Publishing Life Science Data as Linked Open Data: the Case Study of miRBase[*]

Theodore Dalamagas
'Athena' R.C., GR
dalamag@imis.athena-innovation.gr

Nikos Bikakis
NTUA, GR
bikakis@dblab.ece.ntua.gr

George Papastefanatos
'Athena' R.C., GR
gpapas@imis.athena-innovation.gr

Yannis Stavrakas
'Athena' R.C., GR
yannis@imis.athena-innovation.gr

Artemis G. Hatzigeorgiou
'A. Fleming' B.S.R.C., GR
hatzigeorgiou@fleming.gr

ABSTRACT

This paper presents our Linked Open Data (LOD) infrastructures for genomic and experimental data related to microRNA biomolecules. Legacy data from two well-known microRNA databases with experimental data and observations, as well as change and version information about microRNA entities, are fused and exported as LOD. Our LOD server assists biologists to explore biological entities and their evolution, and provides a SPARQL endpoint for applications and services to query historical miRNA data and track changes, their causes and effects.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Scientific Databases

General Terms

Design

1. INTRODUCTION

The technology advances in scientific hardware (sensors, new-generation sequencers, etc.), together with the explosion of Web 2.0 technologies, have completely changed the way scientists create, disseminate and consume large volumes of information and new content. More and more scientific datasets break the walls of private management within their production site, are published, and become available for potential data consumers, i.e., individual users, scientific communities, applications/services. Typical examples include experimental or observational data and scientific models from the life science domain, climate, earth, astronomy, etc.

Linked Data[1] is a compelling approach for the dissemination and re-use of scientific data, realizing the vision of the so-called Linked Science[2]. The Linked Data paradigm involves practices to publish, share, and connect data on the Web, and offers a new way of data integration and interoperability. Briefly, Linked Data is about using the Web to create links between data from different sources. The driving force to implement Linked Data spaces is the RDF technology. The basic principles of the Linked Data paradigm is (a) use the RDF data model to publish structured data on the Web, and (b) use RDF links to interlink data from different data sources. The aim of the Linked Data technologies is to give rise to the Web of Data.

The Web of Data is impelled by the current trend towards an open Web. The open data movement is a significant and emerging force towards this direction. Open science data is open data related to observations and results of scientific activities, which are publicly available for anyone to analyze and reuse.

However, by just converting legacy scientific data as Linked Open Data (LOD), we do not fully meet the requirements of data re-use. To ensure re-use and allow exploitation and validation of scientific results, several challenges related to scientific data dynamics should be tackled. Scientific data are evolving and diverse data. Users and services (a) should have access not only to up-to-date scientific LOD bases but to any of the previous versions of those bases, and (b) should be able to track the changes among versions, as well as their cause and effects.

In this work, we present our LOD services for life science data, and more specifically, genomic and experimental data related to microRNA biomolecules (see Section 2). Legacy data from two well-known microRNA databases are fused and exported as LOD. The first database (see Section 3) provides experimental data and observations, while the second one (see Section 4) provides change and version information about microRNA entities. Our LOD services provide the following facilities:

[*] Funded by COOPERATION 2009 Programme, Praxi I, titled 'MICRORNA - Exploring the role of MicroRNAs in diseases: a computational, experimentally-verified approach', Operational Programmes 'Competitiveness and Entrepreneurship and Regions in Transition', National Strategic Reference Framework NSRF 2007-2013

WOD '12 May 25 2012, Nantes, France
Copyright 2012 ACM 978-1-4503-1404-6/12/05 ...$15.00.

[1] http://linkeddata.org/

[2] http://linkedscience.org/

- Biologists can explore biological entities, their characteristics, and related experimental data with up-to-date information.
- Services and applications can retrieve the same (up-to-date) information as above by using our server as SPARQL endpoint.
- Biologists can retrieve out-of-date resource descriptions, navigate between previous and next versions of the resources, see the changes involved, their causes and their effects on the resources.
- Services and applications can retrieve historic information as above by using our server as SPARQL endpoint.

The system has been built using the D2R LOD infrustructures[3]. All services are available at `http://diwis.imis.athena-innovation.gr/mlod`.

2. RELATED WORK.

The Linked Data paradigm involves practices to publish, share, and connect data on the Web, and offers a new way of data integration and interoperability. Briefly, Linked Data is about using the Web to create typed links between data from different sources. The driving force to implement Linked Data spaces is the RDF technology[4]. The basic principles of the Linked Data paradigm is (a) use the RDF data model to publish structured data on the Web, and (b) use RDF links to interlink data from different data sources. Linked Data technologies have given rise to the Web of Data: a Web of things in the world, described by data on the Web[5]. The Web of Data extents current Web to a global data space connecting data from diverse domains. The Web of Data is impelled by the current trend towards an open Web. The open data movement is a significant and emerging force towards this direction. Open data is public data which are available to people without any restriction. Linked Open Data (LOD) serve a great cause, enabling transparency, accountability and good governance for public administrations.

