Advanced Data Mining and Integration Research for Europe

Advanced Data Mining and Integration Research for Europe

Malcolm Atkinson[1], Peter Brezany[2], Oscar Corcho[3], Liangxiu Han[1], Jano van Hemert[1,*], Ladislav Hluchý[4], Ally Hume[5], Ivan Janciak[2], Amy Krause[5], Dave Snelling[6], Alex Wöhrer[2]

[1]National e-Science Centre, School of Informatics, University of Edinburgh, United Kingdom
[2]Institute of Scientific Computing, Faculty of Computer Science, University of Vienna, Austria
[3]Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Spain
[4]Institute of Informatics, Slovak Academy of Sciences (IISAS), Bratislava, Slovakia
[5]EPCC, University of Edinburgh, United Kingdom
[6]Fujitsu Laboratories Europe Limited, United Kingdom
[*]*Corresponding author:* j.vanhemert@ed.ac.uk

Motivation: There is a rapidly growing wealth of data [1]. The number of sources of data is increasing, while, at the same time, the diversity, complexity and scale of these data resources are also increasing dramatically. This cornucopia of data offers much potential; a combinatorial explosion of opportunities for knowledge discovery, improved decisions and better policies. Today, most of these opportunities are not realised because composing data from multiple sources and extracting information is too difficult. Every business, organisation and government faces problems that can only be addressed successfully if we improve our techniques for exploiting the data we gather.

Strategy: We report the rationale for a new architecture, *DMI architecture*, for combined data integration and data mining under development in the Advanced Data Mining and Integration Research for Europe (ADMIRE) project. The DMI architecture is intended to enable society to make better use of the rapidly expanding wealth of data. The proposed DMI architecture must make all of the stages of DMI process development and enactment as identified in [2] easier and more economic.

The ADMIRE project aims to significantly improve the exploitation of data by delivering three categories of output: a framework, an architecture and a set of use cases that illustrate how they can be used to improve DMI. These will be built on a consistent set of principles:

- The scale and complexity of each data source grows. The DMI architecture addresses this with dataflow technology to reduce data handling and to move data reduction and transformation operations closer to data sources.
- The number and variety of independent data sources increases; warehousing and virtualisation become infeasible at the envisaged scale, which we address by proposing dynamic composition of processes.
- The computational complexity of extracting information grows as a result of the above and of increasingly sophisticated application requirements. The DMI architecture addresses this by enabling the work of *data-aware distributed computing* (DADC) engineers and by supporting the incremental definition and revision of libraries and patterns.
- The number of application domains using DMI grows, becomes more diverse and engages more users. The DMI architecture addresses this by recognising communities of users, by supporting them with their own environments and by delivering packaged production versions of DMI processes.
- The number of experts involved in developing new DMI processes and supporting their application grows. The DMI architecture addresses this by separating support for DMI experts from that for DADC engineers and application-domain users. Support for communities with aligned DMI interests is achieved by sharing between DMI-developers' workbenches via a common registry for their community.
- The number of providers of data and DMI services grows. The DMI architecture separates the organisation of environments for DMI-process development from the complexities of DMI-service provision by interposing DMI gateways using a canonical language.
- The growing sophistication of information extraction from large bodies of data requires ever more complex and refined workflows. The DMI architecture addresses this by structuring collections of components into libraries that correspond to a conceptual structure captured in DMI ontologies and by supporting the incremental refinement of libraries and the DMI processes that use them.
- The providers of data and services autonomously change their offered services and schema at a rate which defies manual adaptation when many data resources are in use. The DMI architecture proposes to exploit type systems, semantic description, community effort and light-weight composition to semi-automatically adapt to change and to pool the intelligence of human interventions.

Overview: The consequences of the above architectural decisions are summarised in this section. The three layers of abstraction lead to a communication via a restricted canonical form between a user-oriented tools layer and a system and provider-oriented platform layer, as is illustrated diagrammatically below.

