Advances in Data Management - Web Data Integration A.Poulovassilis

1 Integrating Deep Web Data

Traditionally, the web has made available vast amounts of information in *unstructured* form (i.e. text). Searching for this information has utilised techniques from Information Retrieval, in *web search engines*.

However, there is an even larger amount of *structured data* also accessible on the web, arising from data that is stored in databases and is accessed via HTML forms — this is known as the *deep web*.

Integrating such data presents similar problems to those we discussed earlier in the context of heterogeneous databases *but on a much larger scale*: the back-end databases use different modelling languages, different data types, and different representations for the same real-world concepts.

These heterogeneities are manifested in the HTML forms provided for accessing the data, and in the results returned in response to user queries submitted via such forms.

The dynamically-generated web content arising from deep web data cannot be discovered and indexed by traditional search engines.

This poses the question: Is it possible to apply and extend techniques developed for heterogeneous data integration to access this kind of heterogeneous web content through a single query interface?

1.1 Recap - Heterogeneous Data Integration

Figure 1 illustrates the general architecture of a heterogeneous multi-database system, as described in earlier Notes.

Figure 2 illustrates virtual data integration on a larger scale — possibly web-scale — in which the manual definition of mappings between a Global Schema and Local Schemas is no longer feasible (even if it is undertaken using an intelligent GUI). Moreover, the data sources may not export a precise database schema in the traditional sense, but instead a looser representation of their content.

In such large-scale data integration scenarios, *automatic schema matching and schema mapping* is required:

- **Schema matching** identifies semantic correspondences between elements from different schemas (these correspondences may be 1-1, 1-n, n-1, n-m).
- **Schema mapping** generates the GAV/LAV/GLAV mappings that link the global schema with the local schemas.

