EPJ Web of Conferences **245**, 08020 (2020) https://doi.org/10.1051/epjconf/202024508020
CHEP 2019

A Software Institute for Data-Intensive Sciences

Joining Computer Science Academia and Natural Science Research

Ian Bird[1,*], *Simone* Campana[1,**], *Pere* Mato Vila[1,***], *Stefan* Roiser[1,****], *Markus* Schulz[1,†], *Graeme A* Stewart[1,‡], and *Andrea* Valassi[1,§]

[1]European Organisation for Nuclear Research (CERN), Espl. des Particules 1, CH - 1211 Meyrin

Abstract. With the ever-increasing size of scientific collaborations and complexity of scientific instruments, the software needed to acquire, process and analyze the gathered data is increasing in both complexity and size. Unfortunately the role and career path of scientists and engineers working on software R&D and developing scientific software are neither clearly established nor defined in many fields of natural science. In addition, the exchange of information between scientific software development and computer science departments at universities or computing schools is scattered and fragmented into individual initiatives. To address the above issues we propose a new effort on a European level, which concentrates on strengthening the role of software developers in natural sciences, acts as a hub for the exchange of ideas among different stakeholders in computer science and scientific software and forms a lobbying forum for software engineering in natural sciences on an international level. This contribution discusses in detail the motivation, role and interplay with other initiatives of a "Software Institute for Data-Intensive Sciences", which is currently being discussed between research institutes, universities and funding agencies in Europe. In addition to the current status, an outlook on future prospects of this initiative will be given.

1 Introduction

We discuss a new idea of a "Software Institute for Data-Intensive Sciences" [1, 2] to connect software and computing activities in natural sciences with academia, such as university institutes and engineering schools concentrating on fundamental research in those fields. In Section 2, more details on what we think the current problem is are given, together with a list of current initiatives that concentrate on software and/or computing in natural science. Section 3 discusses the idea of the institute in terms of main ideas, goals and guiding principles. Finally, Section 4 outlines the next steps to reach the goal of setting up such an institute. A summary in Section 5 concludes this paper.

*e-mail: ian.bird@cern.ch
**e-mail: simone.campana@cern.ch
***e-mail: pere.mato@cern.ch
****e-mail: stefan.roiser@cern.ch
†e-mail: markus.schulz@cern.ch
‡e-mail: graeme.andrew.stewart@cern.ch
§e-mail: andrea.valassi@cern.ch

In this paper, we use examples from the high energy physics (HEP) field for illustration. However, the scope of the institute shall span over all areas of natural science and computing.

2 Problem statement and current initiatives

Data processing using software is a fundamental part of today's scientific research. Planning for future experiments, e.g. in HEP, shows that significant improvements in software are needed to fully exploit the physics potential of new detectors within a realistic computing budget [3]. This is compounded by a rapidly evolving hardware landscape that needs to be adapted to. In addition, according to a study from 2014 in the UK [4], 92% of academics use research software and 56% develop software. Despite these numbers, software engineering for science has not yet gained a high reputation in the academic world, putting the careers of scientists who engage in software engineering at risk. Another article [5] suggests that only 8% of scientists scrutinise the software they use and that "Most scientists [...] continue to emerge from natural science training without formal training in computational methods and software development and/or engineering."

Several initiatives, especially in the field of high energy physics, exist to address one or more parts of the above-mentioned problems. A non-exhaustive list of those initiatives is:

