Next Generation Data Integration for the Life Sciences

Proposal for a three hours tutorial at ICDE 2010, Hannover

Sarah Cohen-Boulakia
Université Paris-Sud 11
cohen@lri.fr

Ulf Leser*
Humboldt-Universität zu Berlin
leser@informatik.hu-berlin.de

Abstract— **Ever since the advent of high-throughput biology (e.g., the Human Genome Project), integrating the large number of diverse biological data sets has been considered as one of the most important tasks for advancement in the biological sciences. Whereas the early days of research in this area were dominated by virtual integration systems (such as multi-/federated databases), the current predominantly used architecture uses materialization. Systems are built using ad-hoc techniques and a large amount of scripting. However, recent years have seen a shift in the understanding of what a "data integration system" actually should do, revitalizing research in this direction. In this tutorial, we review the past and current state of data integration for the Life Sciences and discuss recent trends in detail, which all pose challenges for the database community.**

Keywords: bioinformatics, data integration, semantic web, scientific workflow, ranking, heterogeneous databases

I. Introduction

Biological research is a science which derives its findings from the proper analysis of experiments. But what has changed dramatically over the last three decades is the throughput of those experiments – from single observations to gigabytes of sequences in a single day – and the breadth of questions that are studied – from single molecules to entire genomes, transcriptomes, proteomes, etc. Today, a large variety of experiments are carried-out in hundreds of labs around the world, and their results are reported in a myriad of different databases, web-sites, publications etc., using different formats, conventions, and schemas. The integration of these diverse and distributed databases has been a topic of bioinformatics research for more than 20 years. The goals of data integration (DI) from a biologists point-of-view are mostly cost reduction (less redundant work), quality enhancement (exploiting redundant work), new findings (combining complementary work), and faster discoveries (reusing instead of redoing). From the very beginning, also database researchers have worked on this topic [1], attracted by the importance and challenge of the problem and the multitude of truly heterogeneous and freely available data sets. However, after about a decade of research, efforts declined in the early 2000's. A number of technologically advanced systems had been developed (federated, mediator-based, multi-database languages), but almost none of them achieved a notable impact in the targeted domain. In turn, projects that were well received in the field had been developed by pure practitioners, mostly using home-grown data structures – and large quantities of Perl programming.

However, the need for DI in the Life Sciences has not ceased, but is ever increasing [2, 3]. Systems Biology, aiming at a comprehensive view on cell physiology, inherently depends on data from a multitude of different sources. Translational Medicine targets the transfer of results from basic biological research into medical practice, calling for the integration of genomic and medical data. Related disciplines such as ecology, paleontology, or biodiversity, increasingly need to include data on a level of detail that is only achieved by genomic research, often integrating bio-molecular data with geographic information. This trend is also reflected in the establishment of a proper workshop/conference series (DILS – Data Integration for the Life Sciences), large project calls on national and international level (eScience, cyberinfrastructure etc.), and the establishment of specialized working groups at international organizations (such as W3C-HCLS)

Recent years have seen a revitalization of DI research in the Life Sciences. But the perception of the problem has changed: While early approaches concentrated on handling schema-dependent queries over heterogeneous and distributed databases, current research emphasizes instances rather than schemas, tries to place the human back into the loop, and intertwines data integration and data analysis. Transparency, one of the main goals of federated databases, is not a target anymore; instead, users want to know exactly which data from which source was used in which way in studies (provenance). The old model of "first integrate, then analyze" is replaced by a new, process-oriented paradigm: "integration is analysis – and analysis is integration". These new views on DI, lessons learnt from the past, and the challenges to face are the subject of this tutorial.

II. Biological Data Sets

Designing DI solutions in the context of the Life Sciences must take into account the specific properties of the domain [4, 5]. These are related to (i) the way biological data is produced and stored within biological databases (BDB), (ii) the fact that biology is a science in constant evolution, and (iii) the fact that BDB users are from various communities.

Regarding the first point, *primary* BDB store data that is close to the raw experimental result (e.g., DNA sequences), while *secondary* BDB manage information obtained after various steps of analysis and careful curation, grouped around increasingly complex entities (e.g., genes, diseases, pathways etc.). While (discrete, string-type) sequences were the main kind of raw data to integrate 20 years ago, today various –omics data set (transcriptomics, proteomics,etc.), often consisting of quantitative measurements, are of equal importance. Secondary BDBs heavily import data from other BDB and are mostly maintained by human curators [6]; they may be com-

* Work performed while on leave at LRI, Université Paris-Sud 11, France.

plementary but also could be redundant and divergent (especially when experts disagree). Because humans offer unrivaled precision, manual curation still predominates, despite its high cost and obvious problems of scalability [7].