Heterogeneous DB Integration

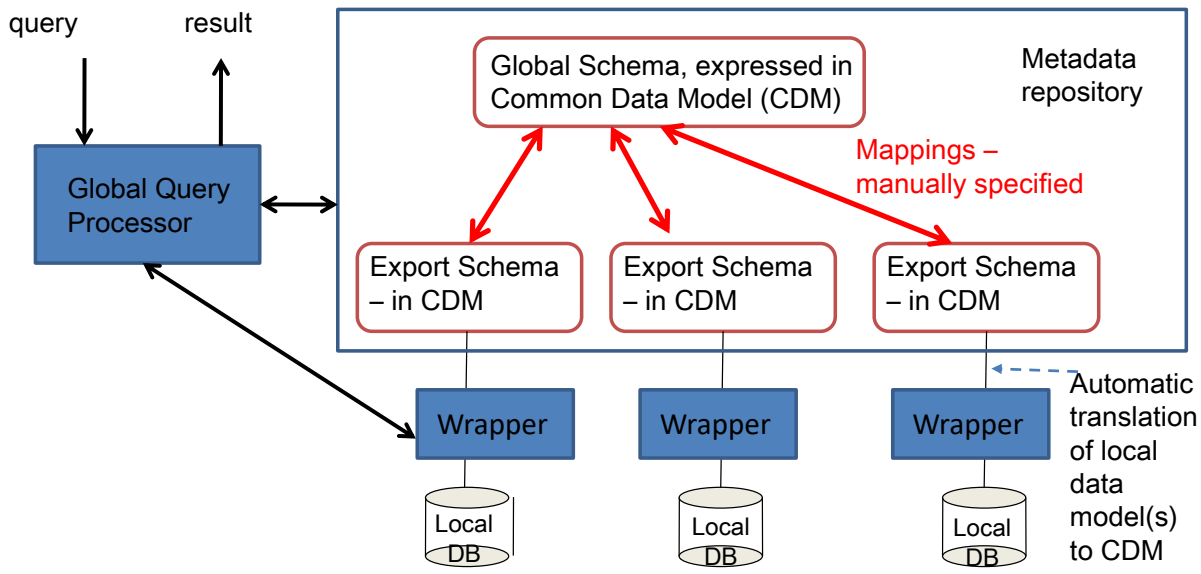


Figure 1: Heterogeneous Database Integration

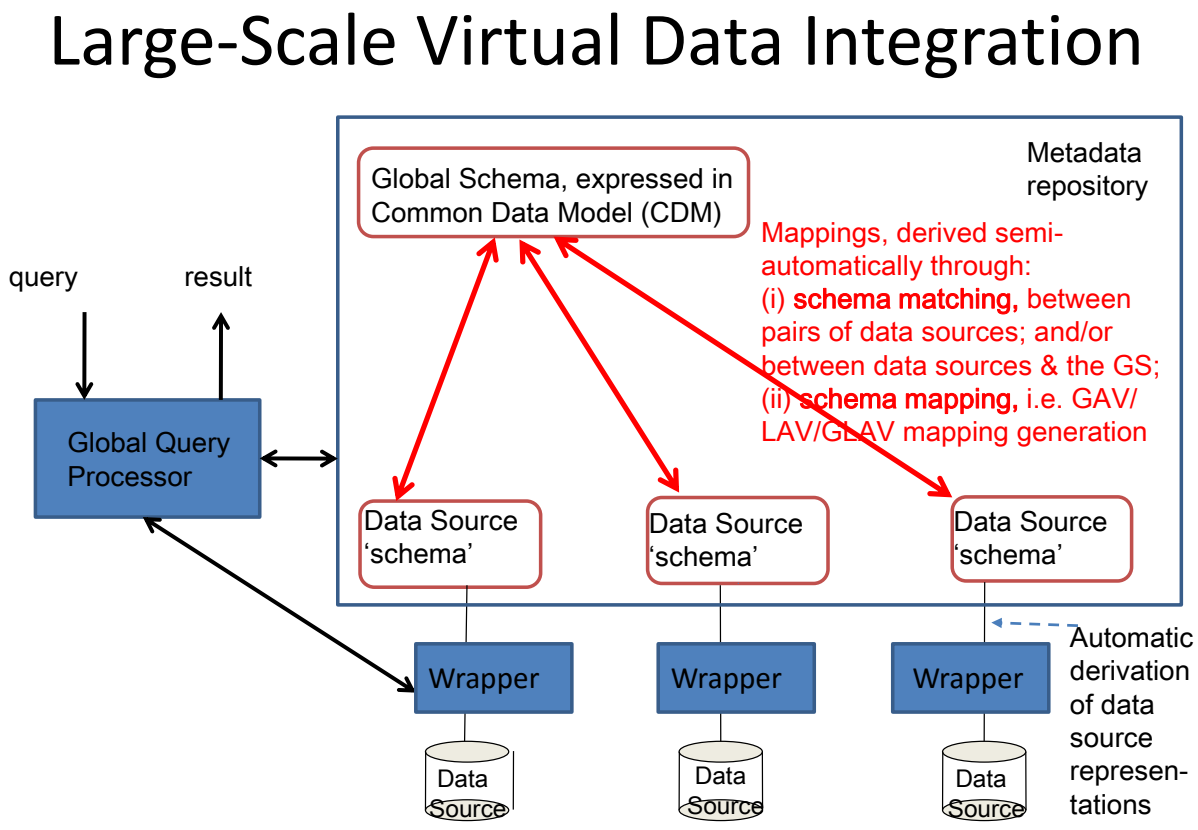


Figure 2: Large Scale Virtual Data Integration

1.2 Domain-specific meta-search engines

One early area of research in web data integration were techniques for automatically generating **domain-specific meta-search engines**:

These aim to provide domain-specific search that utilises multiple source search engines.

One of the major steps involved in automatically generating such tools is the task of **schema matching** and **schema mapping**, as defined earlier.

In this context, the "schema" is the format of the HTML form that allows users to access information retrieved via each source search engine.

The volumes and heterogeneity of Web data require the application of automatic schema matching and mapping techniques.

In the approach by Naz et al. 2009 (see Reading 3), a domain ontology serves as a unifying reference point for the schema matching and mapping process.

Several schema matching techniques are used to automatically resolve semantic heterogeneities between the source search engine schemas and to generate a global unifying schema.

The schema matching techniques applied include a variety of *element-level*, *structure-level* and *ontology-based* techniques.

(See Reading 3 for a description of these techniques, and for references to more general survey articles on schema matching and mapping techniques. See also a more recent overview by Bernstein et al. 2011[1].)

The mappings between the source search engine schemas and the global schema are also derived automatically.

