Ontology Guided Data Integration for Computational Prioritization of Disease Genes

Bert Coessens[1], Stijn Christiaens[2], Ruben Verlinden[2], Yves Moreau[1], Robert Meersman[2], Bart De Moor[1]

[1] Department of Electrical Engineering
Katholieke Universiteit Leuven
bert.coessens,yves.moreau,bart.demoor@esat.kuleuven.be
[2] Semantics Technology and Applications Research Laboratory
Vrije Universiteit Brussel
stijn.christiaens,ruben.verlinden,robert.meersman@vub.ac.be

Abstract. In this paper we present our progress on a framework for collection and presentation of biomedical information through ontology-based mediation. The framework is built on top of a methodology for computational prioritization of candidate disease genes, called Endeavour. Endeavour prioritizes genes based on their similarity with a set of training genes while using a wide variety of information sources. However, collecting information from different sources is a difficult process and can lead to non-flexible solutions. In this paper we describe an ontology-based mediation framework for efficient retrieval, integration, and visualization of the information sources Endeavour uses. The described framework allows to (1) integrate the information sources on a conceptual level, (2) provide transparency to the user, (3) eliminate ambiguity and (4) increase efficiency in information display.

1 Introduction

The ever increasing amount of biological data and knowledge, its heterogeneous nature, and its dissemination all over the Internet, make efficient data retrieval a horrendous task. Biological research has to deal with the diversity and distribution of the information it works with [1]. Yet, access to and integration of a multitude of complementary data sources will become critical to achieve more global views in biology.

Current solutions to these problems are usually handled manually and integration becomes a tedious activity of ad-hoc implementation [2]. Part of these problems originate from the fact that most of these data sources are created as if they exist alone in the world [3]. It is clear that a good framework is a necessity, in which integration can be done through configuration, rather than case-specific implementation. Such a framework will need to be based on semantics, rather than on syntax and structure [4].

Integration on a conceptual level also opens up new opportunities for creating information-rich user interfaces. Presenting a user with the information she needs, augmented with relations to relevant data, is an approach used commonly in semantic web browsers (e.g., OntoWeb [5]) and even in industry software (e.g., Context Browser [6]). Dzbor, Domingue, and Motta [7] call this an interpretative viewpoint or also context. This kind of interface allows a disambiguating stepwise refinement of the user's search [8].

In contrast to warehouse oriented tight integration systems (e.g., Biozon [9]), our approach is to provide efficient navigation through sets of loosely integrated data sources. The purpose of our system is not to allow generic queries on overlapping data sources, but rather to give transparent access to information that is highly relevant and useful for the analysis at hand. Thus, the system fills a very specific need in relieving the researcher of the burden to manually collect or navigate to all necessary information.

According to the classification proposed by Hernandez and Kambhampati [1], our system falls in the category of portal systems (like SRS [10]), but has a structured object-relational data model (like TAMBIS [11], K2/BioKleisli [12], DiscoveryLink [13], etc.). The system is limited to horizontal integration of data sources, but being navigational it does not require critical expertise of a specific query language (like in systems using CPL [14, 11] or OQL [15, 12], for instance).

We start with a description of Endeavour and the problem of computational gene prioritization. In section 3 we give a brief overview of the DOGMA framework for ontology engineering, followed by an explanation how an ontology can be used for mediation and visualization of the information used in the prioritization methodology in section 4. We show how this solves several integration problems. We end this paper with some conclusions and possible future directions in section 5.

2 Endeavour

In the field of linkage analysis and association studies researchers are often confronted with large lists of candidate disease genes, especially when investigating complex multigenic diseases. Investigating all possible candidate genes is a tedious and expensive task that can be alleviated by selecting for analysis only the most salient genes. Also, in the context of high-throughput experiments (like microarray gene expression assays), ever growing amounts of gene-associated data make manual investigation of *interesting* genes nearly unfeasible. It is clear that efficient and statistically sound computational prioritization methods become increasingly important.