The number of BDB is increasing steeply; currently more than 1,200 BDBs are publicly available [8]. This led to the call for DI systems with extreme scalability in the number of sources. However, in reality the usage of these databases is Zipf-distributed [2], and there are only a few projects that work with more than, say, 20 different data sources. The choice of BDB to be integrated is usually is different from project to project, driven by specific user requirements, and very often encompasses integration of local (private) data with publicly available data sets.. Additionally, DI is hardest when applied to the large and constantly changing set of secondary databases; however, these are also the most rewarding targets as they provide condensed knowledge instead of raw data. Another feature of BDBs is that they heavily cross-reference each other, mostly in the form of storing external IDs. These instance-level links probably are the most important source of information for DI and form a dense network of hundreds of databases [4], although they also create severe problems (e.g., no semantics, broken and outdated links, a high level of incompleteness, etc.).

Biology is a science in a state of constant evolution. This leads to a general high level of fuzziness and volatility in concept definitions. Even a fundamental concept such as "gene" has plenty of definitions that change over time, due to scientific discoveries of new phenomena [9]. This has obvious consequences in schema mapping. The second consequence is the inherent difficulty to identify objects. A given gene may be discovered (and named) independently by various groups and it is often hard to establish that different pieces of information are actually related to the same object.

The third class of BDB properties is related to social issues. Biologists need to know the exact origin of any information used in an analysis, as only this gives them an understanding of what they can expect from the data (in terms of quality, completeness etc.). Therefore, provenance and the trust are very important aspects of DI. Other "soft" criteria are definition and usage of standards, and very human problems around the (lack of) willingness to share [10, 11]. This is especially important in a competitive area such as the Life Sciences. Furthermore, integrating BDB is an interdisciplinary topic as users of DI solutions are biologists, while their developers often are computer scientists or bioinformaticians.

III. THE PAST OF DI IN THE LIFE SCIENCES

The first published systems targeting DI in the Life Sciences all followed a similar scheme: They used scripts to parse the database (flat-files with proprietary formats) into a semi-structured form that was indexed and could be searched (SRS [12], Entrez [13], dbGET [14]). Links between entries were used to offer join-like functionality and, more importantly, web-based browsing. All three systems were developed by biologist researchers and used no database technology; they all still exist, and Entrez and SRS arguably still are the most frequently used DI systems in the Life Sciences.

The second generation of systems was built around the concept of federated databases. Examples include OPM [15], BioKleisli/K2 [16], and DiscoveryLink [17]. Later research (third generation) focused on mediator-based systems, often targeting semantic integration by using ontologies [18, 19]. A very influential system of this kind was TAMBIS [19], based on an ontology with more than 1,000 classes and a sophisticated query rewriting algorithm using description logics. However, none of the second or third generations of DI systems (to our knowledge) still exist today. Note that they had a couple of properties in common. They were based on virtual integration, focused on schemas, paid little attention to actual data, and targeted maximal transparency to relieve the user from having to know which source to query.

Notably, the influence of bioinformatics on DI was even greater than the influence of DI on biological discover. Many of the DI projects were performed when the Web was still young. Bioinformatics at that time was one of the few areas were many, large, and heterogeneous data sets were freely available and where there was a strong need for DI. In particular, the Human Genome Project posed unprecedented challenges to DI and fostered new ideas [5].

IV. CURRENT STATUS

In the early 2000's, ad-hoc systems based on materialization and developed by bioinformaticians replaced mediator-based prototypes since they were adapted much better to the analysis and visualization needs of bioinformatics. In a nutshell, the predominant approach to DI is "PERL + MySQL + XML" a.k.a "Data warehouses" (DWH) in this field.