These mappings are then used to:

- automatically generate an integrated meta-search query interface;
- support query processing in the meta-search engine (which interacts with the individual source search engines);
- resolve semantic conflicts arising during result extraction from the source search engines, so as to present an integrated, harmonised set of results to the user.

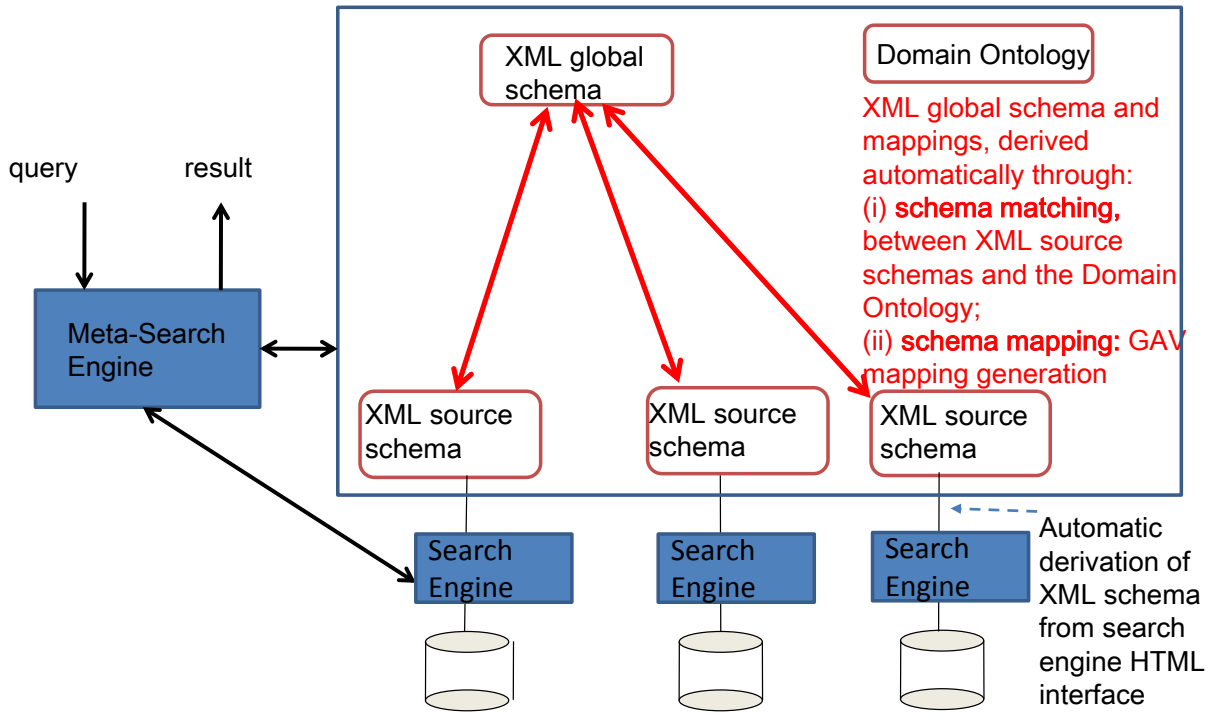
1.3 Web-scale data integration through "surfacing"

Beyond domain-specific techniques, the Pay-As-You-Go (PAYGO) approach proposed by researchers at Google has been every influential[2]. With this approach:

[1] P.A.Bernstein et al. "Generic Schema Matching, Ten Years Later", PVLDB 4(11), pp 695-701, 2011. Accessible at `http://www.vldb.org/pvldb/vol4/p695-bernstein_madhavan_rahm.pdf`

[2] J.Madhavan et al. "Web-scale Data Integration: You can only afford to Pay As You Go", Proc. CIDR 2007. Accessible at `http://www.cidrdb.org/cidr2007/papers/cidr07p40.pdf`