In the context of LOD, numerous approaches have been proposed to study the problems of evolution, versioning, and change detection. In [18], the term dataset dynamics is coined, essentially addressing content and interlinking changes in linked data sources. In [19], a comparative study on the approaches and tools for detecting, propagating and describing changes in LOD resources and datasets is provided. This survey identifies the following interesting problems for LOD dynamics: change detection at several granularity levels (i.e., at the dataset level, at the triple level, etc), common vocabulary for change description across multiple domains, appropriate communication and notification mechanisms for change propagation and finally automatic change (i.e., broken links) discovery. In [16], the authors deal with changes in the linkage between datasets and specifically with the problem of broken links. They propose, DSNotify, a framework able to assist human and machine actors fixing broken links. A similar approach is the Silk linking framework [20], which is used for discovering and maintaining data links between web data sources. It consists of a link discovery engine, a tool for evaluating the generated links and a protocol for maintaining data links between continuously changing data sources. Regarding versioning and temporal approaches to LOD, in [11] the Memento framework is introduced as a resource versioning mechanism for LOD. It is based on HTTP and handles different versions of linked data by attaching time-specific attributes to HTTP requests. Finally, in [9] they propose linked timelines, a temporal representation and management for LOD. This approach augments URIs with temporal attributes and employs temporal reasoning for resolving URIs validity.

Our approach is specially-tailored to the scientific domain of life science data, and more specifically to genomic and experimental data related to microRNA biomolecules. Several attempts have been recently made to provide scientific LOD services. W3C has established the Semantic Web Health Care and Life Sciences Interest Group (HCLS)[6], aiming to exploit Semantic Web technologies for the management and the representation of biological, medicine and health care data. The HCLS group works on Linking Open Drug Data (LODD) project which provides linked RDF data exported from several data sources like ClinicialTriasl.gov, DrugBank, DailyMed, etc. Additionally, Bio2RDF[7] provides linked RDF data produced from over 30 biological data sources. Some earlier efforts include YeastHub [8], LinkHub [17], BioDash [15] and BioGateway[8]. Finally, Chem2Bio2RDF [7] integrates chemical and biological information. Also, several chemogenomics repositories have been transformed into RDF and linked to Bio2RDF and LODD RDF resources.

3. BACKGROUND

Biologists used to consider proteins and DNA as movers and shakers in genomics, seeing RNA as nothing more than a messenger to carry information between the two. This has dramatically changed after the discovery, in early 2000s, of the key role played in gene expression by small RNA molecules, called *microRNAs* (miRNAs). miRNAs can completely silence proteins. They do so by binding themselves to complementary sequences on message RNA (mRNA) transcripts, called *targets*. The knowledge of *miRNA targets* (i.e., which mRNA transcripts are targeted by a miRNA) is important for therapeutic uses. For example, based on such knowledge, biologists can shut off genes by delivering artificial miRNA molecules into cells.

The first miRNA molecules were identified in 1993. Since then, there has been a dramatic increase in the number of miRNAs discovered and registered in *miRBase*[9], a searchable database of published miRNA sequences and annotation. However, there is a lack of high-throughput experimental methods for identifying miRNA targets. Thus, *computational methods* to predict targets have become increasingly important, and led to the experimental identification of many miRNA targets.

Our team in IMIS/Athena R.C. and the DNA Intelligent Analysis (DIANA) group of Alexander Fleming B.S.R.C.[10] have developed a set of advanced Web applications to provide access to computationally predicted miRNA targets. Since its original launch, DIANA Web app has been one of

[3] http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/

[4] http://www.w3.org/RDF/

[5] http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf

[6] http://www.w3.org/blog/hcls/

[7] http://bio2rdf.org/

[8] http://www.semantic-systems-biology.org/biogateway

[9] http://www.mirbase.org/

[10] http://www.fleming.gr/

the most widely used service for miRNA analysis. It includes the following two core services.

microT[11]. The service provides target prediction data for 1884 miRNAs and more than six million predicted target genes, organized in a relational database. Besides the target prediction experimental results, we provide miRNAs and genes functional analysis that goes beyond simple biological pathways, like, for example, relation of miRNAs to functional features, and diseases and medical descriptors. All retrieved miRNAs are associated to diseases, using textual information from PubMed[12], a well-known digital library for biomedical literature.

mirGen[13]. The service provides information about transcripts, and their transcription factors (TF) that correspond to miRNAs. A transcription factor is a protein that binds to specific DNA sequences, thereby controlling the flow of genetic information from DNA to mRNA. MirGen database stores information about 811 human genes, 1270 human miRNAs, 386 mouse genes and 1012 mouse miRNAs, organized in a relational database.

