Ontology-based Registration of Entities for Data Integration in large biomedical Research Projects

Toralf Kirsten[†‡], Alexander Kiel[‡]
[†]Interdisciplinary Centre for Bioinformatics
[‡]Institute for Medical Informatics, Statistics and Epidemiology
University of Leipzig, Germany

tkirsten@izbi.uni-leipzig.de, alexander.kiel@life.uni-leipzig.de

Abstract: Large biomedical projects often include workflows running across institutional borders. In these workflows, data describing biomedical entities, such as patients, bio-materials but also processes itself, is typically produced, modified and analyzed at different locations and by several systems. Therefore, both tracking entities within inter-organizational workflows and data integration are often crucial steps. To address these problems, we centrally register entities and their relationships by using a multi-layered model. The model utilizes an ontology and a typed system graph to semantically describe and classify entities and their relationships but also to access entity data on demand in their original source. Moreover, this integration approach allows to centrally track entities along the project workflows and can be used in explorative data analyses as well as by other data integration approaches using the registered entity relationships. We describe the model, the utilized ontology, and a system implementing this approach, which is applied in a large biomedical research project.

1 Introduction

Biomedical research projects are typically initiated to investigate biological and medical phenomena and their implications. For instance, they study the causes and the therapy and cure process of patients with a specific disease, such as different variants of cancer or HIV. For this purpose, lots of data describing patients, their findings and treatments is captured and analyzed. Additionally, it became quite common in recent years to utilize molecular-biological experiments, such as gene expression and mutation analysis and sequencing based on high-throughput (microarray-based) techniques, to determine the interplay of biological objects (genes, proteins etc.) on the genetic level. The cause is that serious diseases are deeply affected by the molecular-biological conditions for genes and proteins. The resulting molecular-biological data is combined with clinical or patient-related data to study the genotype-phenotype relationship under certain circumstances. Both, patient-related and molecular-biological data is typically produced along inter-organizational workflows. In particular, in large biomedical projects, this data is produced at different locations and by different systems. Therefore, there are several types of data which are usually stored in various sources and need to be integrated to execute comprehensive analyses.

LIFE[1] (*L*eipzig *I*nterdisciplinary Research Cluster of Genetic *F*actors, Clinical Phenotypes and *E*nvironment) is a biomedical project in the described context. The project aims at investigating causes for several civilization diseases including adiposity, diabetes, depression, and allergies by finding factors on the genomic and clinical level but also by considering the environment and the lifestyle of patients. Various partners are involved in this project including several institutions of the University of Leipzig and external organizations, like local hospitals and other research institutions and laboratories. On the one hand, these institutions participate with specific research questions. For instance, one group focuses on nutrition patterns, tobacco and alcohol consumption of patients with adiposity while another group is interested in testing patients according to a specific allergy type. On the other hand, there are institutions providing different biomedical investigation services. Most of such services subsume several types of laboratory analyses according to different types of bio-material, such as blood or urine specimens of patients. Further investigation services include magnetic resonance tomography and radiology analyses to screen special anatomical parts of patients.

Due to these different research directions and services, data is captured and generated in various inter-organizational workflows. While some data is captured in specific patient checkups, there are also structured patient interviews (predefined questions and answer sets) and various material analyses taking place in different laboratories. Many input systems store data in relational databases whereas most laboratory devices generate data in XML, CSV, and proprietary data files. Due to this large number of heterogeneous and distributed sources and the limited IT resources, it is currently impossible to create a large scale global schema and use it to integrate all of this data. However, an integrated database would allow tracking the entities to their origin and to explore their predecessors and successors along the project workflows. This is currently an important requirement in *LIFE* and potentially other projects alike. Therefore, we make the following contributions.

- We designed an integration approach to centrally register entities and relationships between them. The approach utilizes a multi-layered model containing an ontology and a typed system graph to which the entities and their relationships are associated. While the ontology semantically describes entities and their relationships, the system information is used to access the entity data within their source system. Additionally, the ontology and the typed system graph can be used for querying semantical and source-specific entity sets.
- We developed LIO, the *LIFE* Investigation Ontology, and use it on the ontology layer of our model. LIO is based on the General Formal Ontology (GFO) [HBH$^+$06] and comprises concepts according to the two fundamental GFO concepts, namely *presential* and *process*. Presentials are entities that can be described for one point in time, such as study participant, specimen, and data, whereas material and data generation processes are associated with time intervals. The ontology is used to semantically describe and classify entities but also for querying.
- We have implemented a service-based infrastructure realizing the multi-layered registration model. We apply this system in different scenarios including descriptive

[1] `http://www.uni-leipzig-life.de`

and explorative analysis but also for data integration.