Domain-specific web meta-search



- deep web content is discovered by pre-computing the most relevant submissions for each "interesting" HTML form and adding the resulting URLs to the search engine index — this is called *surfacing* the deep web content;
- this allows existing search engine infrastructure to be leveraged;
- when a user clicks on a search result — on the basis of seeing the displayed snippet — the user is directed to the underlying web site and will see fresh content;
- the paper on "Google's Deep-Web Crawl"[3] identifies two main challenges that need to be addressed for effective surfacing, and proposes techniques to address them: (i) deciding which form inputs to fill in when submitting queries to a form, and (ii) deciding on appropriate values for these inputs;
- there is no single global schema encompassing all of this surfaced deep web content of course;
- instead it is expected that, gradually over time, multiple "topic-oriented" schemas will emerge, each playing the role of a "global schema" for a particular domain;
- the mappings between surfaced schemas and topic schemas will typically be approximate, not exact;
- queries can be posed to the topic schemas and automatically routed to the relevant web content;
- the entire integration approach is **"pay-as-you-go"**, in the sense that data sources will become increasingly, and more accurately, integrated over time. Such an approach is in contrast to the "single-shot" integration effort in conventional data integration approaches[4].

The PAYGO proposal for Web-scale data integration is an example of a **Dataspace Management System**[5]. In such systems, a *co-existence approach* to heterogeneous data sources is adopted, rather than full semantic integration:

- no large up-front investment is required for creating mappings in order to support data services over the data sources;
- instead, the data sources are iteratively mapped and integrated, as time and resources allow, to give higher-quality answers to queries;
- effort and resources can be focussed on supporting more effectively the most important queries.

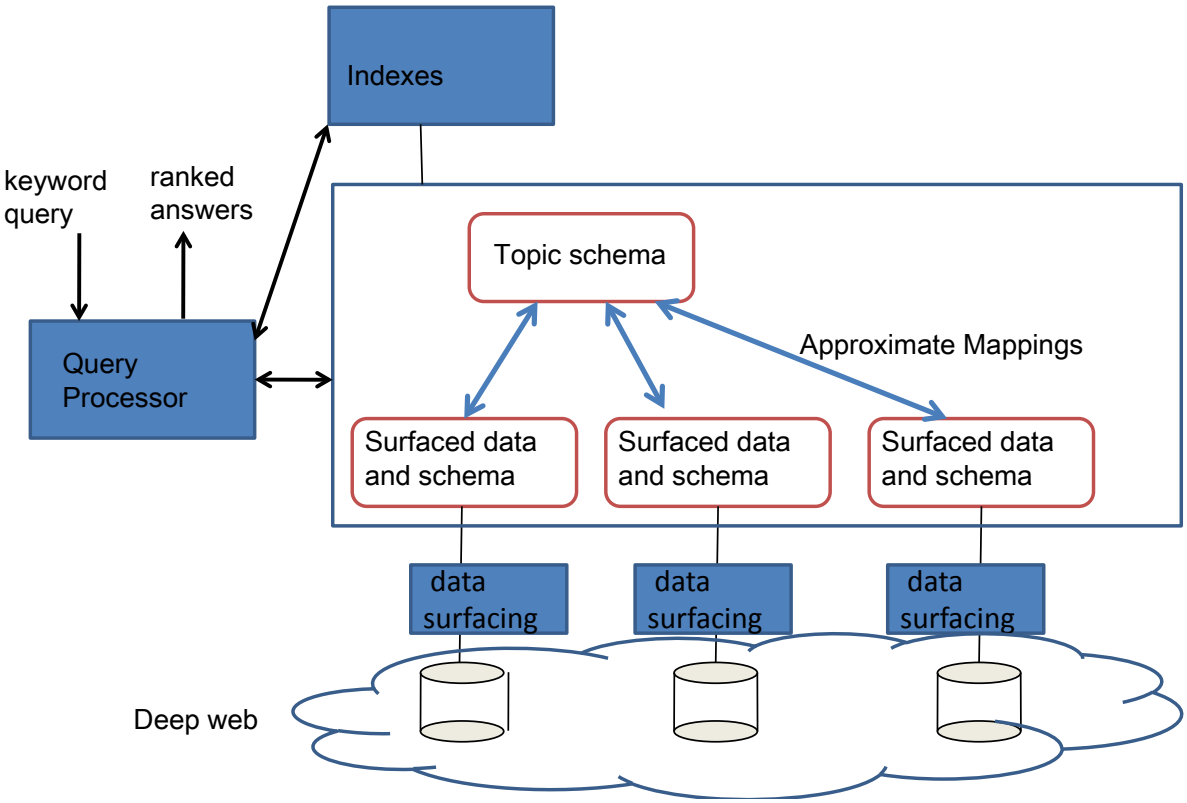
A more recent term roughly synonymous with "dataspace" is "data lake". "Data lake" is understood to mean the centralised storage of large volumes "raw" data, in a variety of heterogeneous formats as

[3] J.Madhavan et al., "Google's Deep-Web Crawl", PVLDB'08, August 2008, pp 1241-1252. Accessible at `http://dl.acm.org/citation.cfm?id=1454163`.

