and so on. This is something that we will look at more in the near future, depending on the data that comes from the Merkle graph. The context information is very powerful. We are able, in some sense, to find source code that are very similar, according to our hashing fingerprints, and put close together to compress together. Which is very powerful. We are very robust with respect to the block size. Now we are using block sizes of 4 megabytes, but could even go smaller and smaller by keeping almost the same performance in compression ratio. We used zstd, but we tried broadly and zstd in this kind of scenario is actually better.

Finally, there are plenty of things to study, other orderings, other partitionings, and this is what we will try to do in the next feature. Also, what about parallelization and distribution of the computation? We have to attack 1 petabyte of data, not just 25 gigabytes of data. Our approaches are very simple to be parallelized. We are using sorting and there are plenty of sorters that are parallel. We are using a compression of individual blocks. This is not something that in some sense worries us. The problem is just to code it, in some sense.

Our hope is to squeeze everything into 200 terabytes and so make software advantage to spend only 4,000 euros for storing the data.

Finally, it's time to attack the real problem. This was a toy example, as we say in algorithmics, but it gave us a lot of information, and a lot of feedback to attack the most challenging and massive problem.

Mr. Roberto Di Cosmo
I would like to say how grateful we are for starting this collaboration. We need top notch research to help in this area. Just a minor correction, regarding this work he says Roberto all the time but Stefano is in the loop there. We are all working together and several people on the team are working on this.

The ENEA Software Heritage mirror (Mr. Giovanni Ponti, Head of Division for Development of Computer Science Systems and ICT at ENEA)

Mr. Roberto Di Cosmo
It was fantastic to have this research but not we need to have the real thing and the real thing is much bigger. How do we do this? This is one of the subjects of the next presentation.

Engineer Giovanni Ponti who is coming from ENEA who is telling us a little bit more what is going on for the creation of the first mirror of software Heritage that will be deployed in Italy in the coming months.

Mr. Giovanni Ponti
I'm here to talk about the ENEA Software Heritage mirror it is a new experimental experience and activity in order to support the Software Heritage Community.

First some words about ENEA (slide). Its a National Agency for New Technologies, Energy and Sustainable Development covering several topics: environment and sustainability, energy and technology, efficiency, and for fusion and nuclear safety. We also have a computing infrastructure and computing resources.

It's a very long story for our infrastructure (slide). ENEA has several research centers in Italy. Each research center is connected with an agreed infrastructure. That is our computing infrastructure in order to support any researcher activities.

There are several topics of research: climate, environment, combustion, nuclear energy, materials, and complex systems. Here there is a trend of growth of ENEA computational resources that is CRESCO clusters. The last installation is CRESCO6 cluster. It is Italian Tier 1 HPC infrastructure. Indeed, the Tier-0 is the reference infrastructure for HPC resources. We have a constant rate of improvement and of growth of our infrastructure. We plan with the findings that are national in the first one, European with the Euro Fusion community, and in the next, with the new installation in maybe this year or in the next one, CRESCO 8 with one of the PNRR projects.

ENEA wants to support Software Heritage in its activities, and for this we signed a cooperation agreement with Inria in order to provide a new mirror for Software Heritage. This provides a very important task for our ICT goals. It's a very challenging activity.

I'm going to describe it (slide). The mirror is not only a mirror of the Software Heritage code and structure. This is based on a new experimental solution regarding the storage strategy and the file system, especially. In fact, we want to provide the new mirror to improve project sustainability, to reduce the data loss risk, and to experiment to try new strategies in storage data.

As Paolo before said, his main issue is about compression data. We want to provide data storage space and strategy. In order to offer our infrastructure and our computational resources for further analysis. Our first installation, our first configuration is based on 8 CRESCO nodes divided in this way. We have 3 nodes, on which are running Docker swarm nodes. One node is the DB server for Merkel graph, and 4 storage servers for object storage activity.

The specific goal is an experimental one. The testing of a new phase system, that is a SEAWEEDFS file system. In fact, this file system is aimed to to be highly scalable in intelligence of performance, in order to also support a large amount of files that are very small. It is one of the main characteristics of the Software Heritage archive. Last but not least, to be a server file as fast as possible.

This is the main architecture of a mirror for Software Heritage and the Software Heritage based infrastructure (slide). We have two replayers. That is the one for object storage. It is for files. The second one is for GR for Merkle graph. These are properly stored in our infrastructure in order to provide the object storage activities, and also updates in the Merkle graph.

I will go fast on these, since this is the diagram of our architecture based on the infrastructure (slide). We have 3, as I said before, 3 Docker swarm nodes. One DB server based on POSTGRESQL to store the Merkel graph. We dedicated the two switches in order to connect this part of the architecture. In fact, facing the snapshot of the architecture of the Software Heritage structure (slide), we can see that the red part of the Merkel graph is storage in the diagram on the left side. Whereas on the right side, these are the nodes for object storage, based on SEAWEEDFS file system.

At this moment we are working to try to replicate the mirror so it will be in production by June of this year. However, we have completed the transfer of the Merkel graph in the storage of the POSTGRESQL server. We are working on the transfer of object storage, in order to completely have the mirror in operation in the early summer of this year.

