# Challenges for Biomedical Information Integration

C. Golbreich, A. Burgun
Laboratoire d'Informatique Médicale, Faculté de Médecine, 35033 Rennes, France
Christine.Golbreich@uhb.fr Anita.Burgun@univ-rennes1.fr

In the last decade significant progress have been done in Information Integration. Most systems for data integration issued from database and AI communities are mediator-based centralized systems. More recently, new approaches [4] [1] emerged, proposing distributed integration, that are quite attractive for Biological Information Integration (BII), such as functional genomics. Their deployment in BII depends on two main features. BII requires *flexible* integration and *expressive* representation languages.

## 1 Flexible information integration

Extensibility and real-time data are crucial requirements for BII. For example, Genomics is a very fast-moving field. Web sources are multiple, with huge and constantly evolving content (versioning of GO and UMLS). New online ontologies and specialized databanks frequently appear. Datawarehouses which can be quite powerful, providing high access performance are not well appropriate to such evolving data. More flexible integration, either centralized mediators or peer-based distributed integration might be more appropriate.

### 1.1 Mediator-based integration

A mediator includes a *global ontology* $G$ (or mediated schema) and a set $M$ of *mappings*, relating the global ontology $G$ to the sources ontologies $S$. The query engine exploits this knowledge to reformulate the user query into queries that refer to the sources ontologies $S$. In bioinformatics or in medicine, new sources constantly appear and shall be added to $S$. Therefore, mainly for their easier extensibility, *local as view* (LAV) mediators defining the content of sources in terms of views over the global ontology, might be more appropriate than *global as view* (GAV) defining the global ontology in terms of views over the sources e.g. Tambis [8]. However, they still raise representation problems (§ 2).

### 1.2 Peer-based integration

Mediators are a significant progress, but for scaling up the Web, centralized integration may be not flexible enough, and distributed systems perhaps even better appropriate. As illustrated for bioinformatics [6], databanks are not only data "sources" but also include precious links and mappings, through their cross-references to general ontologies and to other databanks. Such local relations between sources should be explicitly represented and directly exploited to infer new information. Peer-based integration where "every participant should be able to contribute new data and relate it to existing concepts and schemas, define new schemas that others can use as frames or reference for their queries or define new relationships between existing schema or data providers" [4] is challenging to address the extensibility and distribution encountered in BII.

## 2 Rich languages for ontologies and mappings

Whatever mediator or peer-based integration systems, rich formal languages are required for representing ontologies, queries, and mappings, in the biomedical domain.

### 2.1 A DL extended by rules for ontologies

As advocated in [2] a rich language, that is expressive enough to allow a fine and precise representation of both structural (concepts, properties, and hierarchies) and deductive knowledge, is required in the biomedical domain. The next W3C standard OWL(-DL) is a good candidate for taxonomies, but is not sufficient and should be extended by rules for the deductive part. Rules are particularly needed to represent dependencies between relations, such as mereotopological (part-of) and topological relationships, propagation of relations along transitive role, or consistency constraints [2] etc., for instance location of a disease is inherited across partonymy: "has-location propagates via part-of" [7]. However, as the combination of an expressive DL e.g. $\mathcal{ALNR}$ with rules e.g. Datalog enlarges the search space, a trade-off shall be found in limiting OWL or/and rules expressiveness, in order to remain decidable and to have sound and complete algorithms for subsumption and satisfiability. Second, using OWL as the ontology language in an integration system, fuels additional new questions, about (1) the query language: if rules are wanted to define conjunctive queries, the issue of a logical language combining OWL(-DL) with rules occurs again (2) the mappings language: how the mappings should be

represented; for example, by OWL subsumption or other axioms, by rules? (3) the query answering algorithm: decidability depends on the ontology, query, mapping languages. Thus, an integrated framework including OWL (or sublanguage) were queries reformulation is decidable is a key challenge for BII.

### 2.2 A metamodel and a logical language for the mappings

As illustrated in Bioinformatics (see [6]) the explicit representation of mappings play a key role in mediation. But there are several related problems to solve, in particular two main ones: the *modeling problem* "how to model the mappings between the sources and the global ontology (or between peers)"? and the *representation problem* "how to represent the mappings"?