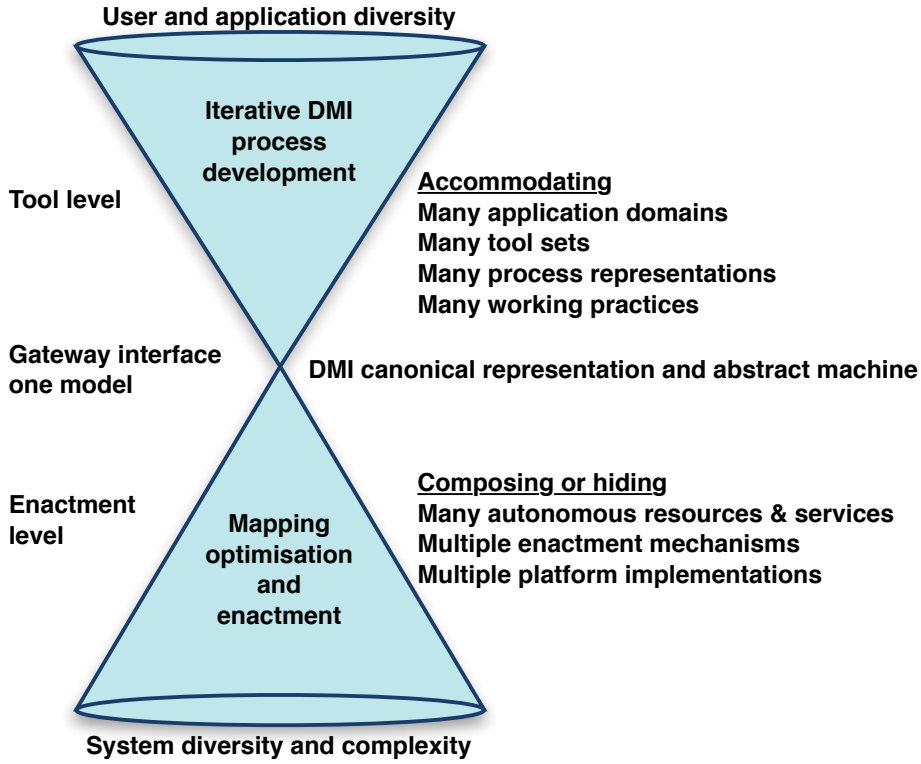

Figure 1: Separating DMI levels of diversity

Figure 1 shows how the complexities of matching the diversity of user requirements at the tools level can be separated from the complexity of the enactment level, accommodating the diversity of data resources and services, by interposing the single canonical domain of discourse represented by an appropriate DMI language. The ADMIRE hypothesis is that, by enforcing this *logical* decoupling, both the tools development and the platform engineering will proceed rapidly and independently. Of course, this depends on the quality of the abstract machine and the language operating at the gateway. Developing that quality is one of ADMIRE's research goals.

To allow providers of DMI services to amortise and smooth their costs over many communities of users and to allow selection of DMI services from multiple providers, the DMI architecture provides a many-to-many relationship between workbenches and portals and the DMI gateways.

A community of developers will use a number of workbenches, e.g. *A, B & C*, and one registry *1* that holds descriptions of the DMI components they are either developing or have obtained from gateways, e.g., *a* and *b*. Some of these descriptions will refer to representations of implementations in a repository *N*. Each gateway has its own registry, which describes all of the resources, services and components it is able to work with. Requests to a gateway can interrogate this information, can update it and can cause DMI processes that use the gateway's capabilities to be enacted. A gateway may present services that are obtained by delegating work to other gateways; shown as *a* delegating to *c* in Figure 2.

Each gateway has the need for its services to evolve independently. We envisage a large, distributed population of DMI gateways with a variety of owners and purposes, and with varying policies on the adoption of new versions. Hence, a unique registry is associated with each gateway. Workbench *C* may be connected to both *a* and *b* so that it can draw on different services provided by these, or to deliver its definitions of DMI components and processes to both *a* and *c*, or to mediate the transfer of descriptions between *a* and *c*.