4. DATABASE OVERVIEW

Next, we present an overview of the miRNA database maintained by our team in IMIS/Athena R.C. and the DIANA group of A. Fleming B.S.R.C., storing info about computationally predicted miRNA targets produced by the target prediction algorithm proposed by DIANA group[14].

To better understand the miRNA domain and the DB schema design, we next clarify some issues. Since the term "miRNA" is nowadays used in a wide scope, it is common to distinguish between `hairpin miRNAs` and `mature miRNAs`, or just `hairpins` and `matures` from now on. The former signifies the genomic location of the latter. A hairpin is actually processed into several matures. Matures bind themselves to transcripts and prevent the creation of functional ribosomes (and, thus, prohibit protein construction). A transcript is a stretch of DNA transcribed into an RNA molecule (messenger RNA, ribosomal RNA, transfer RNA, etc). Figure 1 illustrates miRNA domain and functions.

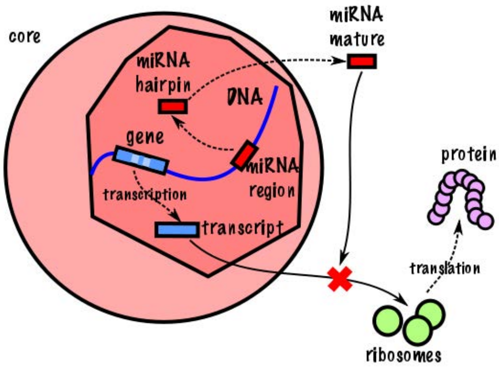


Figure 1: The miRNA domain.

The miRNA database has some core tables to store the key entities of the miRNA domain (hairpins, matures, transcripts and protein-encoding genes) and model their relationships (see Table 1 for a part of miRNA database schema).

Core tables	**Column Description**
Hairpins	id (`mima_id`), name, sequence, species, gene location info, etc.
Matures	id (`mimat`), name, sequence, species.
Transcripts	tid, id given from ensembl.org (`enstid`), species, DNA strand, gene location info, etc.
ProteinGenes	id given from ensembl.org (`ensgid`), name, description.
Keggs	id given from genome.jp (`kegg_id`), name.
Tissues	name, species.
Join Tables	**Column Description**
MatureHairpinConn	It relates matures and hairpins.
MicroT5Interactions	It contains all the experimentally verified gene-mature interactions (bindings).
ProteinGeneKeggConn	It relates genes to kegg pathways.
MatureTissueConn	It relates matures to tissues.

Table 1: Part of miRNA database schema.

There are also tables storing info about `Kegg pathways`[14] and `tissues`. Kegg pathways is a collection of manually drawn pathway maps, with textual descriptions, representing biologists knowledge on molecular interaction and reaction networks.

5. CHANGE AND VERSION MANAGEMENT

The miRBase database is a searchable database of published miRNA sequences and annotation. The miRBase database maintains info for 18443 hairpins and 49670 matures. Each entry in miRBase represents a predicted hairpin miRNA with information on the location and sequence of the corresponding mature miRNA sequence. Hairpins, mature miRNAs and their relationship between them change in time. miRBase maintains a list of files that record successive versions along with the changes between them. A short description for each file follows.

- **miRNA.dat** It maintains info related to all known hairpins (like ID, name, related matures, related publications, sequence, etc.) at the time of each version. Every new version of miRNA.dat contributes to the previous one with all the newly discovered miRNAs, omitting the deleted ones. Example entries are shown in Figure 2 for the hairpin with name `cel-let-7` and id (i.e., key) `MI0000001`. Note that info presented includes publications where this hairpin is referenced, text comments, a list of related matures, and its sequence. For instance, `cel-let-7` is related to two mature miRNAs, namely `MIMAT0000001` and `MIMAT0000001`[15].
- **miRNA.diff** It tracks change operations on hairpins and matures. Each version of miRNA.diff refers to a certain time period and tracks changes only for that period. Example entries of miRNA.diff are shown in Figure 2. For instance, `MI0000001 cel-let-7 NEW` means that the hairpin with ID MI0000001 and name cel-let-7 is created. Also, `MI0004476 mdv2-miR-M29-5p SEQUENCE NAME` means that the hairpin with ID MI0004476 has changed its name (to mdv2-miR-M29-5p) and its