With respect to this need, we developed the Endeavour methodology for the computational prioritization of a group of candidate genes based on their similarity with a set of training genes [16]. The method uses Order Statistics to combine a variety of information sources and clearly has several advantages over other approaches. It solves the problem of missing data and reconciles even contradictory information sources. It allows for a statistical significance level to be set after multiple testing correction, thus removing any bias otherwise introduced by the expert during manual prioritization. It also removes part of the bias towards known genes by including data sources that are equally valid for known and unknown genes. The methodology has been validated in a large scale leave-one-out experiment with 29 diseases and 627 disease genes fetched from OMIM. On top of that, several prioritizations were validated in wet-lab experiments. These analyses were published in Nature Biotechnology by Aerts *et al.* [16].

Endeavour prioritization terminology The central object in the Endeavour prioritization methodology is a *Gene*. This object represents a biological entity and all information known about it. In most of the cases, this entity will be a gene. Biological entities are combined in sets (*GeneGroup*). A training set is a *GeneGroup* that is used to build a model for a process or disease, represented by the *Model* object. A *Model* consists of several *SubModel* objects that each represent a certain data source. Building a *SubModel* means fetching and summarizing all information about the genes in the training set for one particular data source.

Endeavour comes with a set of standard submodels that summarize the following information about the user-specified training genes: KEGG pathway membership [17], Gene Ontology (GO) annotations [18], textual descriptions from MEDLINE abstracts, microarray gene expression, EST-based anatomical expression, InterPro's protein domain annotation [19], BIND protein interaction data [20], *cis*-regulatory elements, and BLAST sequence similarity. Besides these default information models, users can add their own microarray data or custom prioritizations as submodels. Most of the data sources are either vector-based (e.g., textual information, gene expression data) or attribute-based (GO, EST, InterPro, Kegg).

Apart from the *GeneGroup* that contains the training genes, there is a second *GeneGroup* that holds the candidate genes and all their related information. These candidate genes are prioritized during a process called *scoring*. Scoring involves comparing the information of a candidate gene with the information in the *Model* object for every data source. Based on these comparisons, every candidate gene receives a ranking. All rankings of the test genes according to the different available data sources are then combined using order statistics to obtain one overall ranking.

Endeavour information browser The decision was taken to provide the users of Endeavour with a maximal control over the set of training and test genes, as well as over the data sources to include in the prioritization. This idea was conceived from a prospective discussion with many geneticists and biologists, who do not use the existing prioritization methods for their lack of flexibility. This is perhaps best illustrated by the fact that not a single paper has been published reporting the identification of a novel disease gene when using any of the pre-existing methods. Most likely, this relates to the reality that geneticists and biologists, as opposed to bioinformaticians, prefer to have the flexibility to interactively select their own set of genes and the information they want to work with, above an automatic and non-interactive data mining selection procedure of disease characteristics.

In this context, it is of utmost importance to make well-informed decisions about which genes and information sources to include in the prioritization. A user must be able to browse the relevant information efficiently and in accordance with the methodology's demands. To live up to this need, and given the heterogeneous nature of the biological information to be consulted, an information browser was developed based on the Endeavour methodology. The existing data model was extended to a full-fledged ontology with *Gene* as the central object to allow ontology-guided browsing through the available information (see Figure 2).

3 DOGMA Ontology Paradigm

DOGMA[3] is a research initiative of VUB STARLab where various theories, methods, and tools for building and using ontologies are studied and developed. A DOGMA inspired ontology is based on the classical model-theoretic perspective [21] and decomposes an ontology into a lexon base and a layer of ontological commitments [22, 23]. This is called the principle of double articulation [24].

