

PhD Proposal

Data-centric Workflows for *Crowdsourcing Applications*

Loïc Hélouët (CR, Hab.), Zoltán Miklós (MCF) – Laboratoire IRISA

« Crowdsourcing » is a generic term for task-solving techniques that rely on a large group of online users. We can consider for example the success of FoldIt [1], an online game on protein folding, which allowed the crowd to solve a problem left open by specialists. Wikipedia can also be seen as an encyclopedia produced by crowdsourcing. Commercial versions of crowdsourcing also exist, such as Amazon Mechanical Turk [2]. In e-Science, crowdsourcing is used to gather huge data sets (*participative sensing*, for example the « Sauvage de ma rue » [4] project). Systems specifically designed for crowdsourcing are on their way (sCOOP at Stanford [5], crowdDB at Berkeley [6]).

A difficulty addressed by crowdsourcing systems is to build complex applications orchestrating crowd competences. Such applications can be complex processes that need to distribute large sets of data to crowd participants, then aggregate the obtained results, and continue the process differently according to the nature or quality of collected answers. Complex crowd based services are frequently implemented through human management of tasks distribution, or using ad-hoc and low-level programming solutions. The next challenge for crowdsourcing systems is to allow for easy design of applications and services with complex workflows over crowd platforms. This calls for the design of intuitive formalisms to facilitate design, deployment, and runtime management of complex tasks on a crowd platform. The considered models have to handle at the same time data, control (i.e. handle complex tasks progress depending on collected answers), quality of collected answers, and provide mechanisms to distribute work to pools of crowd participants with various competences in order to maximize crowd efficiency [11].

The proposed PhD focuses on some of the above-mentioned aspects. The goal of this PhD is to provide tools and techniques for the development and deployment of complex crowd applications. In particular, it will focus on some the following issues:

- To contribute to the definition of models for complex workflow design over crowd platforms. The starting point for the study of such models can be data-centric declarative formalisms such as datalog, webdamlog [10], or grammars [8] but also more orchestration oriented models: Business artifacts [6,7], process algebras [12], games [13] or transaction oriented models such as [9]. While these models are well suited to describe workflows in general, they are less adapted to deal with imprecisions or inconsistencies that appear in human input.
- To model complex tasks that require interactions between participants, allow complex answering mechanisms or tasks collaboration. Such mechanisms should propose adaptive models allowing for instance a crowd user to easily define a workflow, and return it as an answer to a question.
- To propose deployment schemes for such models, i.e. propose methods to map a complex crowd workflow on a chosen crowd platform.
- To implement and evaluate proofs of concepts for complex workflow models deployment on ad-hoc or existing platforms (Amazon, Foule Factory, CrowdFlower, ...)

This PhD can be focused on its theoretical side (emphasis on models), and/or on its system side, with the implementation of a proof-of-concept. Candidates with theoretical or system skills are very welcome.

Context & Supervision

This PhD takes place in the context of the HEADWORK project (2016-2020), funded by the research agency ANR. The PhD student will work at IRISA, Rennes, France. The thesis is co-supervised PhD, by Loïc Héliouët (CR INRIA, SUMO¹ team) and Zoltan Miklos (Mcf, DRUID² team).

The position is funded for three years, and can start after September 2017.

Loïc Héliouët :

Mail: loic.helouet@inria.fr

Tel: 02 99 84 75 90

Web <http://people.rennes.inria.fr/Loic.Helouet>

Zoltan Miklos :

Mail : zoltan.miklos@irisa.fr,

Tel: 02 99 84 22 54

Web: <http://people.irisa.fr/Zoltan.Miklos/>

Competences

The PhD candidate should hold a master or equivalent degree in computer science. He or she should also have the following competences:

- Fluent in English (written, spoken)
- Basic algorithmic skills

Competences in some of the followings domains are not mandatory but are welcome

- Formal techniques (automata, models checking, algebras, ...)
- Implementation skills
- Databases and data management

Foreign applications are welcome. Knowledge of French is not mandatory.

Application

The PhD candidate should send by mail to loic.helouet@inria.fr and zoltan.miklos@irisa.fr the following documents:

- Complete curriculum vitae
- Motivation letter
- Copy of the master grades, master thesis, and reports (if available).
- Two references.

¹ <http://www.irisa.fr/sumo/>

² <http://www-druid.irisa.fr>

References

- [1] *Foldit, solve puzzles for science.* <http://fold.it/portal/>
- [2] *Amazon Mechanical Turk.* <https://www.mturk.com/mturk/welcome>
- [3] *Sauvages de ma rue.* <http://sauvagesdemarue.mnhn.fr>
- [4] *Projet sCOOP, Stanford.* <http://www-cs-students.stanford.edu/~adityagp/scoop.html>
- [5] M.J. Franklin, D. Kossmann, T. Kraska, S. Ramesh, R. Xin: *CrowdDB: answering queries with crowdsourcing.* 61-72.
- [6] R. Hull, J. Su, and R. Vaculin. *Data management perspectives on business process management: tutorial overview.* SIGMOD, 2013.
- [7] S. Abiteboul, P. Bourhis, A. Galland, B. Marinoiu: *The AXML Artifact Model.* TIME 2009: 11-17, 2009.
- [8] E. Badouel, L. Hérouët, G.-E. Kouamou, C. Morvan, R. Nsaibirni. *Active Workspaces; Distributed Collaborative Systems based on Guarded Attribute Grammars.* ACM SIGAPP Applied Computing Review **15**(3):6-34, 2015.
- [9] E.Badouel, L. Hérouët, C.Morvan, *Petri nets with semi-structured data.* In *Petri Nets'15*, Bruxelles, June 2015.
- [10] S. Abiteboul, E. Antoine, G. Miklau, J. Stoyanovich, J. Testard, *Rule-based application development using Webdamlog,* ACM SIGMOD International Conference on Management of Data, ACM, pp 965—968, 2013.
- [11] P. Mavridis, D. Gross-Amblard, Z. Miklos. *Using Hierarchical Skills for Optimized Task Assignment in Knowledge-intensive Crowdsourcing.* 25th International World Wide Web Conference, 2016.
- [12] R. Pugliese, F. Tiezzi, *A Calculus for Orchestration of Web Services:* J. Applied Logic 10(1), p 2-31, 2012.
- [13] A. Morishima, S. Fukusumi, H. Kitagawa. *Cylog/Game aspect: An approach to separation of concerns in crowdsourced data management.* Information Systems (2016). <http://dx.doi.org/10.1016/j.is.2016.01.004>