The first reason for the success of DWH is the fact that maintaining a BDB almost always involves data curation. In this process, researchers may, for instance, discover that some instances should be grouped together or that others should be split. Values are imputed, tuples are filtered, data is added, etc. The resulting, cleansed data is the input to new analysis and may be redistributed. Such operations require data to be present locally. The second key point is economies of scale. Bioinformatics over the years produced several mature libraries for several aspects of building and managing BDBs (BioPerl, BioJava, BioSQL etc.). Today, parsers for loading data from all major biological sources into a variety of relational schemas are freely available; furthermore, modeling can build on proven schemas (e.g., GUS [16], BioWarehouse [20],, GMOD [21]). Building a DWH integrating dozens of BDB nowadays can be achieved in weeks rather than months [22]; however, an open and major problem is to keep a system in a current and consistent state, despite constant changes in the underlying data sources and cleansing operations being performed in the DWH [23].

A few current used systems work with distributed data. The probably most popular one is DAS, the Distributed Annotation System [24] which essentially is a data exchange protocol and a server. BIRN or caBIG are more heavy-weight systems based on shared schemas and ontologies [25, 26].

Last but not least, an increasingly popular technique for solving semantic DI problems is the usage of ontologies, which are hosted in the hundreds in repositories such as BioPortal [27] or OBOFoundry [28]. In practice these are not used during DI, but before, i.e., as structured, controlled vocabularies for annotating entities. Despite a plethora of work devoted to building inference-based DI solutions, none of these are currently being used in production-level systems.

V. Trends in Data Integration

The recent revitalization of DI research in the Life Sciences is accompanied by a change in the perception of the role and necessary functionality of a DI system. Early systems considered DI as an upfront effort leading to a homogeneous and clean integrated database to be then used by researchers for biological discoveries. This probably never was a good idea, because semantics is context-dependent, implying that different users need different ways of "integration" even for the same data sets. Note that highly successful systems, such as Entrez completely refrain from semantic data.

This paradigm change gives rise to some important research trends. First, the process of integration itself, i.e., the "integration workflow", is becoming a research topic in its own. A second trend is the growing importance of sensible ranking, because data sets grow and grow and crisp results do not properly represent the fuzziness and noise in biological data. Note that both these trends are not unique to the Life Sciences; especially recent work in Data Spaces follows similar lines [29]. A third important trend is the increasing usage of Semantic Web technologies to cope with semantic diversity. This trend is especially fuelled by the simplicity of RDF when used as a global data model.

A. *Scientific Workflows*

Typical analysis processes in the Life Sciences are complex, multi-staged, and large. In contrast to typical ETL workflows, their building blocks are complex user-defined functions rather than relational operators. On the other hand, they are focused on data flow, which contrasts them from business workflows. These differences have accumulated into so-called scientific workflow systems (SWFS) [30], an area largely driven by the bioinformatics community. SWFS aim to provide an environment to guide a scientific analysis process from its design to its execution. The analysis processes are represented at a high level of abstraction which enhances flexibility, reuse, and modularity while allowing for optimization, parallelization, etc. [31]. SWFS may also deal with failure handling, scheduling, and monitoring. All steps plus intermediate results are traced to enhance reproducibility. Using a SWFS for DI implements the paradigm of "integration is analysis".

Challenges: Very interestingly, SWF can be shared, searched, compared etc., opening a door to the exchange of mature and specialized DI solutions (e.g., myExperiment is a portal that hosts about 900 scientific workflows; BioCatalogue is a repository of more than 3,000 web services to be called in workflows [32]). Such non-technical aspects offer a number of open research questions. First, users might want to search for a specific workflow having only weak constraints in terms of requested functionality. Alternatively, users might have a concrete workflow and are interested in finding similar ones. Both problems boil down to approximate queries in SWF repositories. Research in these directions has recently started [33, 34], but many questions such as searching heterogeneous workflow models, proper similarity measures, and scalability are still open.

There are also interesting open questions in the management of SWF. Provenance in SWFS is a key concept since it supports reproducibility and helps assessing the quality of results [35, 36]. SWFS uses various models of computation when executing workflows, providing various kinds of trace while the amount of data produced is enormous. Open research questions in this area include modeling runs [35], designing scalable management and analysis methods for storage frameworks [37, 38], visualization and user-interface topics [34], comparing workflow runs based on their provenance data [33], querying and indexing provenance information [38-40], and orchestrating distributed web services calls in an efficient way [41].

B. *Ranking*

One of the key challenges in DI in the Life Sciences is to help users choose between alternative pieces of information, such as choosing between conflicting updates when maintaining a DWH [23], choosing between different data sources [42], or choosing among several answers when querying a DI system [43]. Such choices are best supported by sensible ranking methods [44]. However, even widely used portals still do not provide any ranking services although queries often produce huge amounts of results. For instance, searching for the set of genes involved in breast cancer returns 1,472 answers in the reference database EntrezGene without any ranking in terms of importance.