However, our mirror, in order to support Software Heritage's activities, is only a starting point for other many important activities (slide). Especially in the field of machine learning and AI to support Big Code as Big Data. This is a very challenging activity. Especially with our situation in Italy.

The new site is in Bologna for Big Data and for HPC (slide). In fact, during this year and in the next one. There is a technopolo, a big structure of a big area in Bologna, that is now reused and re-engineered in order to support activities in computer science in ICT. Especially on topics on data mining artificial intelligence (AI) and Big Data.

Inria is planning to transfer a new technopolo and together with Software Heritage structure and mirror in order to be in an environment of growth and of activities that involves also many others research communities it is IFAB - International Foundation Big Data and Artificial Intelligence for Human Development, ICSC Foundation Research Center in HPC, Big Data, Quantum Computing, and the INFN Data Centre. In China with the new Leonardo supercomputer, that is ranking in the 4th position in the latest top 500 ranking.

"Here is the world software archive, Our data valley to serve people and to facing future challenges."
Stefano Bonaccini
President of the Emilia Romagna Region and Software Heritage (SH) Mirror

This activity is sponsored by Italian institution and government (slide). There is also a welcome by Stefano Bonaccini, the President of the Emilia Romagna Region about Software Heritage activities.

Roberto di Cosmo
Thanks a lot for this presentation. As you have seen, this panel has a clear coherence. We are going from long-term storage, to the need of reducing the size of archive, to the need of having copies for not releasing the data and massive model on top notch infrastructure to test all the possible algorithms one can use, to actually achieve these results. We are really looking forward to see the completion of this work. Let's cross our fingers, you never know what research can bring on the table.

Open Science Panel

Mr. Roberto Di Cosmo

It's a kind of a long marathon but I hope you appreciate the breadth and the diversity of the things that we are looking at today. Let's move to open science. We have a great pleasure in honor to welcome here today Mr. Karel Luyben, President of EOSC (the European Open Source Cloud) Association which is an instrument of European Commission for supporting open science research, Mr. Steve Crawford who is actually leader of the Open-Source and Open Science initiative at NASA and Mr. Bhanu Neupane who is overseeing activities at UNESCO on tracking open access in an open science policy around the world.

Mr. Karel Luyben, President of the EOSC Association

My background is in science and I have worked in the Delft University of Technology for many years and, ultimately, as the Vector Minifigures of that University got very interested in data about 2013 when I was involved in DTL Dutch Tech Center for Life Sciences in the Netherlands became chair of that and then became chair of the executive board in Europe of EOSC and then now I'm the president of the EOS Association

Mr. Steve Crawford, Open-Source Science Initiative Lead, NASA

I'm Steve Crawford. I am the current lead for the Open-Source Science Initiative at NASA. My background is an astronomer. I worked in South Africa for about 11 years on the Southern African Marsh telescope and then moved to the Space Telescope Science Institute where I manage the team developing the calibration software for the James Webb Space Telescope and then moved on to NASA where I help implement open science across all the Science Mission Directorate.

Mr. Bhanu R Neupane, Advisor, ICT and Sciences and Open Solutions, UNESCO

I've been at UNESCO for the last 20 years and before that I was a professor, like many of us, I think, in the room. I was teaching multi-criteria decision making at a university in the US and before that, I was a student at in Canada. I come

from Nepal, the least developed country in the world. I served with UNESCO primarily earlier as a hydro system engineer and I managed the Water Science programme. I graduated into a world of computers and then technology. I serve as an Advisor for ICT Communication Information sector and work primarily on Open Solutions. In a minute, I will be able to explain and what open solutions is all about.

Mr. Roberto Di Cosmo
We've been hearing a lot about open science. There is a drive here in France, for example the Ministry of Research has a National Plan on Open Science which is very strong. You are leading the EOSC Association in Europe. Can you tell us a little bit more about what is open science what is the point of view of the European Commission what is the EOSC Association doing in this aspect?

Mr. Karel Luyben
You called the EOSC Association an instrument of the European Commission. It was created with the help of the European Commission and in, let's say, coordination with the Commission, but it's not an instrument of the Commission. It is an instrument of all the stakeholders that are in Europe. That are stakeholders of the development mainly of what we call fair data rather than only open science.

For me open science is three lines of development, basically. One, is what we call open access, which is a misnomer because if you publish something it's open as long as you can buy the journal. Open is more free access than open access Second, is fair data the development towards findable, accessible, interoperable, and reusable data. These are principles, fifteen actually, underlying how to deal with your data or your software. If I use the word data, I mean this in the broad sense. Data, software and publications and not only the narrow implication of data. The third development is citizen science. Basically, what you see in that is also the development that we open up first and mostly to scientists, gradually to let's say the professionals in society, and ultimately to all of society as much as possible. These are overlapping terms but you do one more after the other.

Then there are at least 5 boundary conditions**.** You need to competences, you need the skills of the people for the data, you need the policies, you need, of course, need the EU infrastructures, and you need the engagement of society and the stakeholders, and others to get this going.