- The *HEP Software Foundation (HSF)* [6] is a relatively newly established forum in high energy physics that aims to bring together researchers of various experiments in order to build a community and establish a discussion forum on central parts of software and computing, such as simulation, data reconstruction, software frameworks, but also more general topics as licensing or software packaging.
- The *World Wide LHC Computing Grid (WLCG)* [7, 8] steers, develops and operates the distributed computing infrastructure used by high energy physics experiments, such as the LHC experiments. It is composed of representatives of experiments, computing sites and operations teams.
- *CERN openlab* [9, 10] engages with industrial partners to test the latest hardware developments in a scientific environment. The partnerships usually last for several years and may also include funding of research personnel.
- The *IRIS-HEP* [11] institute is a US based institute on software engineering and computing. It employs personnel to engage in R&D on various topics and also works on building a community in high energy physics software and computing, e.g. working with the HSF.
- The *Research Software Engineers Association* [12] is a community effort that started in the UK to raise awareness for the job of software engineers working in natural sciences. The association has partners in Germany [13], the Netherlands [14] and in the US [15].
- HEP computing schools, such as the *CERN school of computing*, *CERN topical school of computing* [16] or *Bertinoro school* [17] aim at the training of PhD students and early career researchers in software. Teachers in those schools are renowned experts within high energy physics software and computing or experts from industry.
- National or EU funded projects, such as *ErUM-Data* [18] or *AIDA2020* tackle specific topics in high energy physics with a short to medium timescale. Examples for such topics are trigger and reconstruction algorithms, common software frameworks, usage of compute accelerators for data processing, etc.

Despite these activities, collaboration on software and computing in natural science and academic institutions, concentrating on R&D in those fields, is mostly left to ad-hoc connections made between individuals. Those connections may or may not lead to sustained collaboration or even to R&D work.

Those ad-hoc initiatives connecting the two fields of science are most welcome and will continue. They often happen by chance and success is highly dependant on the expertise and engagement of the people involved in the follow-up on possible collaborations. The main question we try to answer is whether it is possible to set up an environment and structure that helps organize such connections between natural science and academic computer science.

3 The institute

Rather neglected areas of potential resources for future R&D in natural science are computer science departments and software engineering schools. Researchers in those facilities are developing new ideas and paradigms in software and computing that can benefit the software and computing in natural sciences such as high energy physics, biology, medicine, etc.

This paper proposes to set up a structure to establish **interaction on computer science and software engineering in natural sciences and academic institutions**, such as computer science departments or software engineering schools, in the form of an **institute for scientific software in data-intensive sciences**, described also in a one-page statement [1].

3.1 Idea and guiding principles

The main idea of the institute is to tap into the research knowledge of computer science departments and software engineering schools and make it available for natural science developments. The institute shall obtain, curate and disseminate the so obtained knowledge. Furthermore, it aims to:

- Enable R&D resulting from collaborations of computer science and natural science;
- Help establish a career path for scientists and engineers working in software and computing in natural science. E.g. in HEP the recognition of software work and developing and retaining experts is a major concern;
- Cross-fertilise knowledge between different science domains and make the acquired knowledge available across domain boundaries;
- Act as a lobbying organisation and raise awareness of software and computing in natural science.

The setup of an institute will have positive aspects, but also some risks that need to be addressed for both natural science and computer science as described in Table 1.

The institute can work in a complementary fashion to already existing initiatives, as described in Section 2. E.g. it can collaborate with CERN openlab in the context of a scientific / industrial partnership, it can collaborate with IRIS-HEP on a geographical level in a US / European partnership and work with the HSF and WLCG to receive ideas of current problems in high energy physics software and computing. It can furthermore be a partner in R&D projects with special funding and provide input for lecturers to computing schools such as CSC or Bertinoro.