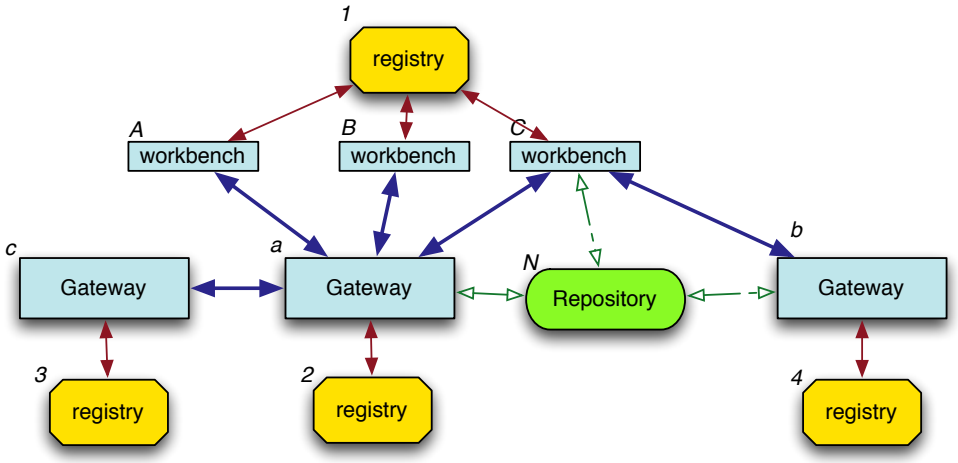

Figure 2: The relationships between workbenches, gateways, registries and repositories

Impact: ADMIRE will develop a model, language and implementation framework that has the following significant practical benefits:

- By providing DMI-experts with a well equipped workbench, a powerful kit of DMI components and an enactment technology that takes responsibility for many distributed system and data heterogeneity issues, ADMIRE will allow those experts to focus on building methods and tools for domain experts. This will improve the quality of systems used, accelerate experiments and exploration of methods and data, and increase the DMI-experts' productivity. *Ingenuity and insight will remain critical; it should be more effectively deployed as distractions are eliminated by consistency and automation.*
- By providing more rapid provision of tailored tool sets and by taking care of operational and performance issues, the ADMIRE strategy will accelerate the pace with which domain experts explore and mine compositions of heterogeneous data. The improved framework should allow them to focus on hypotheses and questions in their domain and improve their productivity. *Knowledge of the domain and expertise in formulating hypotheses and deploying methods will remain crucially important; acceleration of enactment and reduction of extraneous detail should allow these to be better exploited.*
- By providing a computational framework supporting standard computational patterns, multi-scale data flow and composition, parallelisation, distribution, recovery, adaptation, dynamic code placement and optimisation, ADMIRE will present a context that stimulates innovation of data mining and integration methods. This stimulation should subsequently extend to other communities developing algorithms in any application domain that exploits distributed data and that designs and conducts complex analyses.
- Through automation driven by component descriptions ADMIRE will allow abstract descriptions of a very wide range of DMI process requests to be submitted to a DMI service. This will enable the provision of utility services that enact those DMI processes.

The principal innovations of the DMI architecture are: (1) a de-coupling of the enactment technology from the tools used to prepare DMI processes, which in turn enables (2) multiple independent DMI enactment services, some of which may be tightly coupled with curated data collections, and (3) the enactment of each DMI process by distributing it over these services, c.f. distributed queries.

The prototype of the architecture is being evaluated with applications from: customer relationship management, flood modelling and disaster mitigation, gene expression map analysis and data-centre failure reduction. (Experience with the prototype and this evaluation will be reported in the full paper.)

Acknowledgements

The EU Framework Programme 7 FP7-ICT-215024 funding of the ADMIRE project is key to bringing the partners together and to undertaking the research (`www.admire-project.eu`).

References

[1] Gordon Bell, Tony Hey, and Alex Szalay. COMPUTER SCIENCE: Beyond the Data Deluge. *Science*, 323(5919):1297–1298, 2009.

[2] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. The CRISP-DM reference model. Technical report, The CRISP-DM Consortium, August 2000.