[4] For detailed discussions, and techniques, see A. Das Sarma et al. "Bootstrapping pay-as-you-go data integration systems", Proc. SIGMOD'08, pp 861-874; accessible at `http://dl.acm.org/citation.cfm?id=1376702` . H.A. Mahmoud and A. Aboulnaga "Schema clustering and retrieval for multi-domain pay-as-you-go data integration systems", Proc. SIGMOD'10, pp 411-422; accessible at `http://dl.acm.org/citation.cfm?id=1807213`

[5] A. Y. Halevy et al. "Principles of dataspace systems", Proc. PODS 2006. Accessible at `http://homes.cs.washington.edu/ alon/files/pods06.pdf`

Pay-As-You-Go


Indexes
keyword query
ranked answers
Query Processor
Topic schema
Approximate Mappings
Surfaced data and schema
Surfaced data and schema
Surfaced data and schema
data surfacing
data surfacing
data surfacing
Deep web

produced by the data sources, without applying any significant cleansing/transformation/integration effort to the data. This effort is applied later, as necessary, as part of subsequent data analysis processes.

Some of the issues discussed in this Section are explored in Reading 1.

2 Integrating Linked Data

Large volumes of information that was hidden up to now in the deep web is increasingly being published by its owners as sets of RDF triples, in the form of *Linked Data*[6].

Data that stored in relational databases can be published as Linked Data by using relational-to-RDF Wrapper software such as D2R Server (see http://d2rq.org/d2r-server) which allow mappings to be defined between source relational schemas and target RDF graphs.

Similar "RDFizer" tools exist for data that is stored in other structured, semi-structured or unstructured formats (see http://www.w3.org/wiki/ConverterToRdf).

Different RDF data sets may be published by different people, but referencing common URIs and hence providing links between these different data sets.

Federated SPARQL querying tools have also been developed that allow transparent querying of multiple linked data sources.

For example, Reading 4 gives a review of several of the major systems. It presents an example Use Case from Life Sciences data integration, and describes an approach to optimising federated SPARQL queries on the basis of available information about the contents and sizes of the data sources.

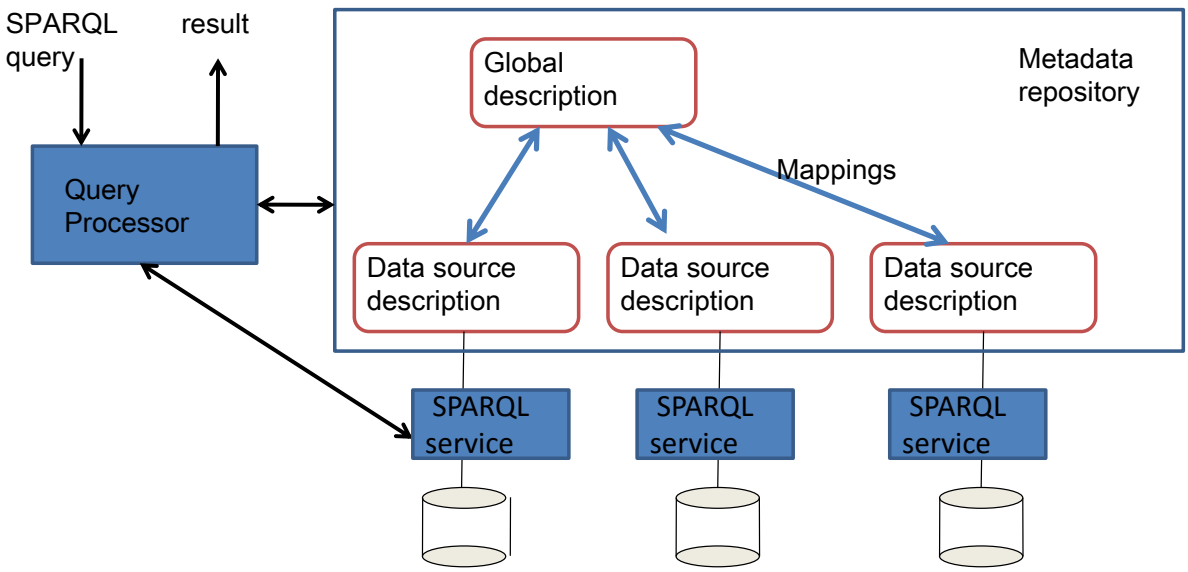
Linked Open Data is linked data that is licensed to be used for free.