I tend to say, and I'm sometimes punished for that by the commission because they want to make it faster, that if we in 2040, something like 17 years from now, would have 50 percent of the relevant research data as fair as possible than I would be let's say very optimistic and hopeful. By relevant I mean those data that are considered relevant by the researchers and as fair as possible in 2040. Which is much fairer than now. It's not fair or unfair it's fair to a certain percentage. So does data and meta data comply with those all those principles are part of those principles.

EOSC is the European, let's say, box on top of a layer of fair data or a web of fair data we hope to create in the world where all the data are connected. We were talking about sending data. I think in the future we will see much more software visiting data rather than sending data because of the large data files and because of the large distribution of input you want to have. This is what the EOSC Association tries to reach for Europe.

Mr. Roberto Di Cosmo
Thank you very much for this discussion. Then we will discuss privately whether software is data or now. As you know, this is another issue.

Mr. Roberto Di Cosmo
Steven, we have all heard that in the US the White House has declared 2023 the Year of Open Science and the organization you are representing here which is a mythical one. I remember was I was 6 years old I was in front of my black and while TV and watching this incredible thing of going to the moon. You have seen here we care a lot about the source code but there it actually showed the way because that software is there because it was in the public domain by decision of the administration back in the 60s. Much earlier than we are talking about open source. Can you tell us a

little bit more about the point of view that you see in the US and what you are doing in NASA for pushing this agenda forward?

Mr. Steve Crawford
I was going to say you've already talked so much about NASA and sharing I don't feel like I have as much work to do here. Which is really fantastic. As you mentioned, this has been really exciting across the US government of being that 2023 is a Year of Open Science. It actually even started a little bit earlier in August with the release of the Office of Science and Technology Policy office's memo on ensuring free, immediate, and equitable access to federally funded research.

With the announcement of a Year of Open Science they're also released the open.science.gov website. I want to read it out because I think they've done a great job of defining open science, "the principle and practice of making research products and processes available to all while respecting diverse cultures maintaining security and privacy and fostering collaborative reproducibility and equity." That captures much of what we're also trying to do at NASA. Of actually trying to make sure that all of our research products and recognizing the wide range of different scientific projects, which are not just publications, but also data and software. Also, recognizing the wide range of different processes.

I love that you mentioned citizen science**.** That's a great way. Especially when we also talk about making those processes more inclusive and actually getting a much greater breadth of the community and the world involved in science. One of the things that we're doing at NASA is the Open-Source Science initiative (OSSI). We're adopting from open-source software not just actually making the science open and the products open but the process of science open as well. To make it more inclusive. With that, the Open-Source Science Initiative **(**OSSI) is a 20 million dollar opportunity hitting on a number of different things: improving open science infrastructure, policy at NASA, open science funding (so directly funding groups in the community), and also open science community building. For that, we have transformed our open science initiative where we're looking to train 20,000 scientists on open science practices, to increase historically underrepresented groups in open science, and accelerate major discoveries through adoption of open science.

Mr. Roberto Di Cosmo
I have to say it is pretty exciting to see what we are doing there. I hope this kind of movement that was underground for a while in the past years is actually going up right so it's being recognized. Maybe even you want to jump on that because you have been observing the policies and open science open access around the world. Can you tell me a little bit more what we have seen?

Mr. Bhanu Neupane
In fact, we have to somehow press a rewind button before we start talking about it. Primarily because in back in 1997 International Cyber-University (ICU) and UNESCO got their acts together and they thought that perhaps the knowledge should become open. This thing was passed through the Tunis Declaration in 2003. When a World Summit and Information Society (WSIS) first was conceived. They say that because data, information, knowledge, software, and everything as part of the broader human knowledge it should be better leverage to power development.

We went on with that one until the Millennium Development Goals (MDGs) was kicked in and it came to an end. It actually concluded one of the things that the Millennium Development Goals (MDGs) failed was because the science was not properly leveraged within the broader framework of development activities.

Now we are in the Sustainable Development Goals (SDGs) days. In 2015, the world decided to move on and brought science within the overall in a broader realm of sustainable development. One of the things which is quite fantastic is out of 17 goals, 19 goals require information sent to them almost in real time. That requires 3 things: openness of visas, something that Karel very rightly mentioned, openness of data, and other tools that will enable that thing to happen.

In 2020, things slightly changed when the pandemic hit, and the world decided that it's not enough. I hope many of you know that the UN has already started to look beyond Sustainable Development Goals (GDGs) and started to talk about our common future and our common agenda. In 2020 the UN Secretary General, António Guterres, coined that we need to talk about outcome and agenda, and that this common agenda will go for another 25 years. Quite interestingly, if you browse this agenda, data drives research as well as the software runs along.

Within these communities, as much as data and research is being talked about as a vehicle to drive development agenda forward, software, unfortunately, it's not. I think this community has a lot to contribute to this overall discussion that has happened.

Mr. Roberto Di Cosmo
The basic principle was many, many years ago now the big question is how do you turn words into action, principles into reality. That's a difficult part. Actually, you are talking about working together. When we started with Software Heritage in 2016, the question was what are the big challenges here?

