Description Logics for Conceptual Design, Information Access, and Ontology Integration: Research Trends

Enrico Franconi
Department of Computer Science, University of Manchester
Oxford Rd., Manchester M13 9PL, UK
Phone: +44 (161) 275 6170; Fax: +44 (161) 275 6204
franconi@cs.man.ac.uk
http://www.cs.man.ac.uk/~franconi/

1 Introduction

In recent years, data and knowledge base applications have progressively converged towards integrated technologies that try to overcome the limits of each single discipline. Research in Knowledge Representation (KR) originally concentrated around formalisms that are typically tuned to deal with relatively small knowledge bases, but provide powerful deduction services, and the language to structure information is highly expressive; research on formal languages for ontologies was originated from KR. In contrast, Information Systems and Database research mainly dealt with efficient storage and retrieval with powerful query languages, and with sharing and displaying large amounts of (multimedia) documents. However, data representations were relatively simple and flat, and reasoning over the structure and the content of the documents played only a minor role.

This distinction between the requirements in Knowledge Representation and Databases is vanishing rapidly. On the one hand, to be useful in realistic applications, such as the applications in the semantic web, a modern ontology KR system must be able to handle large data sets, and to provide expressive query languages. This suggests that techniques developed in the DB area could be useful for ontologies. On the other hand, the information stored on the web, in digital libraries, and in data warehouses is now very complex and with deep semantic structures, thus requiring more intelligent modelling languages and methodologies, and reasoning services on those complex representations to support design, management, retrieval, and integration. Therefore, a great call for an integrated view of Knowledge Representation and Database technologies is emerging.

Description Logics (DL) [BN02] are a very promising research area in KR with applications in DBs. The main effort of the research in DL is in providing both theories and systems for expressing structured knowledge and for accessing and reasoning with it in a principled way [CDLN02, Don02]. Recently, basic progress has been made by establishing the theoretical foundations for the effective use of DL in information systems [Bor95, BLR02]. DL offer promising formalisms for solving several problems concerning Conceptual Data Modelling and Ontology Design (see, e.g., [CLN98, BB02], or the OIL and DAML+OIL efforts [FHvH$^+$00, IH02]), Intelligent Information Access and Query processing (see, e.g., [BB93, LR98, BNP00, Fra00]), and Information Integration (see, e.g., [CGL$^+$98, JQC$^+$00, MIKS00, GLR00]).

This tutorial will have a popular style showing research trends, rather than a strictly theoretical one. Its aim is to let the audience understand why DL and DB technologies could be useful to semantic web research and applications, and it will mostly make use of examples. Nonetheless, precise links to the important theoretical results and to the relevant references will be given.

In the tutorial I will argue that good *Conceptual Modelling* and *Ontology Design* is required to support powerful *Query Management* and to allow for semantic based *Information Integration*. Therefore, the tutorial has been structured into three parts. In the first part (described in Section 2), an extended ontology language and a methodology for conceptual and ontology design will be introduced. A demo of an ontology design tool will be given. In the second part (described in Section 3), the query management problem in the presence of the previously devised conceptual model will be considered: a global framework will be introduced, together with various basic tasks involved in information access. In the last part (described in Section 4), general issues about ontology integration will be presented.

About the tutorial speaker.

Enrico Franconi is Senior Lecturer (Associate Professor) at the University of Manchester, Department of Computer Science, in the Information Management Group. He is currently involved as Principal Investigator of the European IST project "*Semantic Webs and Agents in Integrated Economies*" (SEWASIE), as Principal Investigator of the British Research Council (EPSRC) project "*Knowledge Representation meets Databases*" (KRDB), and as co-investigator in the EPSRC project "*Flexible source integration in distributed knowledge-based query processing for bioinformatic information sources*" (TAMBIS-II). In the past he has been Principal Investigator in various projects on the foundations of information systems and ontologies: European ESPRIT-4 Long Term Research project "*Foundations of Data Warehouse Quality*" (DWQ); Italian National Research Council (CNR) projects "*Ontological tools for the Management and the Integration of Heterogeneous Knowledge*", "*Ontological and Linguistic Tools for Conceptual Modelling*', and "*Hybrid Knowledge Representation Systems*"; Italian Space Agency (ASI) project "*Integration and Access to Heterogeneous Databases*".