Challenges: Despite the large body of work on ranking search results performed in the database community [45, 46], no approach is currently able to take into account features of Life Science data. Here, entries often are text-centric, which requires the inclusion of text mining methods in the ranking [47]. Facts are not always of equal strength. In the example above, some genes may be clearly involved in breast cancer while for others this relationship might still be putative. Links between entries are very important to augment the confidence in results; however, they cannot be considered in isolation, but only in the context of the entire network of linked entities [42]. This may be achieved using PageRank-style algorithms in systems such as Biozon [48] or soft-rank [49] while various criteria (e.g., the reliability of sources providing data or links) should additionally be considered [50]. End users also have different expectations, such as either looking for established or for more surprising results [51]. Also, works on probabilistic databases may play an important role [52] . Another open problem is the proper evaluation of different ranking methods.

C. *Semantic Web*

Semantic heterogeneity on all levels (schema elements, attribute values, object identifiers) has always been a strong impediment to the integration of biological data [5]. Recently, many groups promote the usage of Semantic Web (SW) technologies to alleviate these problems. In their view, the combination of RDF as a common data model, SPARQL as a query language for seamlessly crossing database borders, and the inference capabilities of OWL build a tool stack that is ideally suited for the Life Sciences [53, 54]. This has lead to the creation of a series of prototypes trying to showcase advantages of this approach [54-56].

Challenges: A number of issues are still open and offer ample opportunities for research. First, although the usage of ontologies is commonplace in the Life Sciences (see above), they mostly are not interlinked, most of them are not used as widely as they could be, and/or many of them change rapidly.

Increasing their interconnectedness and their usage calls for research in ontology matching [57] and automated annotation [58-60] (with relationships to work in social networks [61]). Another important line of research tackles the evolution of ontologies [62]. Note that ontologies are far less formal than what the SW community assumes to be an ontology, leading to frequent misunderstandings between researches.

Second, SPARQL still misses essential features for DI, such as user-defined functions, transitive predicates for navigating heterogeneous graphs, and aggregation. SPARQL also does not adequately support distributed queries [63]. Several proposals are on the way to extend SPARQL with such features (e.g. [64, 65]), but for many of them no efficient implementation is currently available.

Third, using RDF as common data model does not by itself solve the problem of semantic heterogeneity, but only postpones it until query time. Current efforts on the large-scale "triplification" of biological data sets [53, 66] must be augmented with methods for object unification and for mapping of ontology terms to unleash the full potential of the SW.

VI. CONCLUSIONS

Research in DI and research in the Life Sciences for long had and still have an intimate relationship. In few (if any) other areas, there is such an abundance of complex, heterogeneous, and freely available data sets coupled with strong scientific incentives. The Life Sciences offer an excellent testbed for DI solutions, and in turn this field also had a strong impact on DI research (e.g., biological data was a main motivator for early work in semi-structured data which then become XML [67]).

However, experiences have shown that to have an impact, DI research must work closely with domain experts and must develop methods that keep these users in the loop, leveraging their knowledge instead of trying to replace it. Also, biologists often remain unimpressed if DI projects stop with the creation of "clean" data sets; instead, the separation between integration and analysis must be overcome. Whether this is best done by empowering systems with more and more semantic knowledge (the SemWeb way) or by empowering people to build their own systems (the SciWF way) remains an open question, leaving ample space for research in both directions.

VII. ABOUT THE TUTORIAL

A. Intended Audience and outline

Researchers and research-oriented engineers in data integration, knowledge management, scientific databases, semantic web, object ranking, and workflow systems. The tutorial will offer a survey over a wide range of topics. Familiarity with biological terms is not necessary.

Attendees will learn about (i) specificities of biomedical databases, (ii) lessons learnt from 20 years of data integration research in this field, and, most importantly, (iii) current hot topics and opportunities for research, especially from a database / data engineering point-of-view. The latter will make up for approximately half of the tutorial.

B. About the Speakers

Sarah Cohen-Boulakia graduated from Université Paris-Sud, is currently an assistant professor in the same University and has spent a post-doc in the Database group at the University of Pennsylvania. Her major research interest is data integration in the Life Sciences, with a focus on provenance in scientific workflows. She collaborates closely with biologists, physicians, and bioinformaticians in European and International projects.

