An Integrated Socio-Technical Crowdsourcing Platform for Accelerating Returns in eScience

Karl Aberer[1], Alexey Boyarsky[123], Philippe Cudré-Mauroux[4], Gianluca Demartini[4], and Oleg Ruchayskiy[5⋆]

[1] Ecole Polytechnique Fdrale de Lausanne, Switzerland
{firstname.lastname}@epfl.ch
[2] Instituut-Lorentz for Theoretical Physics, Universiteit Leiden, Leiden, The Netherlands
[3] Bogolyubov Institute for Theoretical Physics, Kiev, Ukraine
[4] eXascale Infolab, University of Fribourg, Switzerland
{firstname.lastname}@unifr.ch
[5] CERN TH-Division, PH-TH, Geneva 23, Switzerland
oleg.ruchayskiy@cern.ch

Abstract. Progress in science relies nowadays on collaborative efforts of large communities. A single human being has no more the capacity to process all the information necessary to fully comprehend the experimental facts and implications of scientific experiments. We claim that this will result in a fundamental phase transition in how scientific results are obtained, represented, used, communicated and attributed. Different to the classical view of how science is performed, important discoveries will be not only the result of exceptional individual efforts and talents, but alternatively an emergent property of a complex community-based socio-technical system. We even speculate that certain discoveries might be of such a complexity that human individuals might no more be able to fully grasp the underlying models and methods. This has fundamental implications on how we perceive the role of technical systems and in particular information processing infrastructures for scientific work: They are no longer a subordinate instrument that facilitates (or makes more miserable) daily work of highly gifted individuals, but become an essential tool and enabler for performing scientific progress, and eventually might be the instrument within which scientific discoveries are made, represented and brought to use.

1 Introduction

Progress in science relies today on collaborative efforts of large communities. A single scientist has no more the capacity to process all the information necessary to fully understand and comprehend all the experimental facts and models of large-scale scientific endeavors. Even though scientific breakthrough performed by individual scientists or teams of scientists is still at the basis of the innovation process, it is becoming *de facto* impossible for individuals to understand the full implications of their local discoveries in today's networked scientific landscape. As an example, the OPERA collaboration recently claimed to observe neutrino propagations that are faster than the speed of light. Their observations are based on a distributed experimental apparatus, which is so complex that no single expert (inside or outside of the collaboration) can claim to understand all sources of systematic errors in the setup. As a result, the only way to verify the results (as stated by the collaboration itself) is to compare those results with an alternative—and equally complex, challenging and expensive—experimental setting.

Now let us imagine that a high-quality semantic analysis of the description of the neutrino propagation experiment has been performed semi-automatically. It would result in a "semantic scheme"—embedded into a shared-ontology—that combines the various fields of expertise of all participants in

[⋆] Authors are listed in alphabetical order.

the experiment. Such an analysis could be used to reveal a complete list of assumptions, which are often implicit in the experimental setup/data analyses steps. It would then be possible for the system to reason on the experimental setup and results, and to identify new sources of systematic error, previously overlooked due to the high complexity of the system and the diversity of participating experts; the system could then even provide a new workflow and a new set of results by isolating the systematic errors and automatically circumventing them.

Coming back to the bigger picture, it is becoming increasingly clear that the current best practices in sharing scientific results and advancing science are getting obsolete due to the sheer complexity and scale of the problems, models and experiments. We claim that this will result (or is already resulting) in a fundamental phase transition on how scientific results will be obtained, represented, used, communicated and attributed. Different to the classical view of how science is performed, important discoveries will not only be the result of exceptional individuals, but also an emergent property of a complex community-based socio-technical system.

We believe that considerations of the above nature are central when developing next-generation knowledge infrastructures for supporting scientific work. In the rest of this paper, we outline a first set of requirements for such a system by giving initial thoughts to the following questions:

- how are the human participants in such a collaborative scientific ecosystem coordinated and motivated to provide necessary contributions, and how do they cooperate with the automated processes?
- how are the scientific data, processes, and results shared within the system to enable automated cross-pollination?

2 Human Incentives and Scientist-Computer Symbiosis

It is clear that to achieve this goal, ontologies of unprecedented quality and complexity are needed (including plenty of complex, context-dependent concepts and methods, conceptualized hypotheses, etc.). Today's methods and infrastructures for building ontologies are clearly insufficient to achieve this goal. We believe that the only possible approach is to implicitly "*crowdsource knowledge elicitation from the expert community*".

A significant fraction of the scientific knowledge in any given field is not formalized in terms of data or publications, but rather "exists in the heads of the experts". We imagine below a system that could formalize targeted parts of expert activities, making implicit knowledge available for automated use. This system would include facilities for: **(a)** understanding the exact meaning of scientific concepts *within a given context*; distinguishing between directly measured and derived quantities, understanding assumptions and the fundamental differences between observed phenomena and their mathematical abstraction; **(b)** understanding the "mental map of a research area", grasping the main conceptual ingredients of a given field (both of phenomenological and formal nature). This mental map provides a large-scale overview of a more detailed knowledge-graph capturing the advances/concepts along with their non-trivial relationships; **(c)** ranking expert contributors on a case-by-case in a given field (for practical reasons it is often important to identify the best experts in a narrow, precisely defined field; for instance, young researchers, whose contributions in the field so far might be limited, may be the ideal candidates to summarize some advances or to contribute to specific, rapidly evolving fields); **(d)** understanding the methods of analysis within a given field of study and the ability to properly use them (or even develop them beyond the state-of-the-art).