A lexon base holds (multiple) intuitive conceptualization(s) of a particular domain. Each conceptualization is simplified to a *representation-less* set of context-specific binary fact types called lexons. A lexon represents a plausible binary fact-type and is formally described as a 5-tuple <V, term1, role, co-role, term2>, where V is an abstract context identifier, lexically described by a string in some natural language, and is used to group lexons that are logically related to each other in the conceptualization of the domain. Intuitively, a lexon may be read as: within the context V, the term1 (also denoted as the header term) may have a relation with term2 (also denoted as the tail term) in which it plays a role, and conversely, in which term2 plays a corresponding co-role. Each (context, term)-pair then lexically identifies a unique concept. A lexon base can hence be described as a set of plausible elementary fact types that are considered as being true. Any specific (application-dependent) interpretation is moved to a separate layer, i.e., the commitment layer.

The commitment layer mediates between the lexon base and its applications. Each such ontological commitment defines a partial semantic account of an intended conceptualization [25]. It consists of a finite set of axioms that specify which lexons of the lexon base are interpreted and how they are visible in the committing application, and (domain) rules that semantically constrain this interpretation. Experience shows that it is much harder to reach an agreement on domain rules than one on conceptualization [28]. For instance, the rule stating that each gene is identified by an Ensembl Gene ID [26] may hold in the Universe of Discourse (UoD) of some application, but may be too strong in the UoD of another application (e.g., Entrez Gene [27]). A full formalization of DOGMA can be found in De Leenheer, Meersman, and de Moor [29, 30].

[3] Developing Ontology-Grounded Methods for Applications

4 Ontology Guided Mediation

As we are trying to integrate several heterogeneous data sources used in the Endeavour prioritization process and with conceptually overlapping instances, the approach described by Verheyden and Deray [31–33] is suited for our purposes. This approach uses ontologies as a mediating instrument for conceptual data integration.

Each data source is individually mapped (or committed) to the ontology using the Ω-RIDL commitment language [32]. These individual mappings ($\Omega_1 \ldots \Omega_n$ in Figure 1) enable the mediator to access and query the respective wrappers according to its own view on the data. At this point, the data sources were mapped to the ontology manually.

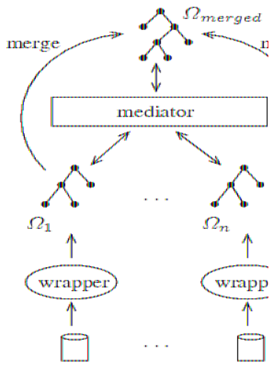


Fig. 1. Mediator Approach for Data Integration.

As a proof of concept, we used this approach to integrate different, but partially overlapping, data sources from Endeavour. In close cooperation with a domain expert, we modeled the ontology based on two major types of data sources in more detail (vector- and attribute-based sources). Since the ontology is aligned directly to the Endeavour terminology, all concepts are unambiguously defined.

The obtained ontology is displayed in Figure 2 in the form of a NORM-tree [34]. Terms (e.g., gene) can be observed multiple times (in darker grey) in this representation. Each node in the tree is extended (using a double click-action) with all its related terms in order to display a complete local context at all times. When a node is selected, it forms a path (darker line in Figure 2) to the root of the NORM-tree. This local context approach is one of the abstraction mechanisms identified by Halpin [35].

Figure 3 shows how the link between the different heterogeneous data sources to the conceptual model can be used to provide transparency, eliminate ambiguity, and increase efficiency when displaying relevant information to the user.