The rest of the paper is organized as follows. In the next section we introduce the developed ontology and their use in the multi-layered registration model. Section 3 shows the system architecture of a software application implementing the sketched approach. In Section 4 we discuss related work and conclude in Section 5.

2 Semantical classification and registration of entities

In this section we firstly focus on the *LIFE* Investigation ontology. This ontology is used in our multi-layered registration model to semantically classify entities. We introduce this model before we describe how it is used for entity tracking and querying.

2.1 The LIFE Investigation Ontology

Recently, ontologies became increasingly important in life sciences. Typically, they are used to semantically describe and classify entities, which are clinically and biologically relevant under certain circumstances. We developed the *LIFE I*nvestigation *O*ntology (LIO) to semantically describe the entities, which are created in the *LIFE* project. Associating ontology concepts to entities enables an entity classification and, thus, an ontology-driven search and navigation in the entity space. Figure 1 gives an overview of LIO. It is based on the General Formal Ontology (GFO) [HBH$^+$06] defining fundamental ontology concepts, such as *item*, *individual*, and *category* as well as their hierarchical order. LIO mainly uses two GFO concepts, namely *presential* and *process*. *Presentials* are entities without temporal extension and, thus, can be described for a point in time whereas *processes* are associated with time intervals.

By now, LIO defines three distinct types of *presentials* and two *process* types. The concept *participant* is used to describe the persons (and further animals) who are included in this large scale project (LIFE), whereas the concept *specimen* specify the bio-material which should be analyzed in a laboratory. This bio-material is typically extracted from the participant in a so-called *material generation process*, e.g., when a patient delivers blood material. At the end, this bio-material is analyzed in a *data generation process* bringing out *data* describing the study participant or the bio-material in a given context. This data can then be iteratively analyzed producing new data. Since we have started to develop LIO, the number of concepts and relationships is relative small. Currently, LIO comprises 41 concepts and 57 relationships. In future, we will extend LIO with further conncepts and relationships depending on the project requirementsin *LIFE*.

Both, presentials and processes are organized in a "*is-a*" hierarchy, i.e., each leaf concept is connected with the general root concept by using the "is-a" relationships on the path from the leaf to the root concept. The entities are associated with the most detailed ontology concepts of LIO, e.g., the tobacco interview or nutrition behavior interview. Therefore, the

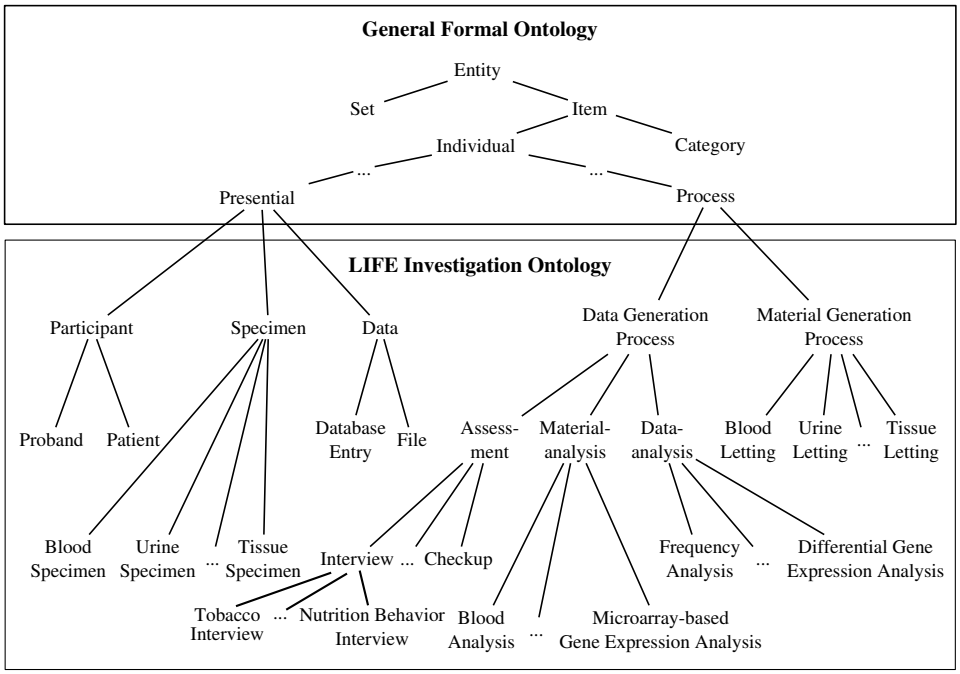

Figure 1: Overview of high-level concepts and their "is-a" relationships in LIO

generalization paths within the *is-a* hierarchy can be used to generate and execute queries such as '*How many patients have been interviewed?*'.

