2nd Workshop on Capturing Scientific Knowledge (Sci-Know): Summary Report

Daniel Garijo[1], Martine De Vos[2]
[1]University of Southern California, Information Sciences Institute, Marina del Rey, CA, U.S.A
[2]Netherlands eScience Centre, Amsterdam, The Netherlands

dgarijo@isi.edu, m.devos@esciencecenter.nl

1 INTRODUCTION AND GOALS

The 2nd workshop on Capturing Scientific Knowledge (SciKnow 2017) [1] aimed to bring together researchers interested in representing and capturing knowledge about different areas of science so that it can be used by intelligent systems to support scientific research and discovery. Although great advances have been made in the last decade, scientific knowledge is still complex and poses great challenges for knowledge capture. SciKnow provided a forum to discuss existing forms of scientific knowledge representation and existing systems that use them, envisioning major areas to augment and expand this field of research.

SciKnow 2017 is the follow up event of a series which started at K-CAP 2015.[2] SciKnow 2017 took place in Austin Texas, December 4th, and had between 17 and 22 attendants during a full day event. Participants had different backgrounds, ranging from computer science and e-Science to bioinformatics and hydrology.

2 WORKSHOP SUMMARY

SciKnow 2017 was divided in three main sessions, which grouped seven paper presentations, together with a keynote presentation and a discussion panel. They keynote was presented by Suzanne Pierce, a hydrologist who explained the importance of decision support based methodologies to successfully communicate scientific outcomes to end users. The keynote described several examples of miscommunication, and their environmental and economic consequences. A key aspect of capturing existing scientific knowledge should be making it understandable and available to potential affected users.

The first session of the workshop focused on domain-specific knowledge capture, in particular when capturing provenance in hydrology [7] and facilitating curation of bioinformatics datasets [1].

The second session contained works related to reproducibility and reusability of scientific methods, with guidelines and metadata recommendations on how to perform abstractions on scientific experiments [6], proper design of computational notebooks [3] and spreadsheet annotation [4].

The final session dealt with the automated support for scientific processes, introducing an approach for capturing the iterative nature of scientific experiments [2] and an approach for representing hypothesis in knowledge discovery systems [5].

3 DISCUSSION SUMMARY

SciKnow 2017 was a participative discussion-oriented workshop. All presentations had several questions and comments, further elaborated at the discussion panel. Below is a summary of the main topics addressed in the workshop.

3.1 From machine in the loop systems to human in the loop systems

Nowadays human scientists are the main drivers of scientific research. Scientists formulate hypotheses, search for appropriate data, prepare and execute experiments, collect results and write down their corresponding conclusions. Intelligent systems are used only as support tools that perform computational experiments set by scientists; or search and store results. In this context, the first topic discussed in the workshop focused on how participants envisioned the role of scientists when doing research with intelligent systems.

Participants distinguished three main roles for researchers. The first one is before an experiment takes place, guiding which sources of information and hypotheses may be most appropriate to do research on. The second one is while doing research, where scientists could interact iteratively with intelligent systems until a solution for the problem at hand is found. Finally, the last type of interaction is after an experiment is performed, in order to be able to find potential errors and curate existing results.

As denoted by the roles, participants agreed for the need of humans to be kept as part of the research lifecycle. Intelligent systems may be used to help and support scientific researchers. However, researchers cannot be blind to the scientific process, otherwise they won't be able to explain the obtained results appropriately and therefore adopt a new proposed solution.

The discussion ended by addressing the need for a better knowledge transfer and capture for intelligent systems. Existing approaches are currently able to extract information and reason with it, but they often do not understand it. Addressing this gap seems like a crucial need for intelligent systems to play more important roles in scientific research.

[1] https://sciknow.github.io/sciknow2017/
[2] https://www.isi.edu/ikcap/sciknow2015/

3.2 Real science versus ideal science

Several of the papers and examples presented at the workshop discussed case studies showcasing the complexity of real scientific problems [1] [7]. In computer sciences, researchers tend to simplify problems and design solutions that are generically applicable. These solutions are often difficult to adapt to real world problems in specific domains.

During the workshop we observed that current 'computer sciences approaches' mainly consider a limited fragment of the scientific process, mostly related to experiments, observations and protocols. Other, less tangible aspects of the scientific process are often left out of scope: formulation of hypotheses and research questions, assumptions and choices made by the scientists and interpretation and use of knowledge. These aspects capture the context of a research experiment, and are key for its understanding by other scientists.