Recently he chaired the 1998 International *Description Logics* Workshop (DL'98); the 6th International Workshop on *Knowledge Representation meets Databases* (KRDB'99); the PC of the 12th European Summer School in *Logic, Language and Information* (ESSLLI'2000); the International Workshop on *Foundations of Models for Information Integration* (FMII-2001), which is the 10th workshop in the series *Foundations of Models and Languages for Data and Objects* (FMLDO).

2 Conceptual Modelling and Ontology Design

For the purpose of this tutorial, an Ontology will be considered as a Conceptual Schema expressed in a suitable conceptual data model (i.e., an Ontology Language). Good *conceptual data models* put their emphasis on the correct and semantically rich representation of *complex* properties and relations that may exist between documents. They should allow for an abstract representation of data which resembles the way they are actually perceived and used in the real world, thus shortening (with respect to the more traditional data models) the semantic gap between the domain and its representation.

Conceptual (or Ontology) modelling deals with the question on how to describe in a declarative and reusable way the domain information of an application, its relevant vocabulary, and how to constrain the use the data, by understanding what can be drawn from it. Recently, a number of conceptual and ontology modelling languages has emerged as de-facto standard, in particular we mention Entity/Relationship (ER) for the relational data model, UML and ODMG for the object oriented data model, and XML, RDF and DAML+OIL for the web semi-structured data model. Still, many such languages do not have a formal semantics based on logic, or reasoners built upon them to support the designer. Not surprisingly, conceptual modelling tasks have always been in the mainstream of KR research – see for example the research on Ontology representation and design – and can be considered now one of the main applications of KR languages and reasoning techniques [BB02]. DL can be considered as an unifying formalism, since they allow the logical reconstruction and the extension of representational tools such as object-oriented data models (e.g., UML and ODMG), semantic data models (e.g., Entity/Relationship and ORM), frame-based ontology languages (e.g., OIL and DAML+OIL) [CLN98, CLN99, CCDGL02, FHvH$^+$00]. In addition, given the high complexity of the modelling task when complex data is involved, in the semantic web field there is the demand of more sophisticated and expressive languages than for normal information systems. Again, DL research is very active in providing expressive ontology languages to capture various aspects of the information (see, e.g., [AF99, FGM00, FS99, BKW02]).

In this tutorial I will present examples using a generic conceptual data model. I will point out how it generalises both the object-oriented data model based on UML class diagrams and the extended Entity-Relationship (EER) semantic data model, and how it is strictly related to OIL and DAML+OIL. The ontology language includes *taxonomic* relations to state containment assertions between entities and between relationships with the possibility to specify additional *covering* and *disjointness* constraints. The most interesting feature of the modelling language is the ability to completely *define* entities and relationships as *views* over other entities and relationships of the ontology [CLN98]. The adopted view language is DLR [CGL$^+$98], a Description Logic over unary and n-ary relationships. DLR is an interesting decidable fragment of first order logic: among others, inclusion dependencies with DLR views can express (a) unary inclusion dependencies, (b) typed inclusion dependencies without projection, (c) existence dependencies, (d) exclusion dependencies, and (e) full key dependencies. DLR is powerful enough to encode the full EER, the UML class diagrams and most of DAML+OIL. An informal introduction to the properties of the DLR Description Logic will be given.

Two additional extensions to the conceptual data model will be also considered. The first one is with multidimensional aggregations – that is, the conceptual data model is able to represent the structure of *aggregated entities* and of *multiply hierarchically organised dimensions*. The ability of representing aggregations at the conceptual level is crucial in modelling structured documents in data warehouses, in the semantic web and in digital libraries. The second one allows for the representation of standard temporal operators for temporal conceptual modelling and of a large class of temporal integrity constraints, useful to model the dynamics in the sematic web.