2.2 The multi-layered registration model

The registration model consists of three interrelated levels. On the top level, an ontology defines the semantical types of entities and their semantical relationships. The second level associates the ontology concepts with physical systems in which the entity data is stored, and the third level associates entities to both, ontology concepts and physical systems.

There are many ontologies available, which can be used in our registration model. Typically, the ontology selection depends on the domain and the purpose it should reflect. Throughout this paper, we use the introduced *LIFE* Investigation Ontology. For simplicity, we model the ontology $\mathcal{O} = (C, R)$ as set of concepts C which are interrelated by the set of relationships R: $C \times C$ and let axioms etc. out of the scope of this paper.

Among the "is-a" organization, LIO also utilizes the relationships "is-used-in" and "generates" to specify the presentials whose instances are the input and output of a process. Figure 2 (upper part) shows an example: a blood specimen is taken from a patient and analyzed in a laboratory process generating a complete blood count. Hence, complex

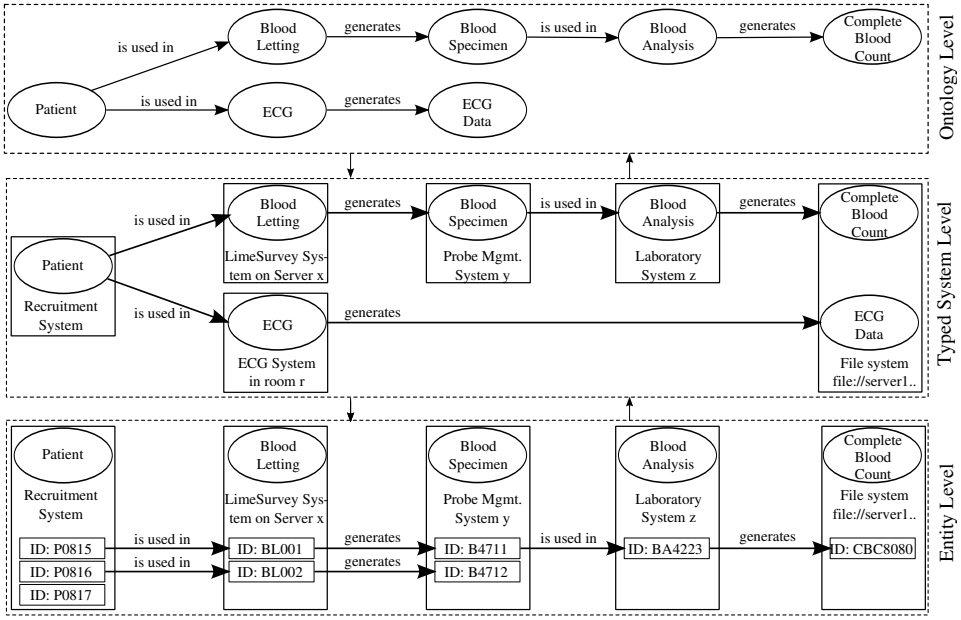

Figure 2: The multi-layered registration model: Exemplified interplay between LIO (upper part), the typed system graph (middle part), and the entity graph (lower part).

workflows can be modeled by combining the inputs and outputs of multiple processes.

Entities are typically generated and modified by a set of systems, which is available in biomedical project environments. Each system has a unique location, which can be specified in some way, e.g., by using an uniform resource identifier (URI). Hence, multiple installations of the same software application result in different systems. However, there are systems, such as laboratory devices modifying and analyzing a given bio-material, which do not create and store data directly. In this case, we assume that the laboratory staff provide minimal information by describing the process using a predefined input schema and system. We also allow to describe processes covering a little organization unit to limit the later input effort. Typically, such organization units are primarily focused on special analyses, such as for the complete blood count or microarray-based gene expression analysis. Due to these analyses require special laboratory equipment and run typically in the same way every time they are executed (within the same laboratory), it is not necessary to describe these processes at a very detailed level, e.g., for every step and every used laboratory device. Hence, there are systems storing data as result of a process but also systems which are used to capture and store metadata about processes. We collect all systems in the set S.

