Abstract. The term Deep Web (sometimes also called Hidden Web) refers to the data content that is created dynamically as the result of a specific search on the Web. In this respect, such content resides outside web pages, and is only accessible through interaction with the web site – typically via HTML forms. It is believed that the size of the Deep Web is several orders of magnitude larger than that of the so-called Surface Web, i.e., the web that is accessible and indexable by search engines.

Usually, data sources accessible through web forms are modeled by relations that require certain fields to be selected – i.e., some fields in the form need to be filled in. These requirements are commonly referred to as access limitations in that access to data can only take place according to given patterns. Besides data accessible through web forms, access limitations may also occur i) in legacy systems where data scattered over several files are wrapped as relational tables, and ii) in the context of Web services, where similar restrictions arise from the distinction between input parameters and output parameters.

In such contexts, computing the answer to a user query cannot be done as in a traditional database; instead, a query plan is needed that provides the best answer possible while complying with the access limitations. In this tutorial, we illustrate the semantics of answers to queries over data sources under access limitations and present techniques for querying answering in this context. We show different techniques to optimize query answering both at the time of the query plan generation and at the time of the execution of the query plan. We analyze the influence of integrity constraints on the sources, of the kind that is usually found in database schemata, on query answering. We present prototype systems that are aimed at querying the deep web, and show their achievements.

1. SCOPE AND AUDIENCE
The tutorial is quite broad, since it touches topics including database theory, logic programming, web information systems, data integration, query answering using views. It is aimed at researchers and practitioners in the areas of web information systems, who want to have a foundational as well as a more practical view on the issues behind the integration and the querying of Deep Web sources. Though some advanced topics are covered, the tutorial is accessible to attendees who have no knowledge of the specific field, and have a generic Computer Science background. However, we select some important topic that we discuss in depth and in detail. Several illustrative examples are shown, that highlight the main issues without going to the deep technical details. At the same time, outlines of the more interesting proofs are given, so as to illustrate relevant techniques that have some general applicability.

2. TUTORIAL ORGANIZATION
2.1 State of the art and introduction

Expected duration: 15 minutes

There are several approaches to querying the Deep Web in the literature.

Different classifications of queries have emerged in this context w.r.t. their potential to retrieve the query answers. In particular, we will distinguish queries whose maximal answer always coincides with the complete answer; these include queries that are executable “as is”, feasible queries (i.e., admitting an executable reordering), and stable queries (i.e., equivalent to an executable query). Stability and feasibility are tightly related with the traditional query containment problem and, in general, can also be cast as an instance of the problem of answering queries using views.

The problem of minimizing the accesses to compute the maximal answer has been addressed for different classes of queries. We shall discuss techniques for dealing with several query classes, including nonmonotonic queries.

Query optimization may also be obtained by query minimization, which is commonly done via rewriting based on query containment. However, the presence of access limitations makes query containment intrinsically harder. We will give hints at techniques for solving the problem and will
point out some of the open issues.

2.2 Modeling and querying the Deep Web

*Expected duration: 25 minutes*

We first present the main formal framework on which most of the tutorial is based. In particular, we model each data source accessible through a web form as a relational table, whose corresponding (relational) predicate has an annotation that specifies a subset of the arguments (attributes) as input (or must-select, or must-bind, hence the term binding pattern for the annotation), and the others as output. In order to access the relation, a query must necessarily make a selection on all the input attributes. An important aspect of modeling the Deep Web is the notion of abstract domain of an attribute: to each attribute (argument of a relational predicate) an abstract domain is associated, that specifies the kind of values appearing at that attribute, in relationship with the system domain. For example, an abstract domain “plate_no” would indicate that the corresponding values are plate numbers; this is essential for optimizing query answering, and the choice of the abstract domain also raises design issues.

Access limitations make query processing significantly harder. In some cases, queries only require a transformation into equivalent but easier-to-execute queries (e.g., by reordering of subgoals); however, in general, an execution via a recursive query plan is needed, even for non-recursive classes of queries, such as conjunctive queries (i.e., select-project-join queries). Moreover, the query answers that can be disclosed (i.e., the maximal answer) are generally only a subset of what would be found if no access limitations were present. We illustrate the basic querying techniques found in the literature, and discuss their relationship with related topics. In particular, Deep Web querying is a form of information integration, where the mediated schema, i.e., the schema that defines the semantics of the whole system, is represented by the sources with access limitations together with the specification of the abstract domains of each attribute.

