# Thesis proposal for a Doctoral position
## 2018-2021

| | | |
|---|---|---|
| **Title** | **ADOPI: A**utomated **D**omain independent **O**ntology **P**opulation using the **I**nternet of things | |
| **Supervisors** | Jean-Marc LE LANN | |
| | Anne-Marie BARTHE-DELANOË | annemarie.barthe@ensiacet.fr<br>+33 (0)5 34 32 33 08 |
| **Laboratory** | Chemical Engineering Laboratory (LGC – Toulouse) | |

## Research project description

The need for *situational awareness* is crucial in various application domains where understanding the environmental elements and events (with respect to time and space) is critical for decision-making. Indeed, **decision support systems** rely on the availability and quality of the information describing the context. The term was initially used for the aircraft pilots community (Endsley, 1996), but it is now widely used for power plant operations, advanced manufacturing systems, supply chain management, crisis management, etc. - in other words for any complex and dynamic system. Providing context information helps the systems to detect and analyze a change (in the system itself and its environment) and to react properly.

With the **rise of the interconnection and communication** between objects and people, the so-called Internet of Things (IoT) (Atzori et al., 2010; Kopetz, 2011), including initiatives like Open Data (Open Data Commons, 2006), crowdsourcing (Gao et al., 2011) and citizen participation (Bremer, 2013; Castillo, 2016; Debin et al., 2013; Meier, 2012), we are facing an exploding amount of emitted data.
This **tremendous stream of raw data** about the environment, people, organizations, processes, activities, etc., known as Big Data (Demchenko et al., 2013; Manyika et al., 2011), can be used as an input for **ensuring situational awareness**. Plus, one interesting characteristic of the IoT is to not be focused on specific applications. It can be seen as a "general purpose sensor network" (Perera et al., 2014), meaning that sensors can be reused, connected and combined with other sensors on demand.

In their extensive literature review about research prototypes and systems related to awareness and IoT, (Perera et al., 2014) pointed out that some of **the significant challenges are**:

1. ***Context discovery***. The question of *how to understand the sensor data and the related context automatically?* is challenging due to the very nature of IoT which covers a wide range of application domains.
2. ***Context modelling***. Acquisition, modelling and reasoning can be achieved by multiple techniques. But the immaturity of the IoT field challenges this step, mainly from an interoperability point of view.
3. ***Context sharing***. Most middleware applications were developed with a silo approach without ensuring inter-middleware communication, whereas there is no central point of control in the IoT. As a consequence, context sharing is not supported.

Given this background, our research proposal envisions to investigate the way to **bridge these gaps between data** (IoT sensing) **and information and knowledge** (situational awareness, actionable knowledge), and how to achieve it **automatically** to support decision making. In the frame of this PhD thesis proposal, we will focus on the following issue:

*How to automate data extraction and analysis from numerous heterogeneous streams of data (considering data format, data semantics, etc.) and turn the obtained datasets into useful information for the decision makers?*

In other words, we are interested in automating the knowledge database population to benefit from the IoT.

Considering that in the last decade ontologies gained popularity in data-intensive domains (Kontopoulos et al., 2017; Skvortsov et al., 2016), we will focus our proposal on knowledge databases based on ontologies.
Research works about automated knowledge base population have been pursued for more than a decade, and more specifically on ontology population (Maynard et al., 2009) in the case of the integration of massive volumes of data for further use at the enterprise level. However, populating ontologies with massive volumes of data remains a challenge. Manual ontology population is no longer an option, as it is not only error-prone and time-consuming but unrealistic regarding the amount of data produced by the IoT.
To tackle this issue, many approaches are proposed in the literature. Three main dimensions can be considered to evaluate existing approaches: (i) the automated or the semi-automated approach; (ii) the degree of domain dependency; (iii) the kind of processed datasets.

**Automated or semi-automated approaches**

These approaches are quite self-explaining. Semi-automated approaches, such as (Karkaletsis et al., 2006), require domain expert intervention: to annotate texts (in the case of manual annotation for Natural Language Processing), to control and validate concepts, etc.

**Domain dependency**

Domain independent solutions often create their own ontology based on the studied corpus (Kaushik and Chatterjee, 2017). They build, enrich and adapt ontologies from a corpus of natural language text (Lehmann and Völker, 2014). Therefore these solutions are more related to ontology learning than ontology population as the latter does not alter the ontology schema. Domain independent solutions are hardly suitable for existing knowledge bases population as they cannot fit the existing structure of an ontology. Domain dependent solutions offer a wider range of ontology population solutions, especially in the biomedical domain. But the classifications rules are generated a priori for a given domain ontology (Faria et al., 2013).