We further associate the ontology concepts (specific presentials and processes) with systems to semantically describe and specify the types of entities which a single system creates or modifies. Hence, such a typed system $s^t = (c, s)$ is a part of a system $s \in S$ for which entities of type $c \in C$ of $\mathcal{O}$ exist. The set of typed systems S^t and their re-

lations M^{S^t}: $S^t \times S^t$ are represented in a typed system graph $G^{S^t} = (\mathcal{O}, S, S^t, M^{S^t})$. The relations in M^{S^t} are associated with relations of the ontology $\mathcal{O}$ and, thus, use their labels. While ontology concepts correlate orthogonally with systems, we assume that not for every combination exists a typed system. Figure 2 (middle part) shows an example for a typed system graph. Each typed system associates an ontology concept from Figure 2 (upper part) with a system of S.

We utilize the typed system graph to describe inter-organizational workflows. These workflows create entities at different locations resulting in an entity graph $G^E = (G^{S^t}, E, M^E)$. The G^E consists of a set of entities E and their relationships $M^E : E \times E$ between them. The entities and their relationships are associated with the typed system graph G^{S^t}. Hence, the entities are semantically described but also their relationships. Figure 2 (lower part) shows an example of an entity graph. A recruitment system stores data about three patients which can be obtained by using their unique identifiers *P0815*, *P0816*, and *P0817*. Two of these patients dispense a blood specimen (identified by *B4711* and *B4712*). These specimens are obtained by a nurse or a doctor who thereafter describe each blood letting process using the specific installation of the LimeSurvey [Lim10] system on server x. While the blood specimens are physically stored in a biobank, they are described in the probe management system y. One of the specimens is utilized to create a complete blood count (in laboratory z) which is a special type of a material analysis. This laboratory process is described by the laboratory staff using a special laboratory system z. The analysis generates a file that is stored within the file system on a special file server. Using both, the knowledge about the processes of obtaining the blood specimen from a patient and the material analysis in the laboratory as well as the location where the resulting data file is stored, allows to track the process of sample and data creation and manipulation on the one hand and to access the data in their source systems (e.g., database, file system) on the other hand.

2.3 Entity registration and querying

We use the described model to centrally register all entities. Each entity is associated with both, a semantical type, which is based on an ontology concept of LIO, and a system, which is used to capture and store data about the entity. This allows classifying and grouping the entities by their semantical type but also to locate the data of the entities. We use a central application managing entity references and their relationships (see Section 3). Within this registry application an entity is represented by the triple (ontology concept, system, entity identifier). The entity identifier may not be globally unique but for the typed system (ontology concept + system). An entity relationship is a directed connection between two entities. We use the form (source entity triple, target entity triple) to specify entity relationships.

The central registration has several advantages. Firstly, it can be used to validate whether data about an entity already exists without accessing any source system. This is especially important when a new entity is created that should be related to another entity in the registration process. For instance, registering a specific blood specimen according to