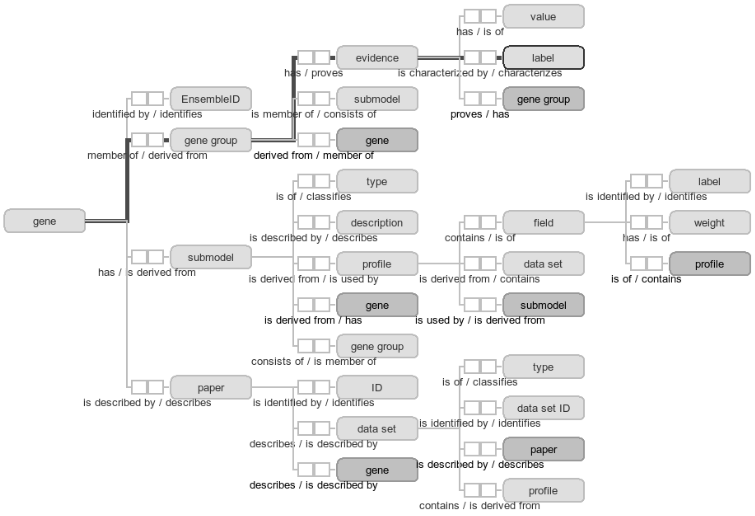

value
has / is of
evidence
label
has / proves
is characterized by / characterizes
EnsembleID
submodel
gene group
identified by / identifies
is member of / consists of
proves / has
gene group
gene
member of / derived from
derived from / member of
type
label
is of / classifies
is identified by / identifies
gene
description
field
weight
is described by / describes
contains / is of
has / is of
submodel
profile
data set
profile
has / is derived from
is derived from / is used by
is derived from / contains
is of / contains
gene
submodel
is derived from / has
is used by / is derived from
gene group
consists of / is member of
type
paper
ID
is of / classifies
is described by / describes
is identified by / identifies
data set ID
data set
is identified by / identifies
describes / is described by
paper
gene
is described by / describes
describes / is described by
profile
contains / is derived from

Fig. 2. Partial Endeavour Ontology visualized using the DOGMA NORM-tree representation.

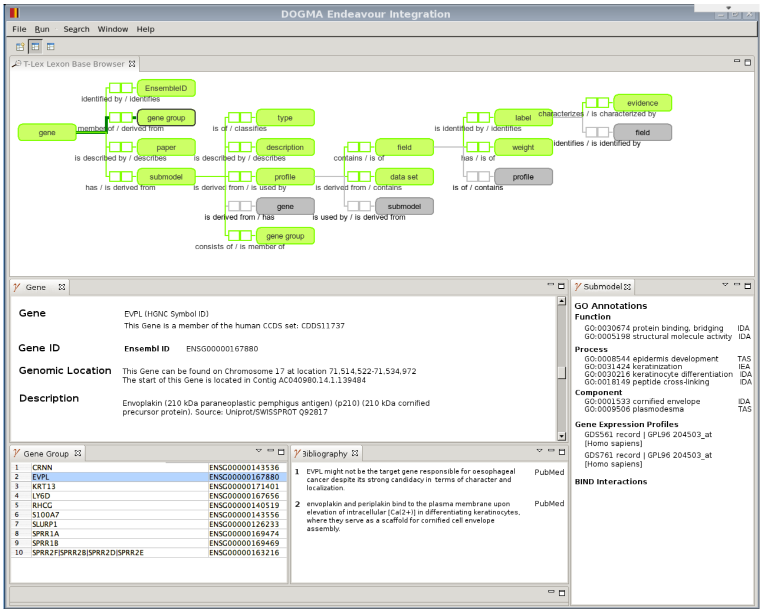
Using the view in the upper part of the screenshot the user can browse the NORM-tree as described for Figure 2. The gene-related information is shown to the user in the lower half of the screen. While the user browses the NORM-tree, she selects relevant ontological paths. This way the active instance of query results provide the information she requests.

As a result of using this integrated interface the user is not confronted with the terminology and jargon used in the different data sources. The user only sees well-known terminology in the ontology when selecting relevant objects, thus eliminating ambiguity. She is also not confronted with the specific source of the data unless explicitly desired, thus providing transparency. By selecting ontological paths in the NORM-tree only information relevant to the local context is shown, thus enhancing efficiency. Eliminating ambiguity, providing transparency and enhancing efficiency to the user when browsing information relevant to a group of genes will allow her to concentrate completely on the prioritization analysis.

5 Discussion and future work

By using a conceptual approach for data integration we obtain several significant advantages. Although tedious, the mediating process is relatively easy to apply on new vector- and attribute-based data sources. The only difficulty arises if the data source (of a yet unmet type) contains data that cannot be mapped to the

[3] The screenshot in Figure 3 is a partial mock-up. The actual link between the gene related information and the interface still needs to be implemented.