If you look at the issues related to open access what you see that in when the internet arrived and was popularized at the end of the 90s everybody was supposing that the problem of publications was solved. Just put your paper on the web somewhere. That's it. Actually, this failed miserably, I believe we didn't do it properly. I'm putting myself in the loop. I'm old enough to carry part of that responsibility. Today, 25 years later, we are still fighting to make this available and then we didn't manage to build the common infrastructure. We actually had tens of thousands of open repositories that are spread around but not coherent situated.

Then you have the data. You have a lot of data silos here and there. Now we are trying to open the silos by putting them together. It's a long story. When we started Software Heritage, we actually said that we are lucky that nobody cares about software. We had one positive thing. That nobody cares means you have a blue sky strategy. Maybe we can try to build a common, shared infrastructure and avoid replication, useless duplication in coherence. This is what we're trying to do here and this is why all these people around the table are here. We hear you okay you and know that you want us to get into that direction. Now the big challenge is after you see what we try to do at Software Heritage we will be able to go to the next level? That is to say to create a global organization where every stakeholder can use this and own this infrastructure without making tons of copies around the world.

Long phrase to get to the questions here to Karel to Steven, etc.

How do you see the possibility of making this dream a reality? I mean, take advantage of the fact that the software was not visible and maybe take this as an opportunity to create a joint universal common shared infrastructure. For once, after failing for data, after failing for publications.

Karel, your thoughts?

Mr. Karel Luyben
Easy question? Earlier you said that an estimate is that 1 percent of the cultural heritage is publicly available more or less. An estimate is that at the moment less than 2 percent of the research data can be found online, so openly accessible. If they are accessible at all they can maybe be founded and they are not accessible but less than 2 percent is the guess.

She also said that there is a lack of positive incentive in order to do this. The same holds in the research domain. We don't have enough positive incentives for our researchers to share the data; to be rewarded for sharing data; to share their software; to be rewarded for sharing the software; and, especially, to get the meta metadata and all the information needed in order to get the software there. If we don't change that culture, if we keep on rewarding people for the wrong things then this is not going to happen. You're lucky, you're right. In software there was less than in the other domain. So, in that sense you chose the right domain.

With respect to the European Open Science Cloud (EOSC), I would be very happy to see an enormous repository or a network of repositories, if needed, that we could hook up to the European Open Science Cloud (EOSC). For me, software is not data, software is just software, and it needs special treatment. If I use the word data, I also a mean including software. Otherwise, we have to repeat all these words. Basically, what we could do is talk about digital objectives. Those people coming from research, and they can be software, they can be data, they can be publications.

I believe in your dream and, as I said, I believe in my own dream. Let me give you an example 35-years-ago I was a professor at Adults University of Technology with PhD students working on the drying of gelatin as a material to mimic food. We were working on the drying of foodstuffs. I made a beautiful piece of software, working beautifully, but I had hardly time to test it with data because I needed a lot of time for the software and had limited time for doing data sets. If I would've had the data from colleague research available, and there were hundreds of them at that time, drying in the world took 15 percent of the world's energy at that moment in time. Not today anymore, by the way. Then I could have mimicked different drying conditions in different equipment with their data sets. How much could I have progressed if I would have had access to that?

This is just a small example of what can be done if you share data properly and with a 100 people in the same domain you have a 100 data sets rather than one data set of your own. It's as simple as that but how to do it? Who will be the first? How am I going to be rewarded for it? Does my boss like it? There's where the problem lies. I think we have to change the culture in order to get it going.

Mr. Roberto Di Cosmo
Steven, you are changing the culture for research?

Mr. Steve Crawford
Well, we hope so. I think that is actually the one of the goals of the Open-Source Science initiative (OSSI).is to help change and transform the culture. I think one of our principles has been that we want to make it easy, and for the sharing to be a part of the research process. Some of that is going to be a mix of different things. Of investing in technology that makes it easy.

Fortunate enough there is wide range of technology that has appeared over the last 10 and 20 years that makes sharing easy. There's also a large number of things that actually make it automatic as well. I think a great example of that is the linking between GitHub and Zenodo. I can put my code up on GitHub and then when I release it gets archived automatically in Zenodo and a Digital Object Identifier (DOI) is produced for it. Then likewise, it will show up in Software Heritage. If it's up on GitHub it's not something where I as a researcher have to think about it. It just becomes part of my normal research process. Making it very easy for people to share their data and their information.

The other thing is we do set policies as well. Sometimes the best pass forward is to mandate it. For all of NASA data we have said that, especially for our mission data, that it must be freely accessible. We're actually saying it should also be released under a Creative Commons license so that anyone can use it, then build off it, and do the science that they want to do, or reuse it for other purposes. Whatever that might be. If it's in fashion, in arts, in outreach, and citizen science, at any wide range of different things.

We're also going to be requiring for our future researchers that they release their software as well. Software is such an important part of earth and space science research. It's such a critical part and has always been such a critical part of what NASA does. To understand that research, and all the nuances which are captured in papers and other aspects, you need that aspect which is captured in the software as well to fully understand to reproduce it, but also to actually enable reuse of it. That we can actually have more benefit from it.

The third part that we're doing is training. A lot of this is new to scientists and new to people. So, making sure they have the time and the space and going to them to engage with them, so that they can actually understand hey the benefits of sharing: that there's more citations, there's more reuse of your software, and there's better access to it. Also, how easy the technology today makes that. Those are three different ways to hopefully, potentially help transform the culture which is definitely a difficult problem.