Ulf Leser holds a CS degree from TU München and a PhD in Query Optimization from TU Berlin. He currently is a Professor for Knowledge Management in Bioinformatics at Humboldt Universität in Berlin. His primary research interests are data integration, information extraction, and bioinformatics. He is part of a number of interdisciplinary projects with Biologists and MDs in which he joyfully experiences the challenges and benefits of directly working together with biologists.

Acknowledgments: We thank S. Davidson, F. Lisacek, B. Ludäscher, F. Naumann, and N. Paton for their feedback on an early version of this manuscript.

REFERENCES

1. Karp, P.D., *Report of the Workshop on Interconnection of Molecular Biology Databases*. 1994, SRI International Artificial Intelligence Center, Stanford, California.
2. Goble, C. and R. Stevens, *State of the nation in data integration for bioinformatics.* J Biomed Inform, 2008. **41**(5): p. 687-93.
3. Searls, D.B., *Data Integration: Challenges for Drug Discovery.* Nature Reviews, 2005. **4**: p. 45-58.
4. Stein, L.D., *Integrating biological databases.* Nat Rev Genet, 2003. **4**(5): p. 337-45.
5. Davidson, S., G.C. Overton, and P. Buneman, *Challenges in Integrating Biological Data Sources.* Journal of Computational Biology, 1995. **2**(4): p. 557-572.
6. Buneman, P., J. Cheney, W.-C. Tan, and S. Vansummeren. *Curated Databases*. in *PODS*. 2008. Vancouver, Canada
7. Baumgartner Jr, W.A., K.B. Cohen, L.M. Fox, G. Acquaah-Mensah, and L. Hunter, *Manual curation is not sufficient for annotation of genomic databases.* Bioinformatics, 2007. **23**(13): p. i41.
8. Cochrane, G.R. and M.Y. Galperin, *The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources.* Nucleic Acids Res, 2010. **38**(Database issue): p. D1-4.
9. Gerstein, M.B., et al., *What is a gene, post-ENCODE? History and updated definition.* Genome Res, 2007. **17**(6): p. 669-81.
10. Schofield, P.N., et al., *Post-publication sharing of data and tools.* Nature, 2009. **461**(7261): p. 171-3.
11. Paton, N.W., *Managing and sharing experimental data: standards, tools and pitfalls.* Biochem Soc Trans, 2008. **36**(Pt 1): p. 33-6.
12. Etzold, T., A. Ulyanov, and P. Argos, *SRS: Information Retrieval System for Molecular Biology Data Banks.* Methods in Enzymology, 1996. **266**: p. 114-128.
13. Schuler, G.D., J. Epstein, H. Ohkawa, and J.A. Kans, *Entrez: Molecular Biology Database and Retrieval System.* Methods in Enzymology, 1996. **266**: p. 141-161.
14. Akiyama, Y., S. Goto, I. Uchiyama, and M. Kanehisa. *WebDBGET: an integrated database retrieval system which provides hyper-links among related database entries*. in *2nd Meeting on Interconnection of Molecular Biology Databases*. 1995. Cambridge, UK.
15. Chen, I.A. and V.M. Markowitz, *An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools.* Information Systems, 1995. **20**(5): p. 393-418.
16. Davidson, S., et al., *K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources.* IBM Systems Journal, 2001. **40**(2): p. 512-531.
17. Haas, L.M., et al., *DiscoveryLink: A System for Integrated Access to Life Sciences Data Sources.* IBM Systems Journal, 2001. **40**(2): p. 489-511.
18. Miled, Z.B., N. Li, G. Kellett, B. Sipes, and B. O., *Complex Life Science Multidatabase Queries.* Proceedings of the IEEE, 2002. **90**(11): p. 1754-1763.