A first challenge is to define a "*metamodel*" for *mappings*, at a conceptual level, independently of the representation language. For example, from the analysis of existing database or DL integration sytems, a first possible simple model[1] is to define, for a source s, *mappings* as triples ($D$, $P$, $C$), where $D$ is a set of assertions relating the kinds of *data* that can be found in the source s to the concepts of the global $G$, where $C$ is a set of *constraints* on its elements expressing restrictions on the data, or integrity constraints in terms of the global $G$, where $P$ is a set of assertions relating *local properties* of the source s *to* $G$ *properties*[2]. For example, for an integration system in genomics, where the global ontology $G$ includes the concepts **Protein**, **Species,** **HumanSpecies** and properties **organism**, mappings for the source SWISS-PROT (SW) are defined as a set of assertions stating 1) that SW entries correspond to instances of **Protein**, 2) to which $G$ entities, its lines are related, e.g. the OS line corresponds to the property **organism** and its content to instances of **Species**[3], 3) constraints e.g. the data of SW file "proteins of the non-redundant human proteome set" contains only human proteins. Thus, SWISS-PROT mappings are defined by the triple ($D_{sw}$, $P_{sw}$, $C_{sw}$), where $D_{sw}$ = {SW-data → **Protein**, ...}, $P_{sw}$: = {SW-OS → **organism**, ...}, $C_{sw}$:= {OS-data → **HumanSpecies**, ...}

A second challenge is to define a logical language for representing mappings and semantics of "→". Most mediators represent mappings as views over databases [3]. But several issues are now re-opened (1) which logical formalism to use, DL (OWL), rule, else? (2) if OWL, then how to represent them? In principle, subclass or subrole axioms e.g. $V^{SP}_{data} \subset$ Protein, $V^{SP}_{OS} \subset$ organism, $V^{SP}_{data} \subset (\forall$ organism HumanSpecies) are possible. Another option, is to represent them by rules e.g. $V^{SP}_{data}(X) \Rightarrow$ Protein(X), $V^{SP}_{OS}(X,Y) \Rightarrow$ organism(X,Y), and to to have a more complex model, for instance allowing to map a local property to a $G$ more complex expression. But the logical formalism to represent mappings with OWL ontologies is still an open issue. Indeed, as well studied [5] [3] the formalism has direct implications on the query reformulation problem, and as the formalism for expressing mappings becomes more expressive, it becomes harder. In conclusion, an hybrid formalism combining a subclass of OWL with rules, that allows to remain decidable and to have sound and complete algorithms for subsumption and satisfiability and if possible with good properties for the reformulation of queries using mappings is another key issue for BII.

Both mediator or peer-based integration raise a major question, that of available tools, ready to be used in BII.

## 3 References


[1] Bernstein P, Giunchiglia F, Kementsietsidis A, Mylopoulos J, Serafini L., Zaihrayeu I. Data management for peer-to-peer computing: A vision, Workshop WebDB 2002.

[2] Golbreich, C., Dameron, O., Gibaud, B., Burgun A. Web ontology language requirements w.r.t expressiveness of taxononomy and axioms in medecine, ISWC 2003, Springer, 2003.

[3] Halevy A. Y. Answering queries using views. The VLDB Journal, 10(4):270-294, 2001.

[4] Halevy, A. Y. Zachary G. Ives, Dan Suciu, and Igor Tatarinov. Schema mediation in peer data management systems. In ICDE, 2003.

[5] Levy A. Y, Rousset MC, The Limits on Combining Recursive Horn Rules with Description Logics, AAAI/IAAI, Vol. 1 (1996)

[6] Marquet G, Golbreich C., Burgun A From an ontology-based search engine towards a mediator for medical and biological information integration, Semantic Integration Workshop, ISWC 2003, Sanibel, Florida, 2003.

[7] Rector A. Analysis of propagation along transitive roles: Formalisation of the GALEN experience with Medical Ontologies, 2002 Int. Workshop on Description Logics DL2002, April 19-21, (2002)

[8] Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, Brass A. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics. 2000 Feb;16(2):184-5.


[1] presented on a LAV mediator, but it can be generalized to other approaches, including Peer-based integration

[2] "concept" and "property" refer to Class and Property in OWL.

[3] organism(s) which was (were) the source of the stored sequence