Mr. Roberto Di Cosmo
We are looking forward to that.

Mr. Roberto Di Cosmo
For Bhanu, after listening to all this, do you feel a bit better than a few years ago?

Mr. Bhanu Neupane
I essentially like to take from Steven and Karel here. When they were making their intervention, I basically thought about UNESCO's Open Science Recommendation, which was agreed in November last year, 2021. That actually changed people's way of thinking about how science can particularly contribute to development processes.

I also look at, like Steve just mentioned, that NASA has spearheaded you know a USD $25 million dollar program. Yesterday, I was just looking at a program that DRC or Democratic Republic of Congo has spirited for themselves and, sadly, they were looking for USD $4,300 just to put one repository together. We are in fact looking at a world that is not…, I basically want to be you know as politically correct as possible.

If we are to move to a different level of participation and bring the culture of sharing and openness into the broad discussion of science, we have to create a level playing field for everyone. How do we do it? We have to create incentive structures around the world. We have to also create some kind of knowledge sharing process that in something that Steve does with his USD $25 million dollars can go and benefit something that the scientists in DRC are doing with their measly USD $4,300. I'm not too sure how that can be done. I think that the world has already has developed a mechanism and we will know in 4-years-time because this is how UNESCO says that you'll know. That there is a recommendation, the world has adapted UNESCO's Open Science Recommendation, but the countries will have to report back to UNESCO, as far as what kind of progress they have achieved.

It's essentially a monitoring exercise, but this monitoring exercise requires very soundly grounded tools. I'm not too sure whether we have it. It requires processes. I'm not too sure whether we have it. Also, contexts because we have to somehow be able to compare the kind of open signs, the open software movement that happen in NASA was fitting, and also is compatible to something that these scientists essentially had done in DRC.

Mr. Roberto Di Cosmo
You have heard a lot of the different things that are going around in open science here and different initiatives. I'm very optimistic after listening to. I agree with Bhanu, we need to monitor. If you cannot measure, you cannot improve.

Perspectives from industry and public administration

Mr. Roberto Di Cosmo
Now we are moving to last part we are giving the floor to industry, to open source, community, to government organization to tell you a little bit more about what they think about software about source code, it's preservation and the value that it has.

Open Source and Open Ecosystems (Mr. Alexios Zavras, Chief Open Source Compliance Officer, Intel)

I work at a very large company that you've probably heard of, it's Intel. We are mostly known for hardware but maybe some of you do not know that we're also doing lots of software. We have around 20,000 software engineers working for us. We're doing everything, and almost everything also contains open source. It should be no surprise to any of you that software nowadays is so complex and complicated. Open source is everywhere. In every product, in everything. Every complex software system is an amalgamation of different components and most of the components that are being reused are under open source licenses.

In the industry we use this wonderful 80/20 rule, the new Pareto Principle, in which every software produced, 80% should not be yours, should be something that is beingreused. You should be focusing on the 20 percent because that's where your competitive advantage is. Nowadays it's more than likely moving towards 90/10. Nobody wants to do more than 10 percent of the final solution.

Intel has a very long history in open source. We've been working on open source for more than 2 decades now. Intel used to have a dedicated Intel Open Source Technology Center that was focused on open source, but, again, because of the change, everything now contains open source. Open source is everywhere. Nowadays, every business unit, every product contains open source. So, there is no need for having specific engineers doing open source things. All software engineers doing everything in any of our products are using open source.

We are also very much contributing to open source and releasing things as open source. We define internally different levels of engagement with the open source communities and ecosystems. I will not bore you with the 5 different levels that we have, starting from just putting the source out there, which is actually not allowed because you have to provide some support, at least. And going all the way up to the level where the software is actually governed by the open source community and it's not yours anymore. The community decides where it's going to go.

More in line with our work and our discussions here today, one of the issues that we are facing is about long-term retention. Nowadays people work in open source and work online and they think "this is the new wonderful new version that appeared, and we should all go and change and use the latest one." We produce products that have very long lives. We ship software that gets inside train engines or satellites. It's not easy that you're going to say "a new version appears, let's go update everything to our latest version." When you have a satellite up there or a train engine that has a working life of decades, you want to have access to the source. Asking engineers what was the exact source that was used 20 years ago, the engineers will not be there. We're living in a real world in a very large company. Business units disappear and everything is being restructured. There is no way that you can easily have this kind of business continuity.

That's why we have worked together with Software Heritage and right now we are integrating, updating, uploading, and using Software Heritage for every piece of our software produced at Intel. Every piece of publicly accessible software for now. We'll see about the proprietary parts later. Definitely this provides us a solution because, again, we have this need for a very long term retention.

Everything that I've talked about is nothing specific to our company. All industry is facing exactly the same problems, all industry is facing exactly the same issues, and all industry is looking at the same solutions. Fortunately, because everyone in the industry does not think that this is a competitive advantage, everyone wants to work collaboratively. So, we are working together in implementing such things as software archiving. The way to actually produce some innovation in all this is only from working openly together, facing these issues.

Mr. Roberto Di Cosmo
Thank you, Alexios. You show us how we highly efficiently optimized the hardware in a company can deliver a super, efficient presentation and talk.