[11] http://diana.cslab.ece.ntua.gr/DianaTools /index.php?r=microtv4

[12] http://www.ncbi.nlm.nih.gov/pubmed/

[13] http://diana.cslab.ece.ntua.gr/?sec=databases

[14] http://www.genome.jp/kegg/pathway.html

[15] Note that IDs starting with "MIMA" refer to matures.

sequence. Note that to find the old name and the old sequence, we should refer to the older version of the miRNA.dat file, where hairpin names and info about sequences are available. Similarly, `MIMAT0000115 dme-miR-10* SEQUENCE NAME` means that the mature with ID MIMAT0000115 has changed its name (to dme-miR-10*) and its sequence.

- **miRNA.dead** It keeps all deleted hairpins at the time of a version. It is maintained incrementally. Deletion means either getting rid of a hairpin (e.g., incorrectly characterized in previous versions) or replacing a hairpin with another one. For the latter case, links to existing hairpins are provided. Contrary to deleted hairpins, deleted mature miRNAs are not stored in miRNA.dead file. Example entries of miRNA.dead are shown in Figure 2. For instance, the hairpin with ID hsa-mir-101-9 and NAME MI0000104 has been deleted. The reason is that it was a duplicate entry (see the comment in CC field). There is a hairpin (MI0000739), though, that replaces the deleted one (see the FW field).
- **miFam.dat** It stores info about hairpin families at the time of a version. Hairpins that produce similar mature miRNAs belong to the same family. It is maintained incrementally. Example entries of miFam.dat are shown in Figure 2. For instance, hairpins with IDs MI0011482 (NAME bta-mir-677) and MI0004634 (NAME mmu-mir-677) belong to the same family with id mir-677.

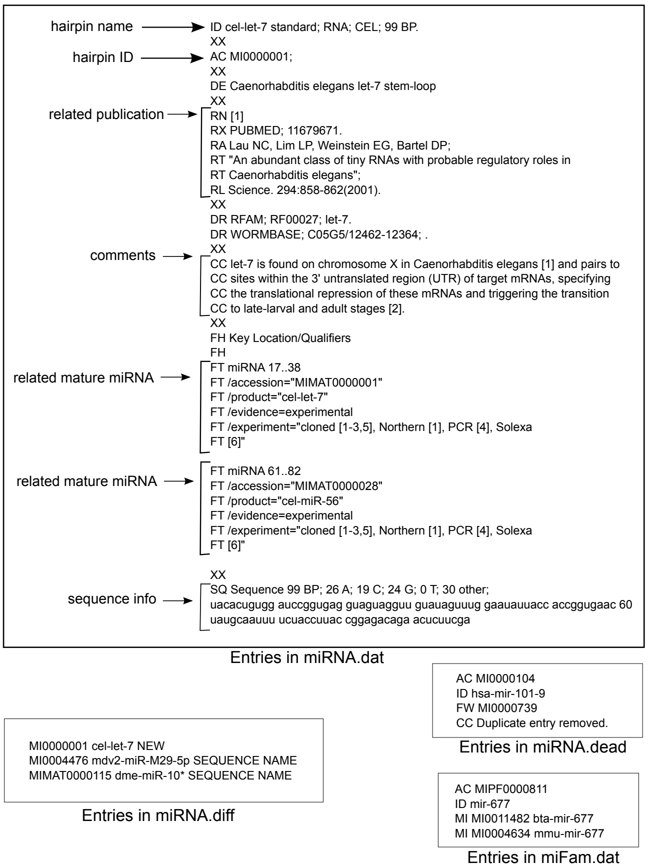
We have examined all files and recorded the following types of changes for hairpins:

- NEW: a new hairpin is created.
- NAME: a hairpin changes its name.
- SEQUENCE (SEQ): a hairpin changes its sequence.
- NAME/SEQUENCE (NS): a hairpin changes both its name and sequence at the same time.
- FORWARD (FW): a hairpin is deleted, but miRBase give a link to another hairpin for replacement.
- DELETE (DEL): a hairpin is deleted (no replacement).

Similarly, we have identified the following type of changes for matures:

- NEW: a new mature is created.
- NAME: a mature changes its name.
- SEQUENCE (SEQ): a mature changes its sequence.
- NAME/SEQUENCE (NS): a mature changes both its name and sequence at the same time.
- ADD PARENT HAIRPIN (APH): a new hairpin is added to the list of hairpins that produces a mature.
- REMOVE PARENT HAIRPIN (RPH): a hairpin is removed from the list of hairpins that produces a mature.
- DELETE (DEL): a mature is deleted.