At the end of this first part, a demo of the i.com tool [FN00, JQC$^+$00] – which implements the above conceptual data model as UML class diagrams or EER schemas – will be given. i.com allows for the specification of multiple EER (or UML) diagrams and inter- and intra-schema constraints. Complete logical reasoning is employed by the tool using an underlying DL inference engine to verify the specification, infer implicit facts and stricter constraints, and manifest any inconsistencies during the conceptual modelling phase.

3 Information Access

Only recently has KR research started to have an interest in query processing and information access. Recent work has come up with advanced reasoning techniques for query evaluation and rewriting using views under the constraints given by the ontology – also called view-based query processing [Ull97, CGLV00]. This means that the notion of accessing information through the navigation of an Ontology modelling the document's domain – which can be seen as a conceptual schema – has its formal foundations.

In this tutorial I will thus consider DL for formalising not only the ontology but also the query processing as well. The (DL-based) conceptual schema as defined in the previous section can be seen as a set of constraints over a vocabulary which is usually richer that the logical schema of the information system it is modelling. In some sense, quite often the conceptual schema plays the role of an general ontology of the domain, very close to the user's rich vocabulary, rather than of a set of constraints over the poor logical vocabulary structuring the data. With this perspective in mind, the user would prefer to query the information system using the richer vocabulary of the ontology. The vocabulary of the basic data (i.e., the logical schema) could be seen in turn either as a subset of the conceptual vocabulary – this is the simplistic view – or more generally as a set of (materialised) views over the vocabulary of the ontology. However, in this case we have to solve the problem of view-based query processing. The problem requires to answer a query posed to a database – the one defined by the ontology – only on the basis of the information in a set of (materialised) views, which are again queries over the same database. In the process, the information contained in the conceptual schema of the database should be of course taken into account.

I will introduce the two approaches to view-based query processing, namely query rewriting (see, e.g., [BLR97]) and query answering (see, e.g., [AD98, CGL00]). In the former approach, we are given a query Q, a set of view definitions characterising the actual data, and a set of (conceptual) constraints – all over the conceptual vocabulary – and the goal is to reformulate the query into an expression, the rewriting, that refers only to the views, and provides the answer to Q. Typically, the rewriting is formulated in the same language used for the query and the views. In the latter approach, besides Q, the view definitions and the constraints, we are also given the extensions of the (materialised) views. The goal is to compute the set of tuples that are implied by these extensions, i.e., the set of tuples that are in the answer set of Q in all the databases that are consistent with the views and the constraints.

This framework can be used to characterise several aspects of an information system. In query optimisation, view-based query processing is relevant because using the views may speed up query processing. In data integration, the views represent the only information sources accessible to answer a query. A data warehouse can be seen as a set of materialised views, and, therefore, query processing reduces to view-based query answering. Finally, since the views provide partial knowledge on the database, view-based query processing can be seen as a special case query answering with incomplete information.

4 Information Integration

In this last part I will show how the technologies introduced in the first two parts, namely a very expressive ontology language and view-based query processing over it, can be used in the framework of Information Integration [CL93, CGL+98, JLVV99, JQC+00].

Let us suppose to have multiple databases to be integrated. Each database will have its own conceptual schema and logical schema, where, as seen in the previous part, the logical schema is just a set of views over the conceptual schema (local-as-view approach). We assume that each symbol of each schema is identified by a unique global symbol, i.e., the various databases have disjoint signatures. Interdependencies between entities and relationships in different schemas are represented by means of integrity constraints involving symbols of the schemas. Such interdependencies are called *inter-model assertions*, and they are of the form of DLR inclusion dependencies. The union of the various schemas with the inter-model assertions and the local views forms the global integrated schema, or the *mediator*. It is worth noting that the integration process is incremental – since the integrated schema can be monotonically refined as soon as there is new understanding of the different component schemas – and that the resulting unified schema is strongly dependent from (actually, it includes) the schemas of the single information sources.

