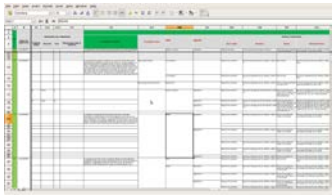
Data Integration in Dataspaces

Knowledge Lab

Birkbeck UNIVERSITY OF LONDON

Research Student
Richard Brownlow

Supervisors
Alex Poulovassilis
Nigel Martin



Birkbeck UNIVERSITY OF LONDON

Department of Computer Science and Information Systems

Research Aims

Our research is motivated by the problems facing data integrators in multi-disciplinary data integration (DI) projects when adopting traditional DI methodologies and tools. Traditional DI approaches require all the semantic mappings between the different data sources to be determined before data services can be supported by the integrated resource, whereas domain data and knowledge are likely to be incrementally gathered and highly evolving, particularly in the early stages of multi-disciplinary DI projects. As a result, such DI projects are often costly and risky. To address these challenges, we are developing techniques for **lightweight data integration** in an **incremental pay-as-you-go methodology.**

Research Methodology

We undertake integration of heterogeneous data sources through the incremental specification of **semantic overlaps** between them, which we term **intersection schemas**, using a graphical tool underpinned by a formal schema transformation language. The data sources are automatically integrated under a virtual global schema or ontology after each iteration of the integration process, and data services can be supported incrementally after each iteration. We are also able to handle scenarios where different domain experts disagree on how to model the knowledge domain, which we term *multiple integration viewpoints*, utilising a rules-based engine to dynamically select the appropriate viewpoint at query runtime (see Figure 1).

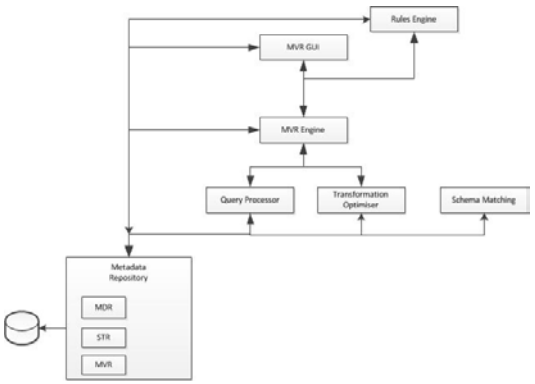


Figure 1. Software Architecture

Research Approach

We have developed a graphical tool (see Figure 2) that supports all steps of this process, and have conducted a preliminary evaluation of our methodology through a case study in the Weaving Domain, utilising a subset of the datasets from the **Weaving Communities of Practice** project. We have demonstrated the potential benefits in terms of increased productivity and the ability to run high priority queries at a much earlier stage of the integration lifecycle. This allows earlier validation of both data and queries by the domain experts, with the potential to reduce both the cost and the risk of DI projects. We are currently carrying out a more extensive evaluation using a fuller dataset from the Weaving Communities of Practice project, aiming to demonstrate its usability and scalability.

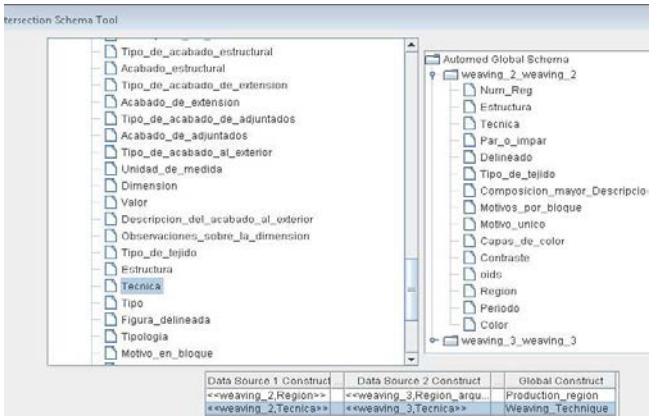


Figure 2. Data Integration tool

Publications

R.Brownlow, A.Poulovassilis: Intersection Schemas as a Data-space integration Technique. Proceedings of EDBT/ICDT Workshops 2014, pp 92-99.

R.Brownlow, S.Capuzzi, S.Helmer, L,Martins, I.Normann, A.Poulovassilis: An Ontological Approach to Creating an Andean Weaving Knowledge Base. ACM Journal on Computing and Cultural Heritage, 8(2), 11:1-11:31 (2015)