2.3 Generation of an optimized query plan for the Deep Web

*Expected duration: 20 minutes*

The techniques for generating a naive query plan computing the maximal answer are well-established by now. However, in a web environment, it is crucial to minimize the accesses to remote sources, which are potentially very slow. Notice that a recursive query plan that finds the maximal answer to a conjunctive query over web sources will in general also need to query sources that do not occur in the query, since they may provide useful data for overcoming access limitations. We illustrate a first optimization, consisting in determining which sources are relevant to a given query, so that non-relevant sources can be excluded, thus saving query execution time. The problem of determining relevant sources in the case of conjunctive queries was left open for several years, and solved only recently by means of a graph-theoretic representation of the schema that we discuss in the tutorial. Based on that, we show how to generate, given a conjunctive query, a query plan that executes a minimal set of accesses to sources, according to a notion of minimality that arises naturally in the context of web sources.

2.4 Query containment

*Expected duration: 10 minutes*

We address the problem of checking containment of conjunctive queries in the presence access limitations, which is highly relevant in query optimization. Checking containment in such a setting would amount to checking containment of recursive Datalog programs of a certain class, while, for general Datalog programs, this problem is undecidable. We describe a recently proposed decision procedure for query containment based on the novel notion of crayfish-chase, for which we analyze the achieved computational complexity thereby providing a new insight into the structure of the problem.

2.5 Impact of integrity constraints and runtime optimization of queries

*Expected duration: 10 minutes*

The execution of a query plan under access limitations can be further improved at run time, by making use of the data being extracted from the web sources. In particular, when integrity constraints are enforced locally on the web sources, some of the accesses to sources planned by a query plan may turn out to be irrelevant w.r.t. the extracted data, i.e., immaterial for the computation of new answer tuples. In such cases, the accesses can be avoided. We illustrate how to formalize relevant notions of access minimization for run-time optimization, and how to achieve such minimization during query plan execution.

2.6 Deep Web and information integration

*Expected duration: 10 minutes*

As mentioned above, querying a set of Deep Web sources is a way of integrating the information stored into them. When integrating heterogeneous data sources, the mediated schema serves as a representation of the application domain or, in other words, as a domain ontology; to express such an ontology, database integrity constraints are commonly used. The specification of the semantic relationship between the individual sources and the mediated schema is expressed by means of views. We show techniques for answering queries using views in the presence of access limitations.

3. SUMMARY AND OPEN ISSUES.

In the tutorial, we will discuss several open issues related with the above mentioned techniques for query processing under access limitations. First, we point out that a solution to the minimization problem for the case of sources allowing multiple access patterns is still not available. State-of-art solutions only account for data sources with a single pattern. Second, static techniques for access minimization are not known when integrity constraints are imposed on the schema. The literature only accounts for dynamic solutions that, based on the data being extracted and the constraints in the schema, avoid some of the accesses that can be deemed as irrelevant for computing the query answer. Third, a tight complexity bound for the query containment problem in the
presence of access limitations is not known. In particular, the best available lower bound is the same as for query containment without access limitations. Last, query containment under access limitations together with integrity constraints in the schema is not yet known to be decidable; we show how different formal tools known from the literature might be combined to attack this problem.

4. PRESENTERS’ SHORT BIO.
Andrea Calì is Research Fellow at the Oxford-Man Institute of Quantitative Finance, University of Oxford; he also holds affiliations with the Oxford University Computing Laboratory and Keble College. Previously, he was Assistant Professor at the Department of Computer Science of the Free University of Bolzano, Italy. He holds a M.Sc. in Electronic Engineering and a Ph.D. in Computer Engineering, both from the University of Rome “La Sapienza”. His research interests include database theory, information integration, web data extraction for economics, description logics, conceptual modelling, matchmaking in e-commerce, and mobile information systems. Andrea Calì has worked in several research projects, funded by the European Union and by the British and Italian Government; he also contributed to the design and development of several prototype information systems, among which: IBIS (data integration), DIS@DIS (data integration), IM3 (recommendation system for tourism), SmartDate (AI-based dating system). He worked as consultant for several IT companies, designing information systems.

Davide Martinenghi received his M.Sc. in Computer Engineering from Politecnico di Milano, Italy, in 1998, and his Ph.D. in Computer Science from Roskilde University, Denmark, in 2005. He is currently Assistant Professor at Politecnico di Milano. Previously, he was with the Department of Computer Science of the Free University of Bolzano, Italy. His main interests are data integrity maintenance, data integration, Web data access, and, in a broad sense, applications of logic to data management. He is also interested in research issues related to Web search, including the development of visual paradigms for the representation of queries over Web sources. He has participated in numerous research projects with funding from the European Union and the Italian Government.

5. ACKNOWLEDGMENTS.
Andrea Calì was supported by the EPSRC project “Schema Mappings and Automated Services for Data Integration and Exchange” (EP/E010865/1). Davide Martinenghi was supported by the “Search Computing” (SeCo) project, funded by the ERC under the 2008 Call for “IDEAS Advanced Grants”.

6. REFERENCES