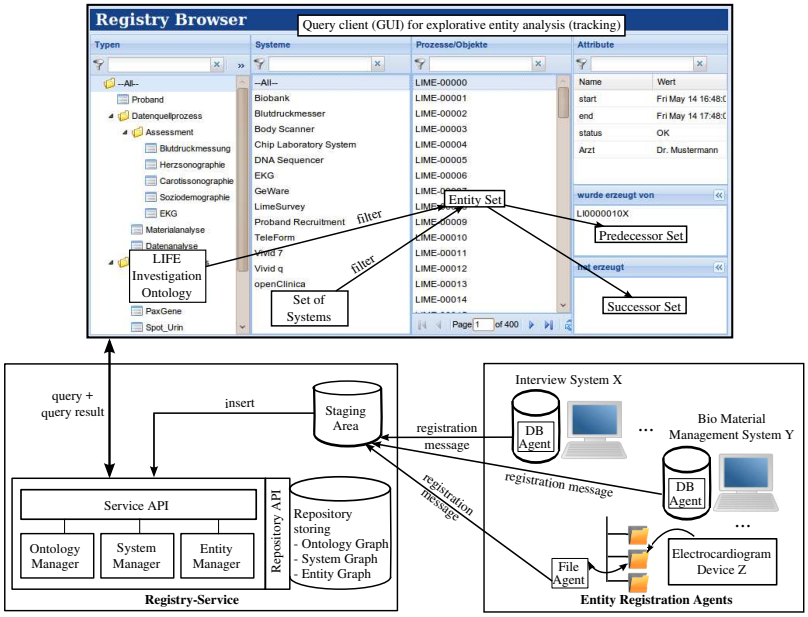

Figure 3: System architecture

the process in which it is created necessitates that the blood letting process is registered before. Secondly, it can be utilized by some data integration approaches. Typically, such approaches require entity references and their relationships for instance data integration. Moreover, the semantical entity classification makes it possible to identify schema fragments of two source systems which need to be included in a schema matching process, e.g., the schema fragments of two systems which have been used to capture interview data. Finally, the central registration can be directly included in several analysis scenarios, e.g., descriptive statistics and explorative analysis. In particular, it allows to track entities from their origin to systems where they have been used.

3 System architecture

We implemented a service-based infrastructure to register entities and their relationships. Figure 3 shows the architecture of the system. It consists of a central *registry service* and several clients. The registry service consists of multiple manager modules behind a uniform application interface (API). These manager modules are responsible for and execute queries on the ontology graph (ontology manager), the typed system graph (system manager), and the entity graph (entity manager). All data, i.e., ontology concepts, typed systems, and entities as well as the relationships between them, are stored in a repository using a relational database.

There are two types of clients, namely *entity registration agents* and *query clients*. The entity registration agents are loosely coupled with input systems, such as the interview input system and the file system directory in that the electrocardiogram device stores data. The goal of these agents is to inform the central registry service when new entities arrive in input systems. Typically, the agents are implemented as triggers or stored procedures when a data input system utilizes a relational database to store the data. In case of input systems producing data files, the agents are implemented as shell scripts and special programs, which regularly look for new files in the system's data directory. While database-based agents can easily and directly use the entity data from a defined database table for the registration, the file-based agents need to parse the generated file(s) to extract entity data. This parse process is short and not very resource-intensive since it only extracts the entity identifier. Each agent is configured with the semantical type and the physical system of entities. Together with the entity identifier, both are used in the registration process.

Normally, the agents register the entities according to predecessor entities within the workflow. For example, an urine specimen is registered according to its urine letting process. This bio-material separation process again is registered for the study participant from which the bio-material originates. An exception of this procedure is the registration of new study participants since they are the root of each entity graph and, thus, have no predecessor.

To keep the data within the registry repository consistent, the central registry service only accepts an entity registration when its predecessor entity has been registered before and, thus, already exists in the repository. However, this would require that the entities are synchronously registered when they are created which can not be guaranteed in most cases. The reason for this is twofold. Firstly, an agent is autonomous, i.e., it has no knowledge about all the other agents generating and sending registration messages to the central registration service. Secondly, the implementation detail of an agent influences the time when new entities are registered. For instance, an agent can be implemented as database trigger registering a new entity when a new entry is inserted into the database table the agent is configured for. Contrary, another agent regularly looks for new files (new entities) in a specific file directory every 15 minutes and, thus, underlie a scheduling. To address the synchronization, all registration messages are firstly collected in a central staging area (relational database) from where they are periodically imported into the registry service by resolving the import order.

The query clients are used to retrieve entity references and their relationships. Currently, we use such clients in two scenarios. On the one hand, they represent specific wrappers of data integration solutions using the central registry to resolve the instance data integration. On the other hand, a query client is utilized by an analysis application. This application allows both, descriptive statistics like "*How many patient have been interviewed?*" and an explorative analysis that gives an overview of participant-specific checkups, taken bio-materials and generated data. We expect to register various data and material generation processes as well as data items and bio-materials for approx. 25,000 participants in the *LIFE* project. Assuming there are on average 20 assessments (interviews, checkups etc.) and materials and data analyses for a single participant, the total number of manged entity references within the registry application is 25,000*20=500,000.

4 Related work

Many integration approaches use entity mappings in the data integration process, in particular for instance data integration. Some integration approaches explicitly model and analyze mappings between entity collections, e.g., [LMNR04, LRV04, KDKR05, KR06]. Like in our registration approach, all these approaches utilize special graph-based structures to represent entities and their mappings. In [LMNR04, LRV04], the authors define a source and object (entity) graph but use them primarily for an algorithm to find the most appropriate path through life science data sources. Similar to our approach [KDKR05] use a domain model to join data from different sources. While this domain model semantically characterizes sources and mappings between them, it does not explicitly use an ontology like our approach with LIO. The integration approach described in [KR06] extracts mapping data from sources and materializes them in a separate mapping database. The database use a star-like schema allowing to efficiently compute mapping compositions which are then used for query processing and to join data from multiple sources. Our approach also materializes entity mappings in a separate repository but uses a more flexible schema and associates ontology concepts to sources and entities. The biggest difference between these integration approaches and our approach is that our solution register entities and their mappings which is a prerequisite for later integration of this data.