Building shared infrastructures (Mr. Gael Blondelle, Chief Membership Officer, Eclipse Foundation)

I'm really amazed to see how Software Heritage is at the crossroads of science and open source.

I'm representing the open source side of it. I'm working for the Eclipse Foundation. I'm the Chief Membership Officer of the Eclipse Foundation. I can tell you that my Executive Director would have loved to be here today but well he was marrying his son last weekend, so that's some priorities. I'm talking about Mike. I think it was in 2014 Paris Open Source Summit when you came to us explaining the Software Heritage project for the first time.

We have been very supportive of Software Heritage since that day because the project is an exciting topic and that resonates with us as developers of open source.

Before I go further, I don't need to try to tell you a bit more about the Eclipse Foundation. We are one of those large open source foundations we have more than 400 projects in plenty of different topics. Of course, the development tools Cloud, IUT, Edge, Automatic are there. We do open source for developers on one side, and we also do open source for the industry. I come back to that, but we have we have lots of connections in my opinion.

The Eclipse Foundation was established 19 years ago in the US, and two years ago we finalized our moved to Europe. We claimed to be the largest open source foundation in Europe, operating under European law, and with the capability towards code in Europe. However, we are a global organization, and we really want to stay a global organization.

Open source is the foundation for modern software. Almost all companies are using open source. It really enables the collaboration of hundreds, even millions of people as part of a successful open source ecosystem. It's interesting to see that as software was eating the world, open source is really now also the foundation for modern business.

The mission of the Eclipse Foundation has always been twofold. Our mission is to foster open source projects and communities, and that's certainly where we have the closest connection, and to foster the commercial ecosystem based on our project.

I also wanted to mention the 90/10 rule. Yes, 90 percent of the software is open source to today. Companies try to focus on the 10 percent that really matters to their users and customers. That also means that open source became the main collaboration framework, not only for society, for volunteers, but also for the industry. Because the freedom to run, study, modify, and redistribute software is a very efficient way to save costs, reduce time to market, and to spread the innovation.

Whether it's for business reasons or to advance Sustainable Development Goals (SDGs), open source enables the free flow of technology, and that's really what enables global collaboration and global innovation. That's where, as an open source foundation, it really matters a lot to us. Whether we are based in Europe, but we work with members and developers all around the world. That was interesting when we moved our events to being virtual events during covid, twe got lots of people attending the event from regions that we didn't meet with before. That's also a very important.

Open source is a simple way of enabling collaboration because the producers decide to publish their software under a specific license, the users consume the software under that specific license, and that's the only moment when you just have to deal with the contract. Which is three pages long and almost everybody all around the world understands the contract in almost the same way.

We were together in Brussels on Friday. In Europe we love we talk a lot about digital sovereignty, but today, in this context, I want to share the following point of view. That it's really important to open source because that moves the needle from having access to proprietary intellectual property, to having to grow the skills to understand use, master, and evolve open truth. This is something that can be done everywhere in the world. You just need a simple laptop, not a fancy MacBook. You just you can start with very simple things and that applies to every individual, every

student, every company all around the world. That's why at Eclipse we try to onboard more of those individuals and people everywhere.

One of our roles at Eclipse is also to ensure that when people do open source, they do it with open source values and best practices in mind. We think that there are three values: transparency, openness, and meritocracy. Those core values are really deeply incorporated in our processes are what ensures vendor neutrality and good collaboration.

With your archive you enable people to build on top of everything that has been done since the beginning and that's certainly important. For example, we have tools like the Eclipse Development Tools, like new cloud-based dev tools and we hope that some plugins will use Software Heritage sooner rather than later to enable a new ecosystem of tools. As we were mentioning, license compliance with your unique identifiers.

That's something very fundamental for license compliance. Because you identify your file, and you know if it's compliant or not. Security with an impact analysis; it evaluates the time required to solve problems; also, to spot components that become a single point of failures.

At Eclipse we have to ensure the sustainability of the projects that are under our stewardship. We have all the process to keep projects active, as long as a group of people are interested. It's really a relief to know that when we terminate a project, when we archive a project, at least we know that you keep it, and we can bring it to life later. Which is also a very important feature.

Thank you for showing the vision and for being an inspiration for all of us.

Software reuse for public services through source code sharing (Mr. Gijs Hillenius, European Commission Open Source Programme Office)

Mr. Roberto Di Cosmo
Now, from the point of view of the Open Source Program Office of the European Commission.

Mr. Gijs Hillenius

We're in the company of rockstars, Roberto and Stefano. You guys have the long-term vision that we really appreciate and really need. To give a bit of perspective from the European Commission on how we see this, how we appreciate the work of Software Heritage, and how we plan to work with Software Heritage.

Many of you will have seen this slide before this is just a history of open source at the European Commission that goes back to the late 90s (slide). It has grown from infrastructure, to using it basically all over the organization, like everybody else, the 80/20 or the 90/10 rule is also at the Commission.

In 2020, we decided to take the next step and the European Commission and took all of its open source strategies, put them together and made it a forward-looking open source strategy (slide). Which was a communication from the Commission to the Commission. Which meant that it went out of Directorate-General for Informatics (DIGIT). It became something where the Commission realized that everything, almost all of political goals, depends on IT, and almost all of IT depends on open source.