19. Paton, N.W., et al. *Query Processing in the TAMBIS Bioinformatics Source Integration System*. in *SSDBM*. 1999. Cleveland, Ohio.
20. Lee, T.J., et al., *BioWarehouse: a bioinformatics database warehouse toolkit.* BMC Bioinformatics, 2006. **7**: p. 170.
21. Stein, L.D., et al., *The generic genome browser: a building block for a model organism system database.* Genome Res, 2002. **12**(10): p. 1599-610.
22. Trissl, S., et al., *Columba: An Integrated Database of Proteins, Structures, and Annotations.* BMC Bioinformatics, 2005. **6:81**.
23. Ives, Z.G. *Data Integration and Exchange for Scientific Collaboration*. in *DILS*. 2009. Manchester, UK.
24. Jenkinson, A.M., et al. *Integrating Biological Data – The Distributed Annotation System* in *DILS*. 2008. Evry, France.
25. Bug, W., et al., *Data federation in the Biomedical Informatics Research Network: tools for semantic annotation and query of distributed multiscale brain data.* AMIA Annu Symp Proc, 2008: p. 1220.
26. Saltz, J., et al., *caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid.* Bioinformatics, 2006. **22**(15): p. 1910-6.
27. Noy, N.F., et al., *BioPortal: ontologies and integrated data resources at the click of a mouse.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W170-3.
28. Smith, B., et al., *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol, 2007. **25**(11): p. 1251-5.
29. Halevy, A., M. Franklin, and D. Maier. *Principles of Dataspace Systems*. in *PODS*. 2006. Chicago, USA.
30. Ludaescher, B., et al., *Scientific workflow management and the Kepler system.* Concurrency and Computation: Practice and Experience, 2005. **18**(10): p. 1039-1065.
31. Wieczorek, M., R. Prodan, and T. Fahringer, *Scheduling of scientific workflows in the ASKALON grid environment.* SIGMOD Record, 2005. **34**(3): p. 56 - 62.
32. Bhagat, J., et al., *BioCatalogue: a universal catalogue of web services for the life sciences.* Nucleic Acids Res, 2010. **38 Suppl**: p. W689-94.
33. Bao, Z., S. Cohen-Boulakia, S.B. Davidson, A. Eyal, and S. Khanna, *Differencing Provenance in Scientific Workflows*, in *ICDE*. 2009, IEEE Computer Society.
34. Scheidegger, C., E. , H. Vo, T. , D. Koop, J. Freire, and C. Silva, T., *Querying and re-using workflows with VisTrails*, in *SIGMOD* 2008, ACM: Vancouver, Canada.
35. Moreau, L., et al., *Special Issue: The First Provenance Challenge.* Concurrency and Computation: Practice and Experience, 2008. **20**(5): p. 409-418.
36. Davidson, S., B. and J. Freire, *Provenance and scientific workflows: challenges and opportunities*, in *SIGMOD*. 2008: Vancouver, Canada.
37. Kumar, A.M., B. Shawn, T. McPhillips, and B. Ludäscher, *Efficient provenance storage over nested data collections*, in *EDBT*. 2009: Saint Petersburg, Russia.
38. Biton, O., S. Cohen-Boulakia, S.B. Davidson, and C.S. Hara, *Querying and Managing Provenance through User Views in Scientific Workflows*, in *ICDE*. 2008.
39. Kumar, A.M., B. Shawn, and L. Bertram, *Techniques for efficiently querying scientific workflow provenance graphs*, in *EDBT*. 2010: Lausanne, Switzerland.
40. Missier, P., N.W. Paton, and K. Belhajjame, *Fine-grained and efficient lineage querying of collection-based workflow provenance*, in *EDBT*. 2010, ACM: Lausanne, Switzerland.
41. De Stasio, A., M. Ertelt, W. Kemmner, U. Leser, and M. Ceccarelli. *Exploiting Scientific Workflows for Large-scale Gene Expression Data Analysis*. in *Int. Symposium on Computer and Information Sciences*. 2009. Cyprus.
42. Cohen-Boulakia, S., S.B. Davidson, C. Froidevaux, Z. Lacroix, and M.-E. Vidal, *Path-based systems to guide scientists in the maze of biological data sources.* Journal of Bioinformatics and Computational Biology, 2006. **4**(5): p. 1069-1095.
43. Birkland, A. and G. Yona, *BIOZON: a system for unification, management and analysis of heterogeneous biological data.* BMC Bioinformatics, 2006. **7**: p. 70.