To manage change and version info, we maintain two history tables: HairpinsHistory and MaturesHistory. Tables 2 and 3 show how change and version info is maintained in history tables. For each hairpin change, HairpinsHistory keeps a record with, the hairpin id, the type of change, hairpin name, the version number where the change occurred (column `first_appearance`), and the version number where the next change occurs (column `last_appearance`). Note that the last two columns actually track time ranges trhoughout the hairpin remanes unchanged.


hairpin name
hairpin ID
related publication
comments
related mature miRNA
related mature miRNA
sequence info

```
ID cel-let-7 standard; RNA; CEL; 99 BP.
XX
AC MI0000001;
XX
DE Caenorhabditis elegans let-7 stem-loop
XX
RN [1]
RX PUBMED; 11679671.
RA Lau NC, Lim LP, Weinstein EG, Bartel DP;
RT "An abundant class of tiny RNAs with probable regulatory roles in
RT Caenorhabditis elegans";
RL Science. 294:858-862(2001).
XX
DR RFAM; RF00027; let-7.
DR WORMBASE; C05G5/12462-12364; .
XX
CC let-7 is found on chromosome X in Caenorhabditis elegans [1] and pairs to
CC sites within the 3' untranslated region (UTR) of target mRNAs, specifying
CC the translational repression of these mRNAs and triggering the transition
CC to late-larval and adult stages [2].
XX
FH Key Location/Qualifiers
FH
FT miRNA 17..38
FT /accession="MIMAT0000001"
FT /product="cel-let-7"
FT /evidence=experimental
FT /experiment="cloned [1-3,5], Northern [1], PCR [4], Solexa
FT [6]"

FT miRNA 61..82
FT /accession="MIMAT0000028"
FT /product="cel-miR-56"
FT /evidence=experimental
FT /experiment="cloned [1-3,5], Northern [1], PCR [4], Solexa
FT [6]"

XX
SQ Sequence 99 BP; 26 A; 19 C; 24 G; 0 T; 30 other;
uacacugugg auccggugag guaguagguu guauaguuug gaauauuacc accggugaac 60
uaugcaauuu ucuaccuuac cggagacaga acucuucga
```

Entries in miRNA.dat

```
AC MI0000104
ID hsa-mir-101-9
FW MI0000739
CC Duplicate entry removed.
```

Entries in miRNA.dead

```
MI0000001 cel-let-7 NEW
MI0004476 mdv2-miR-M29-5p SEQUENCE NAME
MIMAT0000115 dme-miR-10* SEQUENCE NAME
```

Entries in miRNA.diff

```
AC MIPF0000811
ID mir-677
MI MI0011482 bta-mir-677
MI MI0004634 mmu-mir-677
```

Entries in miFam.dat

Figure 2: File examples of tracking miRNA changes.

Consider for example the hairpin with id . . . 1364. It was first created in version 13. In version 16, it changes name from `dre-mir-10b` to `dre-mir-10b-1`. No other change has occurred till version 18, where a change in its sequence has occurred. Another sequence change has occurred in version 20.

Table MaturesHistory has a similar structure. Consider for example the mature with id . . . 9477. It was first created in version 28, getting the name `bfl-miR-79`, and having the parent hairpin . . . 021. In version 30, it changes name (to bfl-miR-9-3p) and sequence.

6. PUBLISHING LOD MIRNA DATA

6.1 LOD technology adopted

To publish miRNA and miRBase databases as LOD, we adopted the "virtual RDF" approach: accessing a non-RDF database using an RDF view. Such an approach enables the access of non-RDF, legacy databases without having to replicate the whole database into RDF. The D2R server [5] is a popular tool that follows the "virtual RDF" approach for publishing the content of relational databases on the Semantic Web. Database content is mapped to RDF using the D2RQ declarative mapping language that captures mappings between database schemas and RDFS/OWL schemas.