**Processed datasets**

Most of the ontology population tools are designed to extract knowledge from natural language text (Kontopoulos et al., 2017; Petasis et al., 2011) and they do not consider other data materials (e.g. (Faria et al., 2014)). They mainly involve machine learning, text mining and Natural Language Processing (NLP) techniques.
Very few research works are trying to take into account other data sources. For example, the HOLMES framework (Remolona et al., 2017), which is also focusing on textual data, envisions to integrate Graphical Image Processing. The PROPheT tool (Kontopoulos et al., 2017) proposes to process Open Linked Data. None of the existing solutions proposes to use more structured but atomic data (e.g. as log files, sensor data) to infer new knowledge (e.g. about real operational processes) in order to feed ontologies.

*The literature on ontology population shows that existing automated data extraction and analysis solutions are mainly domain specific and thus hardly adaptable to a different domain. Most of the domain independent approaches focus on data sources, mainly using textual data, and ignore data sources such as sensors.*

The aim of this PhD thesis proposal is (i) to define and design an Automated Domain-independent Ontology Population framework for existing ontologies using IoT data (textual data as well as raw data from sensors and event logs), (ii) to identify and solve data interoperability and consistency issues among processing components to achieve an automated knowledge management unified process, (iii) to assess the feasibility on use-cases from different business domains.

Thus, to address the above-mentioned research problem, research work will be conducted with a systemic and model-driven engineering approach on (at least):

- Semantics and ontology alignment to find and structure data (Wang, 2015), store information and share knowledge. They could also allow transfer learning and ontology re-use (which try to adapt existing ontologies to new domains) with a syntactic semantic approach,
- Natural Language Processing techniques (supported by Machine Learning approaches) to automate extraction and classification of instances of the concepts and relationships defined in the ontology from texts,

- Complex Event Processing (CEP) (Etzion et al., 2015) and process mining (van der Aalst, 2016). CEP allows to support the collection and the inference of new data through the processing of real-time measurements (sensor data, log files) and information about events that occurred within organizations, machines, systems. CEP can be used as a preprocessing step for process mining. Process mining can be conducted to discover a process based on event logs.

A proof of concept, based on two use-cases, is expected to assess the feasibility of the proposal: a biomass processing use-case based on the BIOCORE ontology (Cecelja et al., 2011), a public health use-case based on the Public Health Document Ontology (Zhang et al., 2017).

This PhD thesis will be directed by Prof. Jean-Marc Le Lann and by Dr. Anne-Marie Barthe-Delanoë.

In the frame of a **collaboration** with the Complex Resilient Intelligent Systems Laboratory (CRIS Lab - University of Columbia), and more specifically with Prof. Venkat Venkatasubramanian, we will also benefit from their skills and knowledge on knowledge management and Natural Language Processing.

| | |
|---|---|
| **Keywords** | Complex Event Processing; Open Data; Big Data; Knowledge Modelling; |
| **Candidate profile** | Data Science; Industrial Engineering; Information Systems Engineering; Computer Engineering; |