Gene EVPL (HGNC Symbol ID)
This Gene is a member of the human CCDS set: CDDS11737

Gene ID **Ensembl ID** ENSG00000167880

Genomic Location This Gene can be found on Chromosome 17 at location 71,514,522-71,534,972
The start of this Gene is located in Contig AC040980.14.1.139484

Description Envoplakin (210 kDa paraneoplastic pemphigus antigen) (p210) (210 kDa cornified precursor protein). Source: Uniprot/SWISSPROT Q92817

1	CRNN	ENSG00000143536
2	EVPL	ENSG00000167880
3	KRT13	ENSG00000171401
4	LY6D	ENSG00000167656
5	RHCG	ENSG00000140519
6	S100A7	ENSG00000143556
7	SLURP1	ENSG00000126233
8	SPRR1A	ENSG00000169474
9	SPRR1B	ENSG00000169469
10	SPRR2F|SPRR2B|SPRR2D|SPRR2E	ENSG00000163216

1 EVPL might not be the target gene responsible for oesophageal cancer despite its strong candidacy in terms of character and localization. PubMed

2 envoplakin and periplakin bind to the plasma membrane upon elevation of intracellular [Ca(2+)] in differentiating keratinocytes, where they serve as a scaffold for cornified cell envelope assembly. PubMed

GO Annotations
Function
GO:0030674 protein binding, bridging IDA
GO:0005198 structural molecule activity IDA
Process
GO:0008544 epidermis development TAS
GO:0031424 keratinization IEA
GO:0030216 keratinocyte differentiation IDA
GO:0018149 peptide cross-linking IDA
Component
GO:0001533 cornified envelope IDA
GO:0009506 plasmodesma TAS

Gene Expression Profiles
GDS561 record | GPL96 204503_at [Homo sapiens]
GDS761 record | GPL96 204503_at [Homo sapiens]

BIND Interactions

Fig. 3. Screenshot of the Endeavour Information Browser. In the top pane, the ontological paths are selected using the DOGMA NORM-tree representation. Relevant data, according to the selected paths, is retrieved on the fly from the different data sources and displayed in the lower panes. The concepts in the ontology align directly to the Endeavour terminology, which results in more efficient browsing through all information related to the analysis.

ontology. In this case, the ontology must be extended in order to integrate the new data. Semi-automated mapping of new data sources lies beyond the scope of the current research.

Another benefit of our approach is found in the visualization of the data. Since all data is linked (mapped) to a certain concept in the ontology, it is possible to enrich the view (e.g., the result of a query) with relevant and meaningful related information. The data is not only presented, it is also displayed in its own local context. This results in a complete transparency to the user and a more efficient visualization, as what needs to be seen, is shown.

The ontology we use can also be used by other applications, who use Endeavour itself as the data source. The conceptual model solves interoperability problems, as from the model alone, it is clear what Endeavour can provide, and how it can provide this data.

The tool is at this point a supporting application for the Endeavour system. We will extend it in order to have it actually send data to Endeavour to further facilitate gene prioritization.

[*] Research supported by: [Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, IDO (Genetic networks), several PhD/postdoc & fellow grants]; [Flem-