A D2RQ mapping specifies how RDF resources are iden-

mima-id	change	name	seq	first_app-earance	last_app-earance
...1364	NEW	dre-mir-10b	...X...	13	15
...1364	NAME	dre-mir-10b-1	...X...	16	17
...1364	SEQ	dre-mir-10b-1	...Y...	18	19
...1364	SEQ	dre-mir-10b-1	...Z...	20	32

Table 2: Table HairpinsHistory: record samples.

mimat	change	name	seq	par. hair-pin	first_app-earance	last_app-earance
...9477	NEW	bfl-miR-79	.X.	...	28	...
...9477	APH	bfl-miR-79	.X.	...021	28	29
...9477	NS	bfl-miR-9-3p	.Y.	...	30	32

Table 3: Table MaturesHistory: record samples.

tified and how RDF property values are generated from database content. Mappings in D2RQ are declared based on *ClassMaps* and *PropertyBridges*. A ClassMap maps a set of database records to an RDF class of resources. Resources are assigned URIs using URI patterns. The pattern `hairpins/`@@`diana_hairpins.mima_id`@@, for instance, produces a relative URI like `hairpins/MI0000005` by inserting a value from the column `mima_id` of table `hairpins` of miRNA database into the pattern. The D2R Server turns relative URIs into absolute URIs by expanding them with the servers base URI. If a database already contains URIs for identifying database content, then these external URIs can be used instead of pattern-generated URIs. The following ClassMap definition creates the class of hairpin resources, and assigns them URIs using their ids from the miRNA database:

```
map:Hairpins a d2rq:ClassMap;
  d2rq:dataStorage map:database;
  d2rq:uriPattern "hairpins/@@diana_hairpins.mima_id@@";
  d2rq:class diana:Hairpin;
  d2rq:classDefinitionLabel "Hairpin";
```

Each ClassMap has a set of PropertyBridges which specify how the properties of an RDF instance are created. Property values can be literals, URIs or blank nodes, and can be created directly from database values or by employing patterns. The following PropertyBridge definition creates the property `diana:name`. Values for that property are created from the `name` column of table `diana_haipins`:

```
map:diana_hairpins_name a d2rq:PropertyBridge;
  d2rq:belongsToClassMap map:Hairpins;
  d2rq:property diana:name;
  d2rq:propertyDefinitionLabel "Hairpins name";
  d2rq:column "diana_hairpins.name";
```

Note that D2R provides flexible mappings of complex relational structures, allowing SQL statements directly in the mapping rules. The resulting record sets are grouped afterwards and the data is mapped to the created instances.

We used D2R as a full-fledge Linked Open Data server. The size of the LOD base is around 100Million triples.

6.2 LOD publishing

The miRNA LOD schema has been designed around four core classes: `Hairpin`, `Mature`, `ProteinGene` and `Transcript` (defined as ClassMap entities in D2R - see previous subsection). Figure 3 shows an overview of the schema adopted and part of the mappings used to publish miRNA data as LOD.

Consider, for example, the class `Mature`. Resources of that class are assigned URIs of the form:
`http://.../resource/matures/[Matures.mimat]`,
where `Matures.mimat` gets values from column `mimat` of Table `Matures`. Some of the class properties are: `name`, `species`, `relatedKegg`, `targetsProteinGene`. These properties are defined as PropertyBridge entities in D2R (see previous subsection). Consider also the property `targetProteinGene` that relates matures with genes (targets). Note that `ProteinGene` resources are assigned URIs of the form:
`http://.../resource/proteingenes/[ProteinGenes.ensgid]`,
where `ProteinGenes.ensgid` gets values from column `ensgid` of Table `ProteinGenes`.

For a given `Mature` URI, to calculate the URIs of related `ProteinGene` resources, the mapping definition should include the following join:
`Matures.mimat=MicroT5Interactions.mimat AND`
`MicroT5Interactions.tid=Transcripts.tid AND`
`Transcripts.enstid=ProteinGenes.enstid.`

To link our LOD to the LOD cloud, we provide `owl:sameAs` links to appropriate biological LOD infrastructures. See, for example, the BIO2RDF[16] data source that provide RDF descriptions for transcripts, tissues, keggs, and species.

Figure 4 illustrates resource descriptions examples for mature miRNA `MIMAT0000001` and kegg `cel00010`.

6.2.1 *Change and version management*

One of the major research problems in LOD publishing is how to deal with linked data that changes over time. While handling changes for information resources is rather straightforward, handling changes for non-information resources is a challenging issue. Key requirements for dealing with changes in miRNA LOD are the following:

- Biologists that care only about the current state of data should be able to browse or query the miRNA LOD base easily to get up-to-date data. Also, up-to-date data should be easily retrieved using SPARQL.
- Biologists should be able to query historic miRNA data, and navigate through versions. Also, miRNA changes should be treated as first-class citizens so that one can form SPARQL queries that involve change resources, and trace those changes and their effects.