In addition, supporting scientists in their work is not trivial. In order to make 'computer science approaches' applicable, scientists have to be open and explicit about their data and methods. At the moment this is perceived by part of the community as risky (fear of losing ownership of data) and time consuming (due to the time required to create all metadata). Intelligent systems should highlight incentives for fostering open reusable data.

4 CHALLENGES FOR SCIENTIFIC KNOWLEDGE CAPTURE

According to the discussions aired in the workshop, three main challenges must be tackled in order to integrate intelligent systems in the research lifecycle conducted by researchers.

The first challenge is **knowledge explainability**. Humans need to be part of the research lifecycle. Scientists should be able to intervene in situations that are surrounded by ambiguity and uncertainty, where automated reasoning may not be applicable. Besides, both scientists and non-scientific parties should be able to use the information produced by intelligent systems. In order to be understood and applied by humans, information in the research lifecycle needs to be clear, unambiguous, explicit and accompanied by semantics. Having intelligent systems help creating explanations of scientific experiments may foster their reusability.

The second challenge is **knowledge transfer**. Intelligent systems may have full access to data and source codes of a research project in order to support any research associated to it. However, access to data and software alone is not sufficient if an intelligent system cannot understand its contents or its functionality. Data, software and methods should be accompanied by metadata and information at a conceptual level. In order to address this issue, intelligent systems may also become part of the metadata annotation process, helping humans curating metadata annotations and abstractions. With these knowledge, intelligent systems would be able to help relating different datasets, software and methods together.

The third challenge is **context capture**. Scientific knowledge of research projects is only valuable when it can be understood by other scientists different from the original 'creators', and reused in different situations. This requires that the context of a research project is captured, including information on hypotheses and research questions, assumptions, design decisions, etc. that are part of the domain of interest. Context is currently captured in publications in human-readable format, but it is generally unavailable in machine readable format appropriate for intelligent systems. A way to tackle this issue may involve highlighting the benefits of such annotations, proposing assistants that use this knowledge to infer and suggest relevant lines of research.

ACKNOWLEDGMENTS

The workshop organizers would like to thank all workshop attendees who participated in the discussions and presentations, including Yolanda Gil, Suzanne Pierce, Takahiro Kawamura, Francesco Osborne, Gully Burns, Natalia Villanueva-Rosales, Danai Symeonidou, Blake Regalia, Joe Raad, Al Idrissou, Mauro Vallati, Iacopo Vagliano, Endris Kemele, Pablo Calleja and Ruben Taelman.

REFERENCES

[1] Burns, G. Vita, R. Overton, J., Fleri, W., Peters, B. "Semantic Modeling for Accelerated Immune Epitope Database (IEDB) Biocuration". SciKnow 2017, Austin, Texas, 2017.

[2]. Carvalho, L. Garijo, D., Essawy, B. Medeiros, C and Gil, Y. "Requirements for Facilitating the Continuous Creation of Scientific Workflow Variants". SciKnow 2017, Austin, Texas, 2017.

[3] Carvalho, C., Wang, R., Gil, Y and Garijo, D. "NiW: Converting Notebooks into Workflows to Capture Dataflow and Provenance". SciKnow 2017, Austin, Texas, 2017.

[4] De Vos, M., Wielemaker, J., Wielinga, B., Schreiber, G. and Top, J. "How plausible is automatic annotation of scientific spreadsheets?". SciKnow 2017, Austin, Texas, 2017.

[5] Garijo, D., Gil, Y. and Ratnakar, V. "Capturing Hypothesis Evolution in Automated Discovery Systems". SciKnow 2017, Austin, Texas, 2017.

[6] Gil, Y., Garijo, D., Knoblock, M., Deng, A., Adusumilli, R., Ratnakar, V., Mallick, P. "A Workflow Design Methodology to Improve Reproducibility andReusability of Computational Experiments". SciKnow 2017, Austin, Texas, 2017.

[7] Villanueva-Rosales, N., Garnica L., Smriti, C., Tamrakar, R., Pennington, D. Vargas-Acosta, R.A., Ward, F. and S. Mayer, A. "Capturing Scientific Knowledge for Water Resources Sustainability in the Rio Grande Area". SciKnow 2017, Austin, Texas, 2017.