This elevated the vision of the Commission on open source and this strategy came with an action plan. As you can read them here (slide). That's what the Commission also is for, to make sure that these 10 action items are carried out.

Regarding software development, one of the first things we did was blow up the internal barriers, because inside the house developers were working in teams. These teams were kind of isolated. These days, if teams start projects the source code is immediately available to all of their colleagues. We're encouraging actively and passively these teams

to switch from the old system, where they wouldn't share, to the new system where they do share. The goal is to create an internal working environment that is entirely based on the principles of open source.

What we also did is work with the internal colleagues who are hosting the software development factory to make sure that there is a place for all the open source software that is going to come out of the Commission. That's what this Roadmap slide shows (slide). At the bottom there is our code repository, code.europa.eu, which was unveiled in September. Where we now have 200 projects from Eurostat, the European Central Bank, the European Data Protection Supervisor and, of course, also Directorate-General for Informatics (DIGIT) ourselves. We have some 700 users. This is where people can work entirely in the open. The goal is to make more and more of Commission inside projects go to the open part of our development factory.

To get this going we removed an internal barrier, which was there for a very long time. Which created an enormous amount of paperwork for the Commission to share open source code, to share source code is open source (slide). This barrier is basically now gone. Projects can decide on their own that their project is useful for other public services. That's mostly our target audience. Or if it should be public for the citizens. Then they can decide to open it there. With a few checks and balances, they can go open almost right away.

What I believe is important with this discussion today is that we want to work with Software Heritage to make sure that A. Our own code repository is there because we share the long-term vision, but B. We want to help Software Heritage get recognition in all of the Member States that we are in touch with. That's the OSPO network.

We're establishing an official network that works with all the national and regional and, maybe, city level in public services across the EU. With other OSPO-like organizations will have regular meetings, we already have regular meetings in person as well as online. Wherein we figure out things like code repositories and where we will also be promoting the work of Software Heritage.

Code development is one thing, but we're also always looking for ways to improve the sharing and reuse of existing solutions (slide). There is a European Parliament pilot project, FOSSEPS - Free and Open Source Software Solutions for European Public Services - EU solutions catalogue/European Open Source Application Catalogue, that we're currently involved with, which is up trying to create an almost automatic European open source solutions catalog. We're in not in the final stages, but we're really doing our best to get this thing out in the open soonish. This is also where we've been looking very hard on how we can use Software Heritage, and use, for example, their unique identifier.

Software Source Code as a key for Innovation (Mr. Florent Kirchner, SGPI, France)

Mr. Roberto Di Cosmo
It was impressive to see this evolution. Also, at the level of the European Commission, it’s really welcome for us to see this kind of change. I would like to give the floor to Mr. Florent Kirchner, who is representing the General Secretariat for Investment (SGPI), Secrétariat général pour l'investissement (SGPI), here in France. Giving his view on what is going on here.

Mr. Florent Kirchner
I recognize many faces. Both the rockstars and the coal miners.

We've seen over the morning how many, and I won't restate those challenges, that we're facing globally. Not only in terms of software development and open source advocacy, but, overall, the questions that we're facing both in terms of societal changes we see in the world around us. As well as ecological, social, economic challenges, technological, industrial, cultural, etc.

We've been talking a bit about culture this morning and it made me very happy. We see those changes coming up as much quicker than what we've seen happening over the past few years. In some sense it’s a little bit scary. In other

senses, it allows us to pack what once fit into a 747 hard drive, now into something that's not even visible to the eye. We're now seeing this both as an advantage, but also as something that challenges us in terms of addressing those new questions around climate change, around conflicts and geopolitics, around the pandemics, around how we manage our digital worlds, and the way that they impact society.

I work in an organization in the French public administration that's called the General Secretariat for Investment (SGPI). The idea of those investments is to give us exceptional means to face those transformations, and not only to face them but also to harness them. To see how we can make this part and how to be a part of the solutions. Rather than just having to face those and, hopefully, survive from them. Those exceptional measures can take and should take many forms and should concerns many domains.

In terms of domains, I want to focus this morning on the question of securing digital artifacts, and building trust in those digital artifacts. Building trust in the fact that we are in control of that heritage.

I really like Roberto's early slides saying, we thought we had things under control until that source code disappeared. Building back that control is a fundamental area where we need to look at, as a community. Where public administration is willing to have a look at that with you, and see how we can move forward.

Trust in those digital worlds encompasses algorithms, software, and data. Putting people in control of those artifacts should come first in those in those areas. Building around this talent, diversity, and culture is something that we want to look at exceptional measures to encourage this type of development.

This does not stop at that, and this does not stop at digital in the in the conventional way, of course. We should look at artifacts, we should look at algorithms, source code, C code, even, God forbid, python.

Then we also need to have a look at, more generally, how we as a society face those changes that encompasses culture, and not only development culture. This means, how do we rethink performing arts; how do we rethink movies,]; how do we think about architecture and preserving designs; how do we support museums through all those transitions? That scope is huge, and, to be frank, it's both a little bit daunting and a little bit exciting to be in those places.