3.2 Organisation and process

The main stakeholders of the institute are:

- **Natural science research laboratories and university institutes** employ people working in scientific collaborations and communities and provide the necessary infrastructure to develop and operate scientific instruments. They will also provide infrastructure to host

Table 1. Potential advantages and risks for the partners in a software institute

Advantages	
Natural Science	***Computer Science***
Inject new ideas from academic research into scientific software, leading to new collaborations	The awareness of new ways to operate natural science experiments and problems specific to natural science fields can lead to the new research in computer science [19, 20]
Open new career path opportunities for scientists and engineers through collaboration with academia	Data gathered within natural science such as scientific or operational data may be used for research work
Raise awareness for the importance of software and computing in natural science via increased visibility and lobbying	Enable new possibilities for supervision of Master and PhD students enrolled with universities
	Benefit from existing programs to fund master and PhD students within natural sciences [21, 22] and supervise those students in the context of the university which will also grant the degree
A stronger and more diverse collaboration across institutions, countries and domains will result in benefits for grant application and lobbying	
Risks	
Common R&D work may fall between the cracks of both science fields and needs to find the equilibrium between research activity on a university level and applicability for a scientific community	
Questions on where to present research work and publish results in journals needs to be addressed; in HEP a newly established peer reviewed journal on "Computing and Software for Big Science" [23] could address this topic	

people for placements of teaching or research work. Employees of those institutes will provide domain-specific expertise which can be used outside the natural science field, e.g. via seminars or block lectures at universities.

- **Computer science and software engineering university departments and schools** provide core software and computing research that can be used in natural science collaborations. Researchers get access to data sets and applied computing, which can benefit their own research fields.
- **Scientific collaborations and science communities** provide research topics that the institute will target. They benefit from access to the knowledge provided by computer science departments. Scientists and engineers forming the collaborations are mostly at the same time employees of natural science laboratories and universities.
- **National and international funding agencies** will provide the means to sustain the efforts of the institute on the long term. They benefit from the activities through synergies and collaborations across locations and domains.

The institute shall be based on 4 main pillars as shown in Table 2.

Table 2. 4 main pillars of the software institute

VIRTUAL	***FUNDAMENTAL***
The institute will not be specific to one place, but a connection of institutes, research labs, schools, etc. A small core team will take care of coordinating activities. Events of the institute such as meetings, workshops etc. will be held in various places within the community and partnering institutions.	The institute will concentrate on fundamental research on topics relevant to scientific computing in computer science departments and engineering schools, with the possibility of application to concrete R&D activities in the natural science fields.
MULTI-DOMAIN	***MULTI-NATIONAL***
With the concentration on fundamental research this knowledge should be also applicable to multiple natural science domains and also allow cross-fertilization between natural science communities.	By nature the institute will engage with partners in various countries such as computer science departments at universities, software engineering schools and natural science communities and experiments.

It is also to mention that many science fields have developed a rich culture of geographical distributed collaboration and remote communication via virtual events. In this respect we think that travel restrictions, as e.g. currently experienced during the COVID-19 pandemic, cannot jeopardise the successful operation of the institute.

With the first steps in this direction being taken, the question of the specific organisational structure of the institute, decision-making processes and legal aspects will be left to a later date, once a conceptualisation phase has been achieved and the full range of partners and stakeholders have a common concrete view on how to proceed (see Sec 4).

4 Next steps

The initial idea of setting up this initiative was first informally discussed in spring 2019. The first step, we want to build a community of interested parties, both in natural science and computer science, who support the idea. This community building can happen by reaching out to potentially interested parties via the direct contacts of already existing partners and further dissemination of the idea at conferences or workshops. It is also possible to join any national gatherings of multiple institutes and present the motivation, idea and guiding principles. On the side of natural sciences, the HEP community can serve as a first starting point to showcase the potential of the initiative. Other natural science fields are of course invited to join.

In parallel to the community building, some first concrete steps can be launched already with a limited set of partners in the form of topical workshops, where again both natural science and computer science shall meet and explain problems on the one hand and present research in a specific field on the other hand. These workshops should also act as a seeding ground for establishing connections and launching concrete R&D work. Apart from the organisation of workshops, very little financing is be needed in this first phase.

Once a critical mass of interested partners has been reached a conceptualisation phase of the institute shall be launched. This will include the setup of an organisational structure and discussions with funding agencies and/or multi-national founders (e.g. European Commission) to raise sufficient money to sustain the activity in the long-term.