## References


- Atzori, L., Iera, A., Morabito, G., 2010. The Internet of Things: A survey. Computer Networks 54, 2787–2805. https://doi.org/10.1016/j.comnet.2010.05.010
- Bremer, R., 2013. Fluthelfer mit Google Maps - eine Geschichte aus Halle. Der offizielle Google Produkt-Blog.
- Castillo, C., 2016. Big Crisis Data: Social Media in Disasters and Time-Critical Situations. Cambridge University Press.
- Cecelja, D.F., Kokossis, P.A., Du, D.D., 2011. Integration of ontology and knowledge-based optimization in process synthesis applications, in: Pistikopoulos, E.N., Georgiadis, M.C., Kokossis, A.C. (Eds.), Computer Aided Chemical Engineering, 21 European Symposium on Computer Aided Process Engineering. Elsevier, pp. 427–431. https://doi.org/10.1016/B978-0-444-53711-9.50086-9
- Debin, M., Turbelin, C., Blanchon, T., Bonmarin, I., Falchi, A., Hanslik, T., Levy-Bruhl, D., Poletto, C., Colizza, V., 2013. Evaluating the Feasibility and Participants' Representativeness of an Online Nationwide Surveillance System for Influenza in France. PLoS ONE 8, e73675. https://doi.org/10.1371/journal.pone.0073675
- Demchenko, Y., Grosso, P., De Laat, C., Membrey, P., 2013. Addressing big data issues in scientific data infrastructure, in: Collaboration Technologies and Systems (CTS), 2013 International Conference on. IEEE, pp. 48–55.
- Endsley, M.R., 1996. Automation and situation awareness. Automation and human performance: Theory and applications 163–181.
- Etzion, O., Fournier, F., von Halle, B., 2015. "The Event Model" for Situation Awareness. Data Engineering 105.
- Faria, C., Girardi, R., Novais, P., 2013. Analysing the Problem and Main Approaches for Ontology Population, in: 2013 10th International Conference on Information Technology: New Generations. Presented at the 2013 10th International Conference on Information Technology: New Generations, pp. 613–618. https://doi.org/10.1109/ITNG.2013.94
- Faria, C., Serra, I., Girardi, R., 2014. A domain-independent process for automatic ontology population from text. Science of Computer Programming, Special Issue on Systems Development by Means of Semantic Technologies 95, 26–43. https://doi.org/10.1016/j.scico.2013.12.005
- Gao, H., Barbier, G., Goolsby, R., Zeng, D., 2011. Harnessing the Crowdsourcing Power of Social Media for Disaster Relief.
- Karkaletsis, V., Valarakos, A., Spyropoulos, C.D., 2006. Populating ontologies in biomedicine and presenting their content using multilingual generation. Programme Committee 51.
- Kaushik, N., Chatterjee, N., 2017. Automatic relationship extraction from agricultural text for ontology construction. Information Processing in Agriculture. https://doi.org/10.1016/j.inpa.2017.11.003
- Kontopoulos, E., Mitzias, P., Riga, M., Kompatsiaris, I., 2017. A Domain-Agnostic Tool for Scalable Ontology Population and Enrichment from Diverse Linked Data Sources, in: Data Analytics and Management in Data Intensive Domains. Presented at the XIX International Conference on Data Analytics and Management in Data Intensive Domains, Springer, Cham, Moscow, pp. 184–190.
- Kopetz, H., 2011. Internet of Things, in: Real-Time Systems, Real-Time Systems Series. Springer US, pp. 307–323.
- Lehmann, J., Völker, J., 2014. Perspectives on Ontology Learning. IOS Press.
- Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A., 2011. Big data: The next frontier for innovation, competition, and productivity. URl: http://www. mckinsey. com/Insights/MGI/Research/Technology_and_Innovation/Big_data_ The_next_frontier_for_innovation 156.

- Maynard, D., Funk, A., Peters, W., 2009. SPRAT: a tool for automatic semantic pattern-based ontology population, in: In: International Conference for Digital Libraries and the Semantic Web.
- Meier, P., 2012. Crowdsourcing a Crisis Map of the Beijing Floods: Volunteers vs Government. iRevolution.
- Open Data Commons, 2006. Open Database License (ODbL) v1.0.
- Perera, C., Zaslavsky, A., Christen, P., Georgakopoulos, D., 2014. Context Aware Computing for The Internet of Things: A Survey. IEEE Communications Surveys & Tutorials 16, 414–454. https://doi.org/10.1109/SURV.2013.042313.00197
- Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E., 2011. Ontology Population and Enrichment: State of the Art, in: Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 134–166. https://doi.org/10.1007/978-3-642-20795-2_6
- Remolona, M.F.M., Conway, M.F., Balasubramanian, S., Fan, L., Feng, Z., Gu, T., Kim, H., Nirantar, P.M., Panda, S., Ranabothu, N.R., Rastogi, N., Venkatasubramanian, V., 2017. Hybrid ontology-learning materials engineering system for pharmaceutical products: Multi-label entity recognition and concept detection. Computers & Chemical Engineering. https://doi.org/10.1016/j.compchemeng.2017.03.012
- Skvortsov, N.A., Kalinichenko, L.A., Kovalev, D.Y., 2016. Conceptualization of Methods and Experiments in Data Intensive Research Domains, in: Data Analytics and Management in Data Intensive Domains, Communications in Computer and Information Science. Presented at the XVIII International Conference on Data Analytics and Management in Data Intensive Domains, Springer, Cham, pp. 3–17. https://doi.org/10.1007/978-3-319-57135-5_1
- van der Aalst, W.M.P., 2016. Process Mining: Data Science in Action. Springer.
- Wang, T., 2015. A study to define an automatic model transformation approach based on semantic and syntactic comparisons. Ecole nationale des Mines d'Albi-Carmaux.
- Zhang, Z., Gonzalez, M.C., Morse, S.S., Venkatasubramanian, V., 2017. Knowledge Management Framework for Emerging Infectious Diseases Preparedness and Response: Design and Development of Public Health Document Ontology. JMIR Res Protoc 6. https://doi.org/10.2196/resprot.7904