



2. Publishable summary

LarKC in a nutshell

The LarKC platform is a pluggable RDF-aware workflow platform in which it is possible to exploit techniques and heuristics from diverse areas such as databases, logic, machine learning and information retrieval in order to provide flexible and scalable data processing for very large scale Semantic Web applications. LarKC supports both local and remote execution of plugins. Remote execution can be either through lightweight servlets on regular web-servers, or through dedicated grid-protocols for execution on large compute clusters. LarKC reduces the startup cost of experimenting with and applying web scale reasoning.

Main Innovations

In LarKC, large-scale, distributed and possible incomplete reasoning is performed over web-scale knowledge sources. Massive inference is achieved by distributing problems across heterogeneous computing resources and coordinated by the LarKC platform

This is achieved by:

- Releasing an evolving open-source, publically developed platform and plugin code, available at LarKC.sourceforge.net and www.larkc.eu/plugin-marketplace.
- Enriching current logic-based Semantic Web reasoning with methods from information retrieval, machine learning, information theory, databases, streamed reasoning, distributed reasoning, and probabilistic reasoning.
- Employing cognitively inspired approaches and techniques such as spreading activation, focus of attention, reinforcement, habituation, exploration and stopping rules, relevance reasoning, and bounded rationality.
- Achieving scalability through giving up completeness: web search does not guarantee to produce all possible relevant pages, and web reasoning cannot, and need not, attempt to produce all possible answers. Significant speedups can be obtained by incompleteness in many stages of the reasoning process, ranging from selection of the axioms to incomplete reasoning over those axioms.
- **Achieving scalability through parallelization.** LarKC enables platform users to explore both tight integration of parallel processes on a high-performance computing cluster, or less tightly coupled local-area and wide-area distributed computing.

Result

LarKC is not just a single reasoning engine, but a configurable platform that supports all phases of reasoning about knowledge on the web, by allowing components that are plugged in to implement use cases with different scale/efficiency trade-offs. The resulting plugin architecture has been used to support: real-time analysis and interpretation of the data concerning a city infrastructure, data integration for early clinical drug development, and the production of reference works on carcinogenic risk factors. By supporting public use of the platform and plugin market place, we expect to see increasing use of LarKC outside the consortium to drive research and innovation in applications of parallel and approximate reasoning over the web. LarKC is currently being used as the infrastructure for OpenPhacts, a large EU-funded funded under the Innovative Medicine Initiative (IMI), as well as in Data2Semantics, a project funded under the Dutch national COMMIT programme.