How do we do this? What types of exceptional measures do we take? First of all, we support emerging ideas and emerging approaches. The ones that we've seen this morning are very good examples of this. We've been concretely supporting the MoleculArXiV project, that you've seen previously, both financially and in terms of making sure that the right connections are being made and the right exposition is happening. We need to be bold about how we think about those ideas, and we need to leverage everything that we can in terms of collaboration, helping literacy, helping out diversity, and building diversity in those collaborations, and helping develop trainings.

I think about normalization and how it can contribute to the overall picture. I think about legal aspects we've heard about this morning, but also in incentives. Those things don't spring and come up spontaneously. So, thinking about how we invest in those different levels is something that is very key. Of course, we do this both humanly and through processes, but also financially.

The program itself is called France 2030 and it manages 54 billion euros in investment over 5 years. Of those 54 billion euros, the first year, which was last year, invested 8.4 billion euros. Now you do the math, but there's quite a lot of significance left in terms of capacity of investment. So, through these we support new developments.

What's exciting about Software Heritage and the reason why I'm very happy to be here is that we see that software as an enabler and the archive as an enabler of human endeavors. It can be a playground for many of those developments. We've seen the initial thing that comes to mind as nerds is, oh can I contribute to the infrastructure, to the core of it and I love it. I'm including myself in this. I want to see if that compression algorithm really works. Hey, can I have a look at those base 4 encodings? Those types of questions are really good.

The core infrastructure the tools that surround them and the new tools that Gaelle mentioned a few minutes ago are really exciting things to have a look at. Also, this should spawn new applications and new ways to think about how we do research. How we do science. Not only how we include this in our in our practice, but what types of new fields are happening.

You're creating an archive, you're creating a museum, you know you need an archivist, you need an archaeologist, maybe, to have a look at those new programmes can we support, can we encourage to create those new areas of work around the world to invent the new jobs of tomorrow. I want to be a software archaeologist one day and you know have a have an Indiana Jones hat to go with that.

The final point for me is, this is my call for action. Have a look at what this enables, in terms of new fields of new disciplines. Have a look at this and come and talk with me about what you want to do in terms of sociology, for instance, of development. Archeology is one of the leading examples. Have a look at how we can lead the way through and reach new societal goals in France, in Europe, and in the world. Let me know how we as a public administration, can help support you as a community.

International Working Group

Mr. Fackson Banda
As you know we've been working with Inria for quite some time now and within UNESCO itself we have been repositioning ourselves in relation to how we programme ourselves around *Software source code as documentary heritage for sustainable development*. Because of that development and in trying to build on the main discussions that we have had which resulted in the Paris Call: Software Source Code as Heritage for Sustainable Development, for example. The symposium that we've been having, the summits that we've been having, and given the one issue coming up all the time the need for international recognition of software source code as an enabler for sustainable development. We thought that, perhaps, we could come up with something a little bit more structured around this issue. We propose and we might come to some of you later your own to consult. We propose to set up an International Working Group on the promotion of software source code as documentary heritage for sustainable development.

We have developed a draft concept that tries to drill into what we really want to do with this international working group but it will have a very narrow mandate which can be captured in terms of these three things. Three things, which are more related to what the Memory of the World (MoW) programme does. As part of this programme we have what we call the Memory of the World (MoW) International Register. It's a listing of items of documentary heritage that are regarded to be of world significance. World significance is defined in the foundational of documents of the programme. So based on that and given the need to drum up this international recognition.

We propose that this International Working Group focus on three things. First, identify landmark software which could be proposed for inscription in the World International Register, as a way of calling attention to its existence, world significance, and need for preservation of such landmark software. Secondly, propose for possible inclusion in the companion to the General Guidelines of the Member of the World (MoW) programme criteria for inscription of software source code as a way of enhancing existing criteria for digital documents in general. Thirdly, explore the possibility and desirability of setting up a more permanent organizational structure dedicated to providing support towards the identification, preservation, and promotion of software source code as digital documentary heritage for sustainable development. Related to this might be an exploration of a series of activities that could potentially be undertaken by such an organizational structure, including the possibility of a global computing museum.

Those are the three objectives and we will be reaching out at some point to some of you. We are going to exercise executive autonomy in setting this up. So, it's not entirely up to you, but we will seek your support as we go along.

Mr. Roberto Di Cosmo

It's really important to see that UNESCO takes all software as a new, valuable part of our cultural heritage. I salute this initiative.

Conclusion

Mr. Roberto Di Cosmo
A few words of conclusion. We are all together because we trying to realize a kind of a Grand Vision. We really would like to bring together all facets of society, from academia to industry, to governments, to other specific society. The kind of a vision of building a global infrastructure which is in the service of societies of the world, building better software for a better sort of warehouse. What we do provide is the first brick for this initiative. Which is basically a common shared infrastructure, which is vendor neutral, which is open source, which is non-profit, which is from the beginning thought as a worldwide initiative and long-term initiative. It provides many, many components already today: archive reference, integrity, etc. We are acting as a catalyzer. We need all of you and all the people you can reach out to, to make this a reality. Let's try to work together and build together an infrastructure we can all use and share.