6.2.2 *Browsing and querying up-to-date miRNA data.*

Using the D2R browsing facilities, biologists can navigate through the miRNA LOD base, exploring hairpin, mature, gene or transcript resources and their descriptions. All data provided refer to the current version of miRNA database. Also, any resource URI refers to the current version of that resource. This is ensured because all triples involving resources from `Hairpin`, `Mature`, `ProteinGene` and `Transcript` classes are populated from the core and join tables of Table 1 that are up-to-dated.

Using the D2R SPARQL end-point facilities, biologists can pose SPARQL queries to the miRNA LOD. Whenever a resource URI is used in a query, it refers to the current version of that resource. To get up-to-date results, a property should be used to avoid the retrieval of out-of-date triples. For example, the following SPARQL query retrieves all hair-

[16] http://bio2rdf.org

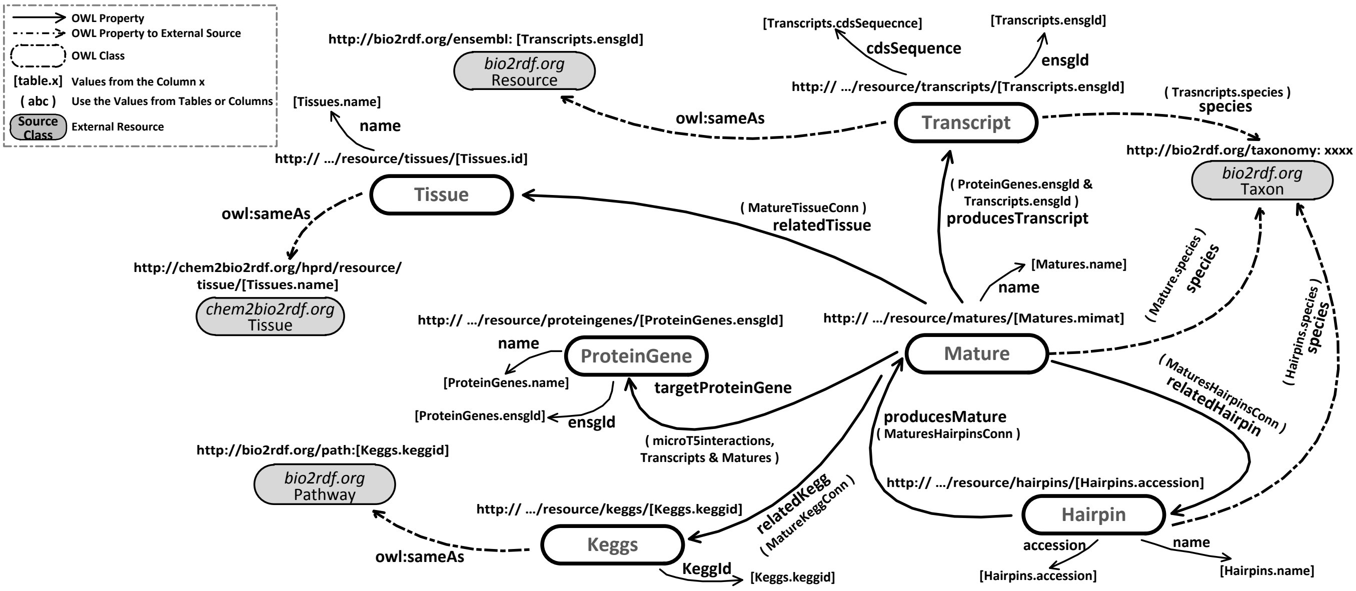
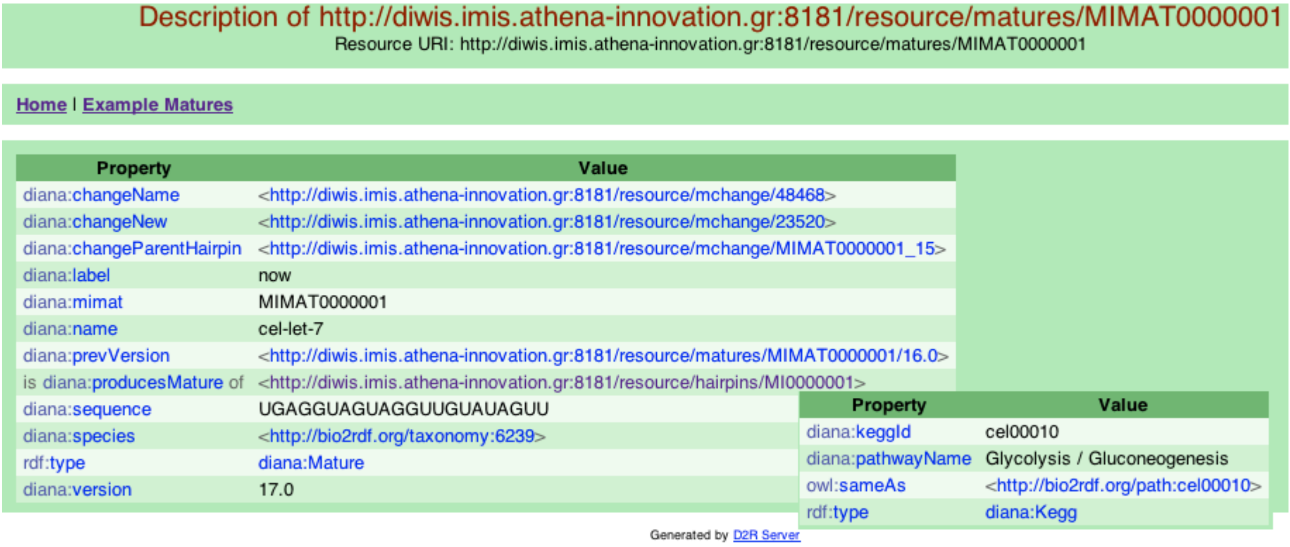