Novel and related approaches from semantic web community has been collected under the term *Linked Open Data* (LOD) [BHBL09, Lin10]. These approaches make heavily use of Semantic Web techniques, like the OWL and RDF languages as well as those allowing to store (typically known as triple stores) and to query (e.g., SPARQL query language) such data. Although our approach can also be implemented using these techniques, the repository of our system utilizes a relational database to store data and SQL for querying.

Similar to LIO, the Ontology for Biomedical Investigation (OBI) [CBG+08] has been provided to uniformly describe clinical and life-science investigations. OBI is an Open Biomedical Ontology (OBO) that is based on Basic Formal Ontology (BFO) [GSG04] and is managed by the OBO Foundry [SAR+07]. While it is possible to map some LIO concepts to OBI concepts, e.g., *material generation process* to *OBI_0000659 – specimen creation*, OBI has not the focused view of a clinical trial as we need it. We see LIO as a domain specific ontology which serves us well in *LIFE*. We let an ontology mapping to OBI open to future work.

5 Conclusions

In this paper we presented an integration approach to centrally register entities and their mappings. The approach utilizes a multi-layered model in which entities are semantically described by combining concepts of an ontology and systems where data about the entities is stored. Therefore, entities can be classified and queried using the provided ontology whereas the associated system information is used to access the entity data within the sources. The entities are explicitly registered according to their predecessor entities along

the workflow in which the entities have been generated. Using these chains of relationships allow tracking entities to their origin and resolving the set of predecessors that have been used to create an entity. Moreover, the approach can be utilized in various data integration approaches for instance data integration. We implemented and applied a service-based infrastructure realizing this approach in a real and large scale biomedical project.

Acknowledgment We thank Christoph Engel and Michael Kleinert for discussions according to clinical and laboratory workflows and special implementation aspects of the service clients.

References

[BHBL09] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data - The Story So Far. *International Journal On Semantic Web and Information Systems*, 5(3):1–22, 2009.

[CBG+08] Mélanie Courtot, William Bug, Frank Gibson, et al. The OWL of Biomedical Investigations. In *Proc. 5th Intl. Workshop on OWL Experiences and Directions, Karlsruhe*, 2008.

[GSG04] Pierre Grenon, Barry Smith, and Louis Goldberg. Biodynamic ontology: applying BFO in the biomedical domain. *Stud Health Technol Inform*, 102:20–38, 2004.

[HBH+06] H. Herre, B. Hellerand P. Burek, R. Hoehndorf, F. Loebe, and H. Michalek. General Formal Ontology (GFO): A Foundational Ontology Integrating Objects and Processes. Part I: Basic Principles. Research Group Ontologies in Medicine. Technical report, Onto-Med Goup, University of Leipzig, 2006.

[KDKR05] Toralf Kirsten, Hong-Hai Do, Christine Körner, and Erhard Rahm. Hybrid integration of molecular–biological annotation data. In *Proc. of the 2nd International Workshop on Data Integration in the Life Sciences (DILS)*, 2005.

[KR06] Toralf Kirsten and Erhard Rahm. BioFuice: Mapping–based data intergation in bioinformatics. In *Proc. 3rd Intl. Workshop on Data Integration in the Life Sciences (DILS)*, 2006.

[Lim10] LimeSurvey - The Open Source Survey Application. `http://www.limesurvey.org`, Last online access, May, 12, 2010.

[Lin10] Linked Data - Connect Distributed Data across the Web. `http://linkeddata.org`, Last online access, May, 12, 2010.

[LMNR04] Zoé Lacroix, Hyma Murthy, Felix Naumann, and Louiqa Raschid. Links and Paths through Life Sciences data sources. In *Proc. 1st Intl. Workshop on Data Integration in the Life Sciences (DILS), Leipzig*, 2004.

[LRV04] Zoé Lacroix, Louiqa Raschid, and Maria-Ester Vidal. Efficient techniques to explore and rank paths in the life sciences. In *Proc. 1st Intl. Workshop on Data Integration in the Life Sciences (DILS), Leipzig*, 2004.

[SAR+07] Barry Smith, Michael Ashburner, Cornelius Rosse, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*, 25(11):1251–1255, Nov 2007.