44. Bleiholder, J., et al., *BioFast: Challenges in Exploring Linked Life Science Sources.* SIGMOD Record, 2004. **33**(2).
45. Kaushik, C., G. Venkatesh, H. Jiawei, and X. Dong, *Ranking objects based on relationships*, in *SIGMOD*. 2006: Chicago, USA.
46. Agrawal, R., R. Rantzau, and E. Terzi, *Context-sensitive ranking*, in *SIGMOD*. 2006: Chicago, USA.
47. Lee, W.-J., L. Raschid, H. Sayyadi, and P. Srinivasan. *Exploiting Ontology Structure and Patterns of Annotation to Mine Significant Associations between Pairs of Controlled Vocabulary Terms*. in *DILS*. 2008. Evry, France.
48. Shafer, P., T. Isganitis, and G. Yona, *Hubs of knowledge: using the functional link structure in Biozon to mine for biologically significant entities.* BMC Bioinformatics, 2006. **7**: p. 71.
49. Varadarajan, R., et al., *Flexible and efficient querying and ranking on hyperlinked data sources*, in *EDBT*. 2009: Saint Petersburg, Russia.
50. Cohen-Boulakia, S., et al. *Selecting Biomedical Data Sources According To User Preferences*. in *Int. Conference on Intelligent Systems in Molecular Biology (ISMB/ECCB)*. 2004. Glasgow, UK.
51. Hussels, P., S. Trissl, and U. Leser. *What's new? What's certain? Scoring Search Results in the Presence of Overlapping Data Sources*. in *DILS*. 2007. Philadelphia, US.
52. Li, J., B. Saha, and A. Deshpande, *A unified approach to ranking in probabilistic databases.* Proceedings of the VLDB Endowment, 2009. **2**(1): p. 502-513.
53. Slater, T., C. Bouton, and E.S. Huang, *Beyond data integration.* Drug Discov Today, 2008. **13**(13-14): p. 584-9.
54. Ruttenberg, A., et al., *Advancing translational research with the Semantic Web.* BMC Bioinformatics, 2007. **8 Suppl 3**: p. S2.
55. Post, L.J., M. Roos, M.S. Marshall, R. van Driel, and T.M. Breit, *A semantic web approach applied to integrative bioinformatics experimentation: a biological use case with genomics data.* Bioinformatics, 2007. **23**(22): p. 3080-7.
56. Sahoo, S., O. Bodenreider, K. Zeng, and A. Sheth. *An Experiment in Integrating Large Biomedical Knowledge Resources with RDF: Application to Associating Genotype and Phenotype Information*. in *Workshop on Health Care and Life Sciences Data Integration for the Semantic Web* 2007. Banff, Canada.
57. Udrea, O., L. Getoor, and R.J. Miller. *Leveraging Data and Structure in Ontology Integration*. in *SIGMOD*. 2007. Beijing, Chiina.
58. Rhee, S.Y., V. Wood, K. Dolinski, and S. Draghici, *Use and misuse of the gene ontology annotations.* Nat Rev Genet, 2008. **9**(7): p. 509-15.
59. Groth, P., B. Weiss, H.-D. Pohlenz, and U. Leser, *Mining phenotypes for gene function prediction.* BMC Bioinformatics, 2008. **9**: p. 136.
60. Karaoz, U., et al., *Whole-genome annotation by using evidence integration in functional-linkage networks.* Proc Natl Acad Sci U S A, 2004. **101**(9): p. 2888-93.
61. Markines, B., et al. *Evaluating similarity measures for emergent semantics of social tagging*. in *WWW*. 2009. Madrid, Spain.
62. Hartung, M., T. Kirsten, and E. Rahm. *Analyzing the Evolution of Life Science Ontologies and Mappings*. in *DILS*. 2008. Evry, France.
63. Quilitz, B. and U. Leser. *Querying Distributed RDF Data Sources with SPARQL*. in *ESWC*. 2008. Teneriffa, Spain.
64. Kochut, K. and M. Janik. *SPARQLeR: Extended Sparql for Semantic Association Discovery*. in *ESWC*. 2007. Innsbruch, Austria.
65. Schenk, S. and S. Staab. *Networked graphs: a declarative mechanism for SPARQL rules, SPARQL views and RDF data integration on the web*. in *WWW*. 2008.
66. Belleau, F., M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette, *Bio2RDF: Towards a mashup to build bioinformatics knowledge systems.* Journal of Biomedical Informatics, 2008. **41**(5): p. 706-716.
67. Buneman, P., S. Davidson, G. Hillebrand, and D. Suciu. *A Query Language and Optimisation Techniques for Unstructured Data*. in *SIGMOD*. 1996. Tuscon, Arizona.