A Plan to Set Up an Institute for Software in Data Intensive Sciences

Ian Bird, Simone Campana, Pere Mato, Stefan Roiser, Markus Schulz, Graeme Stewart, Andrea Valassi
European Organization for Nuclear Research (CERN), Geneva, Switzerland

Data processing using software is a fundamental part of today's scientific research. Planning for future experiments, e.g. in high energy physics (HEP), shows that significant improvements in software is needed to fully exploit the physics potential within a realistic computing budget [1]. This is compounded by a rapidly evolving hardware landscape, which needs to be adapted to. According to a study from 2014 in the UK [2], 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 [3] 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”.

To address these problems we propose to create an European Institute with a mission to promote excellence in software engineering and best practice in natural sciences based on **a strong collaboration with computer science departments and software engineering schools**. To ensure a sustainable impact the institute will gather, curate and disseminate software engineering and computer science knowledge in the long term.

In addition the institute also aims to:

- Encourage R&D resulting from collaborations of natural science researchers and computer scientists.
- Promote a career path for data scientists within scientific collaborations, communities and academia by raising awareness on technical, sociological and political levels.
- Cross-fertilize between different science fields, make knowledge retrievable and accessible across domain boundaries.
- Provide complex, large applications and data from natural sciences to computer scientists to serve as input for innovative engineering techniques or studies.
- Act as a lobbying forum for software engineering for natural sciences on national and international levels.

Following a Trans-European strategy, the institute will also help to link and bridge gaps amongst different European national and international initiatives which already have been funded or are being proposed in various countries such as the Software Sustainability Institute (UK), CDCS (DE), IRIS-HEP (US), HSF, etc.

On the path forward, the institute will first enter a conceptualisation phase where the above ideas will be implemented in the scope of HEP and astrophysics, reaching out to a limited set of interested countries, experiments and computing and software scientists. A series of topical workshops will be held to initiate the process and gather feedback on the most useful areas for the institute to be active in. At the same time concepts for funding and governance shall be developed. The conceptualisation phase shall not last longer than 2 years from the start of the initiative. In a second phase, after having proven to be successful, the institute shall reach out to additional data intensive sciences and operate in a sustained mode. More details can also be found in [4].

[1] The HEP Soft. Foundation, A Roadmap for HEP Software and Computing R&D for the 2020s, DOI:10.1007/s41781-018-0018-8
[2] S.J. Hettricket et al, UK Research Software Survey 2014, DOI:10.5281/zenodo.1183562
[3] L.N. Joppa et al, Troubling Trends in Scientific Software Use, DOI:10.1126/science.1231535
[4] I. Bird et al, Memorandum on a Software Institute for Data Intensive Sciences, http://cern.ch/go/f7lq