Some natural sciences have already established computational natural science curricula at universities. A long term goal of the institute may also be to inject ideas of establishing such tracks also for other natural science fields.

5 Summary

We presented a new initiative of a "Software Institute for Data-Intensive Sciences" aiming at establishing a better collaboration between natural sciences and computer science. While ad-hoc connections between the two fields have been established from time to time, the aim of this institute is to provide a structure which will facilitate such connections and also enable R&D work between the two fields and across natural science domains. At the time of writing this document the work on launching this initiative is at an early stage, aiming to build a community around the idea and then proceed with a conceptualisation phase where the two domains will come together through various start-up activities designed to demonstrate the huge potential of working together.

6 Acknowledgements

The authors are grateful for discussion and input for this document and the overall process of setting up this initiative to Jakob Blomer[1], Jiri Chudoba[2], Peter Clarke[3], Dirk Düllmann[1], Peter Elmer[4], Maria Girone[1], Michel Jouvin[5], Thomas Kuhr[6], Alfons Laarman[7], Gonzalo Merino[8] and Axel Naumann[1].

References

[1] I. Bird et al, A plan to set up an institute for software in data intensive sciences, DOI:10.5281/zenodo.3466586

[2] https://cern.ch/sidis [visited Jan 2020]

[3] The HEP Software Foundation, A Roadmap for HEP Software and Computing R&D for the 2020s, DOI:10.1007/s41781-018-0018-8

[4] S.J. Hettricket et al, UK Research Software Survey 2014, DOI:10.5281/zenodo.1183562

[5] L.N. Joppa et al, Troubling Trends in Scientific Software Use, Science 17 May 2013, Vol. 340, Issue 6134, pp. 814-815, DOI: 10.1126/science.1231535

[6] https://hepsoftwarefoundation.org/ [visited Nov 2019]

[7] http://wlcg.web.cern.ch/ [visited Nov 2019]

[8] K. Bos et al, LHC computing Grid: Technical Design Report. Version 1.0 (20 Jun 2005), CERN-LHCC-2005-024

[9] https://openlab.cern/ [visited Nov 2019]

[10] Purcell et al, CERN openlab annual report 2018, DOI:10.5281/zenodo.3234404

[11] https://iris-hep.org/ [visited Nov 2019]

[12] https://rse.ac.uk/ [visited Nov 2019]

[1]CERN, CH [2]Czech Academy of Sciences, CZ [3]The University of Edinburgh, UK [4]Princeton University, US [5]Centre National de la Recherche Scientifique, FR [6]Ludwig Maximilians Universität, DE [7]Leiden University, NL [8]Port d'Informació Científíca/CIEMAT, ES

[13] https://www.de-rse.org [visited Nov 2019]
[14] https://nl-rse.org/ [visited Nov 2019]
[15] https://us-rse.org/ [visited Nov 2019]
[16] https://csc.web.cern.ch/ [visited Nov 2019]
[17] http://cs.unibo.it/projects/biss2019/index.html [visited Nov 2019]
[18] M. Erdmann et al, Challenges and Opportunities of Digital Transformation in Fundamental Research on Universe and Matter, http://www.ketweb.de/stellungnahmen/e300611/Strategiepapier_ErUM-Data_Final_2019-04-29.pdf
[19] P Buncic et al, Technical Design Report for the Upgrade of the Online-Offline Computing System, Apr 2015, CERN-LHCC-2015-006. ALICE-TDR-019, https://cds.cern.ch/record/2011297
[20] S. Benson et al, The LHCb Turbo Stream, 2015 J. Phys.: Conf. Ser. 664 082004, DOI:10.1088/1742-6596/664/8/082004
[21] https://careers.cern/students [visited Nov 2019]
[22] https://careers.cern/special-programmes [visited Nov 2019]
[23] https://link.springer.com/journal/41781 [visited Nov 2019]