This approach gives both a clear semantics to the integration process of ontologies, and a calculus for deriving inconsistencies and checking the validity of integrity constraints in the integrated schema. Most importantly, in this framework global queries can be defined as views over single ontologies, or they can be generalised to span over multiple ontologies. The view-based query processing mechanism will guarantee the correct answer to the global query from the local sources. In the tutorial a complete worked out example will be given.

The particular but important case of designing a Data Warehouse Conceptual Schema will be presented. In this case it is assumed to have a privileged schema – called the *Enterprise Model* – which is the conceptual representation of the global concepts and relationships reconciled and abstracted in the data warehouse, and it is not necessarily a complete model of all the source information. Such schema is integrated with the different source schemas. The crucial point is that not only the interrelationships between the source schemas and the Enterprise Model are modelled, but also the interdependencies between the source schemas themselves. Moreover, the global integrated schema – the Data Warehouse Conceptual Schema – is composed not only by the Enterprise Model, but also by the various source schemas and by the inter-model assertions. Global data warehouse queries are formally seen as views over the Enterprise Model.

In the tutorial a comparison will be given between the above local-as-view approach to processing global queries and the global-as-view approach, which is more common in current information integration architectures.

References

[AD98] S. Abiteboul and O. Duschka. Complexity of answering queries using materialised views. In *Proc. of the 17th ACM Symp. on Principles of Database Systems (PODS'98)*, pages 254–265, 1998.

[AF99] A. Artale and E. Franconi. Temporal ER modeling with description logics. In *Proc. of the International Conference on Conceptual Modeling (ER'99)*. Springer-Verlag, November 1999.

[BB93] Alexander Borgida and Ronald J. Brachman. Loading data into description reasoners. In *Proc. of 1993 ACM SIGMOD International Conference on Management of Data*, pages 217–226, 1993.

[BB02] A. Borgida and R. J. Brachman. Conceptual modelling with description logics. In Baader et al. [BMNPS02].

[BKW02] Franz Baader, Ralph Kuesters, and Frank Wolter. Extensions to description logics. In Baader et al. [BMNPS02].

[BLR97] C. Beeri, A. Y. Levy, and M.-C. Rousset. Rewriting queries using views in description logics. In *Proc. of the 16th ACM Symp. on Principles of Database Systems (PODS'97)*, pages 99–108, 1997.

[BLR02] A. Borgida, M. Lenzerini, and R. Rosati. Description logics for databases. In Baader et al. [BMNPS02].

[BMNPS02] F. Baader, D. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *Description Logic Handbook: Theory, Implementation and Applications.* Cambridge University Press, 2002.

[BN02] F. Baader and W. Nutt. Basic description logics. In Baader et al. [BMNPS02].

[BNP00] Paolo Bresciani, Michele Nori, and Nicola Pedot. A knowledge based paradigm for querying databases. In *Proc. of DEXA-00*, pages 794–804, 2000.

[Bor95] A. Borgida. Description logics in data management. *TKDE*, 7(5):671–682, 1995.

[CCDGL02] Andrea Calì Diego Calvanese, Giuseppe De Giacomo, and Maurizio Lenzerini. A formal framework for reasoning on UML class diagrams. In *Proc. of the 13th Int. Sym. on Methodologies for Intelligent Systems (ISMIS 2002)*, 2002.

[CDLN02] D. Calvanese, G. De Giacomo, M. Lenzerini, and D. Nardi. Reasoning in expressive description logics. In Baader et al. [BMNPS02].

[CGL$^+$98] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi, and Riccardo Rosati. Information integration: Conceptual modeling and reasoning support. In *Proc. of the 6th Int. Conf. on Cooperative Information Systems (CoopIS'98)*, pages 280–291, 1998.

[CGL00] D. Calvanese, G. De Giacomo, and M. Lenzerini. Answering queries using views over description logics knowledge bases. In *Proc. of the 16th Nat. Conf. on Artificial Intelligence (AAAI 2000)*, 2000.

[CGLV00] D. Calvanese, G. De Giacomo, M. Lenzerini, and Moshe Y. Vardi. View-based query processing and constraint satisfaction. In *Proc. of the 15th IEEE Sym. on Logic in Computer Science (LICS 2000)*, 2000.

