Prof. Mathieu d'Aquin
Full Professor of computer science
LORIA/IDMC
Université de Lorraine, Nancy, France

22nd May 2023

The habilitation thesis of Dr. Jakub Klímek presents a collection of research works realised over the last few years on the topics of linked data production and consumption, especially focusing on tool support to facilitate the adoption of linked open data technologies and principles by experts and non-experts alike. The thesis starts with an overview of the field and of the main contributions from Dr. Klímek's work, followed by 10 articles detailing those contributions.

I am being asked to comment on the results of the plagiarism check included with the thesis for review in this report. I find that most of what is reported points to the published articles included in the thesis. In other words, the introductory and overview content included in the thesis is clearly original in its writing, while, unsurprisingly, the part of the thesis made up of already published material clearly refers to that already published material.

Concerning the research itself, Dr. Klímek's work looks into aspects of linked data that are critical to its adoption and development from a pragmatic and practical point of view. It has indeed, for many years, been identified that a key barrier to the full realisation of the vision of a truly linked and open data-based web is that it requires significant effort and expertise from a large number of people and users to produce and use linked data sources. This comes from a number of factors, including in particular that kinked data, in its most common form, is based on technologies that are not well understood by even IT professionals and that do not integrate well with common tools and workflows. Another key aspect, well identified in Dr. Klímek's work, is the difficulty in exploiting data in a linked data format. Indeed, while producing and publishing linked open data requires significant investment, the incentive for it can be hampered by the fact that more time and effort are needed to create applications and demonstrate the usefulness of the process. The work presented in this habilitation thesis addresses those field-wide issues through mostly three lines of work: 1- Tool support for the production of linked data from different sources, 2- tool support for the creation of visualisation and applications of existing linked data sources and, 3- practical use-cases demonstrating and showcasing the process, and allowing researchers to further explore the issue of linked open data adoption.

Regarding the first line of work, an overall process for linked data provisioning is presented, showing the need for support at various levels and through various components. It is clear that the most significant overall contribution to this field is materialised by the design and development of the LinkedPipes ETL tool, enabling the creation of reproducible workflows for data transformation (in particular from more traditional formats, such as CSV files, into RDF, the main data model for linked data). The tool in itself is impressive in the sense that its development must have required taking into account many different challenging aspects, from useability to efficiency in manipulating large amounts of data. On that last point, I particularly enjoyed the work on applying data chunking to enable robust processes and large datasets to be manipulated in ETL workflows without requiring unattainable amounts of memory. This work has clearly shown its value through comparing with baselines and with other approaches (especially based on streaming data). I would actually have been interested in further discussions on the theoretical aspects, including under which conditions those approaches are applicable. Another aspect on which I would have appreciated more discussion is on the choice of creating an entirely new tool. While I do not doubt that valid reasons exist, there are many commercial ETL tools in existence that would all have to deal with a subset of the challenges that had to be tackled in the creation of LinkedPipes ETL too. Those tools are generally entirely based on a tabular representation of data and are, in many cases, highly optimised for complex manipulations of large scale data from a large number of potential sources. This is obviously too much to ask, but I would be interested in exchanging with Dr. Klímek on the value of the LinkedPipes ETL tool as it exists, in comparison with the possibility to create extensions and plugins to existing, popular ETL tools. LinkedPipes ETL is evaluated through user satisfaction, and some technical aspects and comparisons with other Linked

Data-related approaches are provided. Adding the view from traditional, non-linked data ETL could help further validate some of the fundamental choices made in its development.

On the linked data consumption side, an overview of the components, processes and challenges is also provided in the thesis. In addition, a survey of existing tools (dubbed “Linked Data Consumption Platforms”) is provided where the authors systematically review multiple tools (including their own) designed to help experts and non-experts in making use of linked (open) data sources. This survey is really interesting as it is based on identifying a set of components that could be included in such platforms, and of criteria to evaluate them. The main message I retain from it is also that the entire field is very scarce, with very few tools existing and most of them supporting a very small part of the overall process. Another interesting aspect of this work is that the tools were evaluated by both linked data experts and non-experts, further demonstrating how linked data suffers from a high barrier of entry and how, in their current state, tools do not yet achieve to overcome this issue to a satisfying degree. I therefore hope and expect that this survey is used by tool developers as a guideline for where to focus their effort. Other than the survey, the focus of Dr. Klímek's work has been on the creation of tools to support the easy creation of visualisation, applications and extraction of/from linked data sources, namely LinkedPipes Visualisation, LinkedPipes Applications, LinkedPipes DCAT-AP Viewer and Simplod. The last two are, not in a negative way, simpler utilities that help in exploiting linked data sources by making it convenient to see and comprehend what exists and to extract data in a format that is more readily usable with common tools. An interesting aspect of DCAT-AP Viewer is that it is another tool that is comparable to other, highly visible and popular platforms (e.g., CKAN) that have been built without relying on linked data technologies and principles. On the other hand, the idea behind LinkedPipes Visualisation and Application is very appealing: To have a mechanism to quickly apply linked data in ways that are useful to users who are not necessarily interested in knowing and understanding about linked data principles. I find the idea of using the schema, and therefore to an extent the semantics attached to the data, and to align it with the expected input of available applications particularly interesting. From the provided material, it is not completely obvious how easy the process is in practice, especially considering that an ETL process is being put in place to transform the data from its linked data sources to the expected format from the application, but the potential of being able to quickly know and to semi-automatically match data to their possible exploitation is certainly high. All of those works are evaluated considering different aspects, including use satisfaction. Those tools have also been used in a number of use cases (see below) and it would have been interesting to have a deeper discussion on their impact in those use cases.

Finally, a key aspect of Dr. Klímek's work is the realisation of a number of use cases for the processes and tools presented, the most prominent of which being, I believe, the release and use of Open Data from the Czech Government. Those have obvious value as case studies, in enabling others to rely on the lessons learnt from those works. An aspect that I find very interesting also is the evident co-evolution of the use cases and the tools, demonstrating how tackling concrete, specific issues in particular scenarios can help devising general and yet practical solutions to field-wide research issues.

The field of Linked Open Data has been evolving over the last decade, with the objective to demonstrate global societal and economic impact. There are many barriers in achieving that, which stem from the specifics of the technologies and principles used. The research conducted by Dr. Klímek's is important in the sense that it is tackling those barriers from a practical and pragmatic standpoint, providing tools to lift them and exemplar cases where they are being addressed.

Prof. Mathieu d’Aquin

mathieu.daquin@loria.fr - @mdaquin - +33749630574