This 'Web of Data' or 'Linked Open Data Cloud' is founded on pay-as-you-go principles with the aim of semantically integrating increasing numbers of data sets over time.

The ultimate vision is to be able to interact with Linked Open Data as if it were a single global database.

[6] See `http://linkeddata.org/` for datasets, tools, guides, tutorials etc.
Also `http://www.w3.org/standards/semanticweb/data` for related W3C standards and examples.

Integrating Linked Data



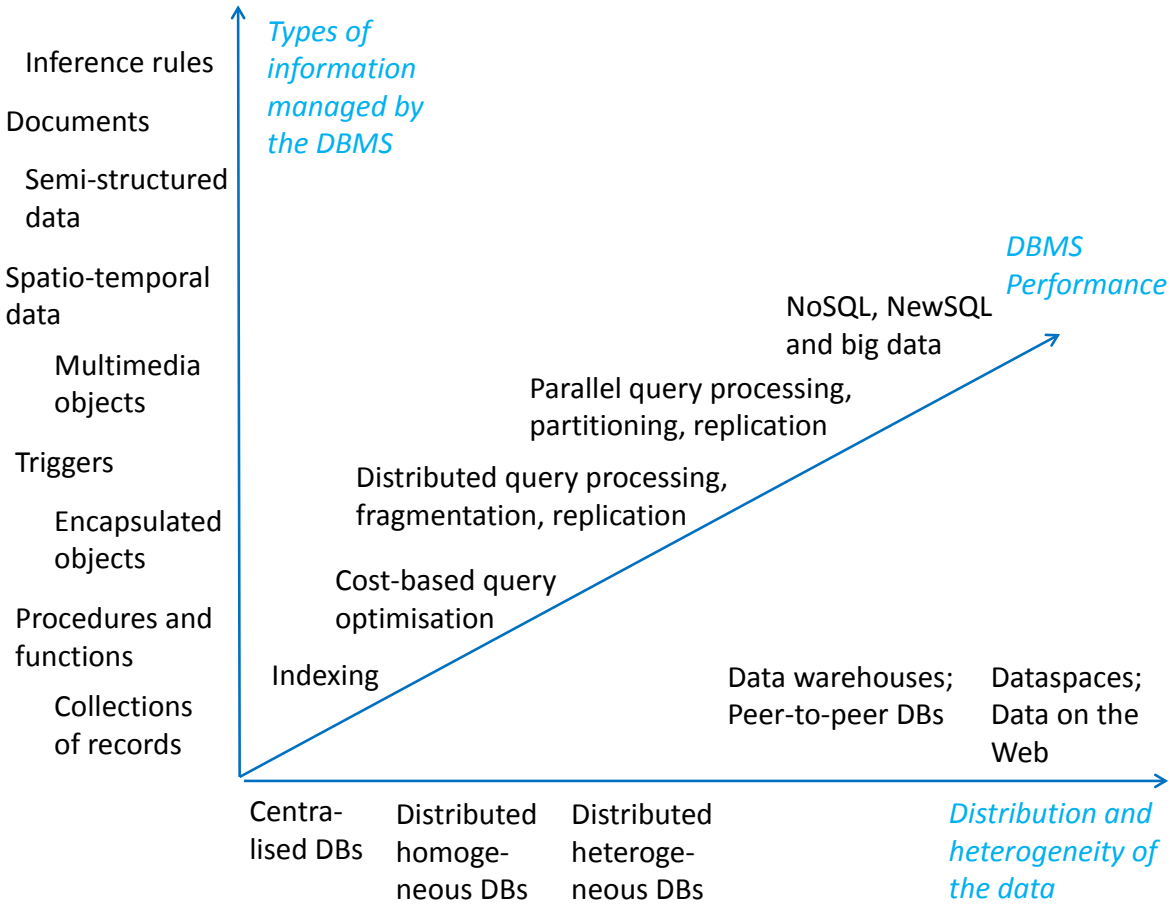


Figure 3: ADM Summary

3 ADM Round-Up

Recall from Week 1 that this course aimed to explore three major directions of developments and advances in DMBS technology:

- the degree of distribution and heterogeneity of the DBMS
- DBMS performance
- the variety of information managed by the DBMS

The figure summarises the topics we have covered along these three axes. Not shown, but cross-cutting all three axes, are the issues of consistency guarantees and support of ACID or BASE transactional properties.