[CL93] T. Catarci and M. Lenzerini. Representing and Using Interschema Knowledge in Cooperative Information Systems. *Journal of Intelligent and Cooperative Systems*, 2(4):375–398, 1993.

[CLN98] D. Calvanese, M. Lenzerini, and D. Nardi. Description logics for conceptual data modeling. In J. Chomicki and G. Saake, editors, *Logics for Databases and Information Systems.* Kluwer, 1998.

[CLN99] D. Calvanese, M. Lenzerini, and D. Nardi. Unifying class-based representation formalisms. *J. of Artificial Intelligence Research*, 11:199–240, 1999.

[Don02] F. Donini. Complexity of reasoning. In Baader et al. [BMNPS02].

[FBSV99] E. Franconi, F. Baader, U. Sattler, and P. Vassiliadis. Multidimensional data models and aggregation. In Jarke et al. [JLVV99], chapter 5, pages 87–106.

[FGM00] E. Franconi, F. Grandi, and F. Mandreoli. A semantic approach for schema evolution and versioning in object-oriented databases. In *Proc. of the 1st International Conf. on Computational Logic (CL'2000), DOOD stream.* Springer-Verlag, July 2000.

[FHvH$^+$00] D. Fensel, I. Horrocks, F. van Harmelen, S. Decker, M. Erdmann, and M. Klein. OIL in a nutshell. In *Proceedings of the European Knowledge Acquisition Conference (EKAW-2000)*, Lecture Notes In Artificial Intelligence. Springer-Verlag, 2000.

[FK99] E. Franconi and M. Kifer, editors. *Proceedings of the 6th International Workshop on Knowledge Representation meets Databases (KRDB'99).* Linköping University Technical Report, July 1999. Also electronically available as CEUR Publication, Vol. 21, RWTH Aachen, Germany.

[FN00] E. Franconi and G. Ng. The ICOM tool for intelligent conceptual modelling. In *Proc. of the 7th International Workshop on Knowledge Representation meets Databases (KRDB'2000)*, 2000.

[Fra00] Enrico Franconi. Knowledge representation meets digital libraries. In *Proc. of the 1st DELOS (Network of Excellence on Digital Libraries) workshop on "Information Seeking, Searching and Querying in Digital Libraries"*, 2000.

[FS99] E. Franconi and U. Sattler. A data warehouse conceptual data model for multidimensional aggregation. In *Proceedings of the Workshop on Design and Management of Data Warehouses (DMDW'99)*, 1999.

[GLR00] Francois Goasdoue, Veronique Lattes, and Marie-Christine Rousset. The use of CARIN language and algorithms for information integration: the picsel system. *International Journal on Cooperative Information Systems*, 2000.

[IH02] Christopher A. Welty Ian Horrocks, Deborah L. McGuinness. Digital libraries and web based information systems. In Baader et al. [BMNPS02].

[JLVV99] M. Jarke, M. Lenzerini, Y. Vassilious, and P. Vassiliadis, editors. *Fundamentals of Data Warehousing*. Springer-Verlag, 1999.

[JQC$^+$00] Mathias Jarke, V. Quix, D. Calvanese, Maurizio Lenzerini, Enrico Franconi, S. Ligoudistiano, P. Vassiliadis, and Yannis Vassiliou. Concept based design of data warehouses: The DWQ demonstrators. In *2000 ACM SIGMOD International Conference on Management of Data*, May 2000.

[LR98] Alon Y. Levy and Marie-Christine Rousset. Combining horn rules and description logics in CARIN. *Artificial Intelligence*, 104(1-2):165–209, 1998.

[MIKS00] Eduardo Mena, Arantza Illarramendi, Vipul Kashyap, and Amit P. Sheth. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. *Distributed and Parallel Databases*, 8(2):223–271, 2000.

[Ull97] J. D. Ullman. Information integration using logical views. In *Proc. of the 6th Int. Conf on Database Theory (ICDT'97)*, pages 19–40, 1997.