In the end, the entire socio-technical system would act as a giant crowdsourcing conceptualization machine for complex scientific fields, that continuously elicits all hypotheses, concepts, and contextual information related to a scientist's daily activities. The only feasible way to achieve all those goals is to provide a scientific infrastructure that can implicitly and automatically follow the entire life cycle of the experts' workflows, while explicitly helping him in his daily routine (e.g., by providing him with effective and integrated search tools, editors and semantic-aware data processing frameworks to assist him throughout his scientific discovery process.) The main objective is thus to heavily invest in performant, user-friendly and customizable scientific tools to help the scientists save time and effort in their daily work, while building complex ontological networks to capture their scientific work in the background.[6]

The machine-processable information elicited in this process can then be used to automate as many routine operations as possible for the individual scientists. In that context, the experts would *constantly but implicitly train the system* through their daily scientific activities, but would directly benefit from the implicit elicitation process by being able to quickly automate all of their highly repetitive tasks.

3 Automated Cross-Pollination

How can we design a generic system capable of (re)interpreting and combining disparate results obtained through heterogeneous experimental settings into something novel and potentially useful to the scientific community? A first step in that direction is to take all local data, processes and results used by the scientists locally as well as the hypotheses, conceptualizations and contextual information made explicit by the process described above, and to make them available in an open, networked system. This represents a somewhat disruptive scientific model—where the norm is to instantaneously share all artifacts created throughout the scientific process—which seems however imperative in order to enable automated cross-pollination throughout our socio-technical system.

We believe that current and future Semantic Web formalisms will play a key role in this context. Highly-expressive machine-processable formats must be used to describe and represent all input data, conceptualizations, hypotheses, workflows, and results in the shared infrastructure. In addition to those rather obvious self-descriptive requirements, related pieces of information must be explicitly connected across the entire system; this demands for highly effective and scalable methods to:

- relate semantically similar but syntactically heterogenous conceptualizations and entities; current and future advances in ontology matching, entity resolution, and linked data enrichment can be used in this context.
- maintain fine-grained lineage information in order to trace back, and possibly re-generate or modify (for instance through speculative and automated workflow generation) output data; in this context, although existing provenance languages can be used, novel formalisms are needed to represent lineage using expressive constructs to adequately capture all scientific processes, and to refrain the information explosion that current solutions lead to.
- discriminate conflicting information; as important as the inter-linking and integration points above, the system needs to isolate conflicting facts, and cluster data based on incompatible concepts, hypotheses, or experimental setups; this is essential in order to correctly discriminate sets of scientific experiments, to drastically reduce the search space when trying to automatically compose existing experiments, and to propose entirely new scientific experiments based on already available hypotheses, data and workflows.
- reason upon all available pieces of information, in order to infer new data, concepts, hypotheses or processes based on available information. Such capabilities should also be extended to draw entirely new

[6] Our current platform `http://ScienceWISE.info` provides a first step in this direction. Its semantic bookmarking and annotation gives a simple but important example of such a "symbiosis": scientists use the system for their own work, but also enrich the scientific ontology and provide services to the whole community while doing so.

conclusions from disparate sets of preexisting data (e.g., to automatically create new conceptualization based on semantic-aware workflows and their corresponding experimental results).

In the end, the whole system would act as a gigantic *entropy-reduction machine*, scrutinizing all creative steps performed by individual scientists in the context of the overall system and trying to classify, corroborate, enrich and ultimately combine local data, hypotheses, workflows and conclusions in light of all other scientific artifacts contributed to the system.

4 Conclusions

Developing the system described above will not only make it possible to overcome the complexity crisis in natural sciences; it will also open a new era for social sciences (complex by their nature) and make a unique step towards transforming the Web from a presentation platform into an integrated, collective intelligence engine.

Our speculative perspective has fundamental implications on how we perceive the role of technical systems and in particular information processing infrastructures in a scientific context. Information systems in this case are no longer subordinate instruments that facilitate (or even make more miserable) daily work of highly gifted individuals, but become an (the) essential catalyst for performing scientific progress, and eventually will be the instruments within which scientific discoveries are made, represented and brought to use. This implies fundamental questions and requirements for such an infrastructure, including: *how is the presence and correctness of a scientific discovery that is represented in an implicit way within the supporting infrastructure verified?*

We can even speculate that certain discoveries made in such a context might be of such a complexity that individual scientists may no more be able to fully appreciate the underlying models and methods used by the system. Humans might only be able to interpret a subset of the full results that indicate the existence of such a discovery in the scientific system, which would call for self-awareness qualities for the socio-technical system (i.e., mechanisms to observe the processes in place and semi-automatically detect and summarize the emergence of important properties, data and theories across the knowledge ecosystem.)

On more practical grounds, it is important to stress that the creation of such a complex, self-organizing socio-technical infrastructure—providing a new symbiosis between a crowdsourcing information system and an integrated, semantic-aware platform for eScience—is a singularly challenging task. It requires significant efforts in capturing and automatically processing non-trivial scientific activities, and in identifying and deploying next-generation information management techniques to share, reason upon, and automatically recombine arbitrary scientific artifacts in vey large-scale, heterogeneous settings. The development efforts for such an infrastructure does *not* represent a "service to the community" *per se*. Rather, it materializes as a coordinated investment into a crucial instrument for future advances in science. It can thus be compared to large-scale investments of the physics community in facilities like sophisticated accelerators (e.g., the LHC at CERN) or cosmic missions. Without such efforts, the future described above may be significantly delayed or even rendered impossible.

5 Acknowledgments

This paper was supporter (in part) by the Swiss National Research and Education Network SWITCH as part of the "AAA/SWITCH – e-infrastructure for e-science" program, and by funds from the ETH Board and the Swiss National Science Foundation under grant number PP00P2_128459.