References

1. Thomas Hernandez ans Subbarao Kambhampati : Integration of biological sources : current systems and challenges ahead. In ACM Sigmod Record, 2004, 33(3), pp. 51-60
2. L. Stein : Creating a bioinformatics nation. In Nature 6885, 2002, 417, pp. 119-120
3. M. Stonebraker : Integrating Islands of Information , EAI Journal, Sept 1999. http://www.eaijournal.com/DataIntegration/IntegrateIsland.asp.
4. Sheth A. 1998 : Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics, in Interoperating Geographic Information Systems, M.F. Goodchild, M.J. Egenhofer, R. Fegeas, and C.A. Kottman (eds) Kluwer Publishers
5. P. Spyns, D. Oberle. R. Volz, J. Zheng, M. Jarrar, Y. Sure, R. Studer, R. Meersman : OntoWeb - a Semantic Web Community Portal. In Proc. Fourth International Conference on Practical Aspects of Knowledge Management (PAKM), December 2002, Vienna, Austria, 2002.
6. Context Browser : http://www.agilense.com/p_artifactmgt.html
7. M. Dzbor, J. Domingue and E. Motta : Magpie - Towards a Semantic Web Browser. In International Semantic Web Conference 2003, LNCS 2870, pp. 690-705
8. E. Garca and M.A. Sicilia : User Interface Tactics in Ontology-Based Information Seeking. In Psychology e-journal 2003 1(3):243-256.
9. Birkland, A and Yona, G : BIOZON: a system for unification, management and analysis of heterogeneous biological data. In BMC Bioinformatics, 2006, 7, pp. 70-70
10. T. Etzold and A. Ulyanov and P. Argos : SRS: information retrieval system for molecular biology data banks. In Methods Enzymol, 1996, 266, pp. 114-128
11. R. Stevens and P. Baker and S. Bechhofer and G. Ng and A. Jacoby and N.W. Paton and C.A. Goble and A. Brass : TAMBIS: transparent access to multiple bioinformatics information sources. In Bioinformatics, 2000, 16, pp. 184-185
12. S.B. Davidson and J. Crabtree and B.P. Brunk and J. Schug and V. Tannen and G.C. Overton and C.J. Stoeckert : K2/Kleisli and GUS: Experiments in integrated access to genomic data sources. In IBM Systems Journal, 2001, 40, pp. 512-531
13. L.M. Haas and P.M. Schwarz and P. Kodali and E. Kotlar and J.E. Rice and W.C. Swope : DiscoveryLink: A system for integrated access to life sciences data sources. In IBM Systems Journal, 2001, 40, pp. 489-511
14. L. Wong : The Collection Programming Language - Reference Manual. Technical Report Kent Ridge Digital Labs, 21 Heng Mui Keng Terrace, Singapore 119613

ish Government: FWO: PhD/postdoc grants, projects G.0407.02 (support vector machines), G.0413.03 (inference in bioi), G.0388.03 (microarrays for clinical use), G.0229.03 (Bonatema), G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel) research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-SQUAD (quorum sensing), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos]; [Belgian Federal Science Policy Office: IUAP P5/22 ('Dynamical Systems and Control: Computation, Identification and Modeling, 2002-2006)]; [EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain]