4 Ongoing Research Topics (Optional Reading, for interest)

- light-weight integration of heterogeneous datasets in dataspace environments, leveraging AutoMed's fine-grained data integration capabilities;
  see "Intersection schemas as a dataspace integration technique", Richard Brownlow, Alex Poulovassilis, Proc. EDBT/ICDT Workshops 2014, pp 2-99, Moodle reading 5.

- querying of graph data using Regular Path Queries, with automatic query approximation & relaxation, and ranking of query answers;
  see `www.dcs.bbk.ac.uk/~ap/talks/nldb2013.pdf` and "Implementing Flexible Operators for Regular Path Queries", P.Selmer, A.Poulovassilis, P.T.Wood Proc. GraphQ 2015, at ICDT/EBDT, March 2015, Moodle reading 6.

- speeding up graph query processing by supporting *path indexes*;
  see "Efficient regular path query evaluation using path indexes", George H. L. Fletcher, Jeroen Peters, Alex Poulovassilis, Proc. EDBT 2016, pp 636-639, Moodle reading 7.

- adding flexible querying to SPARQL 1.1;
  see `www.dcs.bbk.ac.uk/~ap/talks/AUT2014.pdf` and "Flexible Querying for SPARQL", A.Cali, R.Frosini, A.Poulovassilis, P.T.Wood. Proc. OTM Conferences 2014, pp 473-490, Moodle reading 8.

- going beyond federated SPARQL query processing to support semantic P2P integration of Linked Data;
  see "Peer-to-Peer Semantic Integration of Linked Data", M.Dimartino, A.Cali, A.Poulovassilis, P.T.Wood. Proc. LWDM 2015, at ICDT/EDBT, pp 213-220, Moodle reading 9.

- a variety of interdisciplinary projects at the Birkbeck Knowledge Lab; see my home page and follow the links, e.g.:

  - "Mapping Museums" and "Weaving Communities of Practice": semantic integration, querying, visualisation of rich humanities data sets; using ontologies, SPARQL.
  - Learning analytics; see e.g. "Visualisation and analysis of students' interaction data in exploratory learning environments", M.Mavrikis et al. Proc. Web-based Education Technologies workshop, at WWW 2015, May 2015, Moodle reading 10.
  - Event-based services for Awareness in P2P Groupware Systems; implemented using peer-to-peer triggers; see `www.dcs.bbk.ac.uk/~ap/talks/3PGCIC2013Pres.pdf`
  - Flexible querying of lifelong learners' metadata;
    see `http://www.dcs.bbk.ac.uk/~ap/talks/nldb2013.pdf`, pages 6 - 49

Readings

1. Golshan, B., Halevy, A., Mihaila, G., & Tan, W. C. Data integration: After the teenage years. In Proceedings of the 36th ACM Symposium on Principles of Database Systems (pp. 101-106), 2017. Moodle Reading 1.

2. W. Kuhn et al., Linked Data - A Paradigm Shift for Geographic Information Science. Proc. International Conference on Geographic Information Science (pp. 173-186), 2014. Moodle Reading 2.

   This paper discusses the impacts and benefits of Linked Data and Semantic Web technologies, from the perspective of an application domain - that of Geographical Information Science.

3. (Optional) T.Naz, J.Dorn and A.Poulovassilis, A Hybrid Approach to Schema and Data Integration for Meta-search Engines. Technical Report BBKCS-09-02, Birkbeck, February 2009. Moodle Reading 3.

4. (Optional) O. Gorlitz and S. Staab, SPLENDID: SPARQL endpoint federation exploiting VOID descriptions. Proc. COLD at ISWC, 2011. Moodle Reading 4.