15. A. M. Alashqur and Stanley Y. W. Su and Herman Lam : OQL: A Query Language for Manipulating Object-oriented Databases In Proceedings of the Fifteenth International Conference on Very Large Data Bases, August 22-25, 1989, Amsterdam, The Netherlands, pp. 433-442
16. S. Aerts and D. Lambrechts and S. Maity and P. Van Loo and B. Coessens and F. De Smet and L.C. Tranchevent and B. De Moor and P. Marynen and B. Hassan and P. Carmeliet and Y. Moreau : Gene prioritization via genomic data fusion. In Nat Biotechnol, 2006, 24, pp. 537-544
17. M. Kanehisa and S. Goto and S. Kawashima and Y. Okuno and M. Hattori : The KEGG resources for deciphering the genome. In Nucleic Acids Res., 2004, 32, pp. D277-D280
18. The Gene Ontology Consortium : Gene Ontology: tool for the unification of biology. In Nature Genet, 2000, 25, pp. 25-29
19. N.J. Mulder and R. Apweiler and T.K. Attwood and A. Bairoch and A. Bateman and D. Binns and P. Bradley and P. Bork and P. Bucher and L. Cerutti and R. Copley and E. Courcelle and U. Das and R. Durbin and W. Fleischmann and J. Gough and D. Haft and N. Harte and N. Hulo and D. Kahn and A. Kanapin and M. Krestyaninova and D. Lonsdale and R. Lopez and I. Letunic and M. Madera and J. Maslen and J. McDowall and A. Mitchell and A.N. Nikolskaya and S. Orchard and M. Pagni and C.P. Ponting and E. Quevillon and J. Selengut and C.J. Sigrist and V. Silventoinen and D.J. Studholme and R. Vaughan and C.H. Wu : InterPro, progress and status in 2005. In Nucleic Acids Res, 2005, 33, pp. D201-205
20. D. Gilbert : Biomolecular interaction network database. In Brief Bioinform, 2005, 6, pp. 194-198
21. Reiter, R. : Towards a Logical Reconstruction of Relational Database Theory. In Brodie, M., Mylopoulos, J., Schmidt, J. (eds.), On Conceptual Modelling, Springer-Verlag, 1984, pp. 191-233.
22. Meersman, R. : The Use of Lexicons and Other Computer-Linguistic Tools in Semantics, Design and Cooperation of Database Systems. In Zhang, Y., Rusinkiewicz, M., Kambayashi, Y. (eds.), Proceedings of the Conference on Cooperative Database Systems (CODAS 99), Springer-Verlag, 1999, pp. 1-14.
23. Meersman, R. : Ontologies and Databases: More than a Fleeting Resemblance. In d'Atri, A., Missikoff, M. (eds.), OES/SEO 2001 Rome Workshop, Luiss Publications.
24. Spyns, P., Meersman, R. and Jarrar, M. : Data Modelling versus Ontology Engineering. SIGMOD Record: Special Issue on Semantic Web and Data Management, 2002, 31(4), pp. 12-17.
25. Guarino, N., and Giaretta, P. : Ontologies and Knowledge Bases: Towards a Terminological Clarification. In Mars, N. (ed.) Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing, IOS Press, Amsterdam, pp. 25-32.
26. A. Kasprzyk and D. Keefe and D. Smedley and D. London and W. Spooner and C. Melsopp and M. Hammond and P. Rocca-Serra and T. Cox and E. Birney : EnsMart: a generic system for fast and flexible access to biological data. In Genome Res, 2004, 14, pp. 160-169
27. D. Maglott and J. Ostell and K.D. Pruitt and T. Tatusova : Entrez Gene: gene-centered information at NCBI. In Nucleic Acids Res, 2005, 33, pp. 54-58
28. Meersman, R. : Semantic Web and Ontologies: Playtime or Business at the Last Frontier in Computing? In NSF-EU Workshop on Database and Information Systems Research for Semantic Web and Enterprises, 2002, pp. 61-67.
29. P. De Leenheer and R. Meersman : Towards a formal foundation of DOGMA ontology: part I. Technical Report STAR-2005-06, VUB STARLab, 2005.

30. De Leenheer, P. and de Moor, A. and Meersman, R. : Context Dependency Management in Ontology Engineering. Technical Report STARLab, Brussel, 2006.
31. Deray T. and Verheyden P. : Towards a semantic integration of medical relational databases by using ontologies: a case study. In, R. Meersman, Z. Tari et al.,(eds.), On the Move to Meaningful Internet Systems 2003: OTM 2003 Workshops, LNCS 2889, pp. 137 - 150, 2003. Springer Verlag.
32. Verheyden P. Deray T. and Meersman R., Semantic Mapping of Large and Complex Databases to Ontologies: Methods and Tools. Technical Report 25, STAR Lab, Brussel, 2004.
33. Verheyden P., De Bo J. and Meersman R. : Semantically unlocking database content through ontology-based mediation . In, Bussler C., Tannen V. and Fundulaki I.,(eds.), Proceedings of the 2nd Workshop on Semantic Web and Databases (SWDB 2004), LNCS 3372, pp. 109 - 126, 2005. Springer Verlag.
34. Trog D. and Vereecken J. : Context-driven Visualization for Ontology Engineering. Master thesis, Vrije Universiteit Brussel, 2006
35. Halpin T. : Information Modeling and Relational Databases. Morgan Kaufmann Publishers Inc. , 2001
36. Barriot, R and Poix, J and Groppi, A and Barré, A and Goffard, N and Sherman, D and Dutour, I and de Daruvar, A : New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. In Nucleic Acids Res, 2004, 32(12), pp. 3581-3589