OWL Property
OWL Property to External Source
OWL Class
[table.x] Values from the Column x
(abc) Use the Values from Tables or Columns
Source Class
External Resource
http://bio2rdf.org/ensembl: [Transcripts.ensgId]
bio2rdf.org Resource
[Transcripts.cdsSequecnce]
cdsSequence
[Transcripts.ensgId]
ensgId
http:// …/resource/transcripts/[Transcripts.ensgId]
owl:sameAs
Transcript
(Trasncripts.species)
species
http://bio2rdf.org/taxonomy: xxxx
bio2rdf.org Taxon
[Tissues.name]
name
http:// …/resource/tissues/[Tissues.id]
Tissue
owl:sameAs
http://chem2bio2rdf.org/hprd/resource/ tissue/[Tissues.name]
chem2bio2rdf.org Tissue
(MatureTissueConn)
relatedTissue
(ProteinGenes.ensgId & Transcripts.ensgId)
producesTranscript
[Matures.name]
name
(Mature.species)
species
(Hairpins.species)
species
http:// …/resource/proteingenes/[ProteinGenes.ensgId]
name
ProteinGene
[ProteinGenes.name]
[ProteinGenes.ensgId]
ensgId
targetProteinGene
(microT5interactions, Transcripts & Matures)
http:// …/resource/matures/[Matures.mimat]
Mature
(MaturesHairpinsConn)
relatedHairpin
producesMature
(MaturesHairpinsConn)
http://bio2rdf.org/path:[Keggs.keggid]
bio2rdf.org Pathway
http:// …/resource/keggs/[Keggs.keggid]
relatedKegg
(MatureKeggConn)
http:// …/resource/hairpins/[Hairpins.accession]
owl:sameAs
Keggs
KeggId
[Keggs.keggid]
Hairpin
accession
[Hairpins.accession]
name
[Hairpins.name]

Figure 3: Publishing miRNA data as LOD data: RDFS to database mappings (up-to-date data).



Description of http://diwis.imis.athena-innovation.gr:8181/resource/matures/MIMAT0000001

Resource URI: http://diwis.imis.athena-innovation.gr:8181/resource/matures/MIMAT0000001

Home | Example Matures

Property	Value
diana:changeName	<http://diwis.imis.athena-innovation.gr:8181/resource/mchange/48468>
diana:changeNew	<http://diwis.imis.athena-innovation.gr:8181/resource/mchange/23520>
diana:changeParentHairpin	<http://diwis.imis.athena-innovation.gr:8181/resource/mchange/MIMAT0000001_15>
diana:label	now
diana:mimat	MIMAT0000001
diana:name	cel-let-7
diana:prevVersion	<http://diwis.imis.athena-innovation.gr:8181/resource/matures/MIMAT0000001/16.0>
is diana:producesMature of	<http://diwis.imis.athena-innovation.gr:8181/resource/hairpins/MI0000001>
diana:sequence	UGAGGUAGUAGGUUGUAUAGUU
diana:species	<http://bio2rdf.org/taxonomy:6239>
rdf:type	diana:Mature
diana:version	17.0

Property	Value
diana:keggId	cel00010
diana:pathwayName	Glycolysis / Gluconeogenesis
owl:sameAs	<http://bio2rdf.org/path:cel00010>
rdf:type	diana:Kegg

Generated by D2R Server

Figure 4: Examples of miRNA and kegg RDF resource descriptions.

pins, and their sequences, that are located in chromosome X from the current version of miRNA LOD:

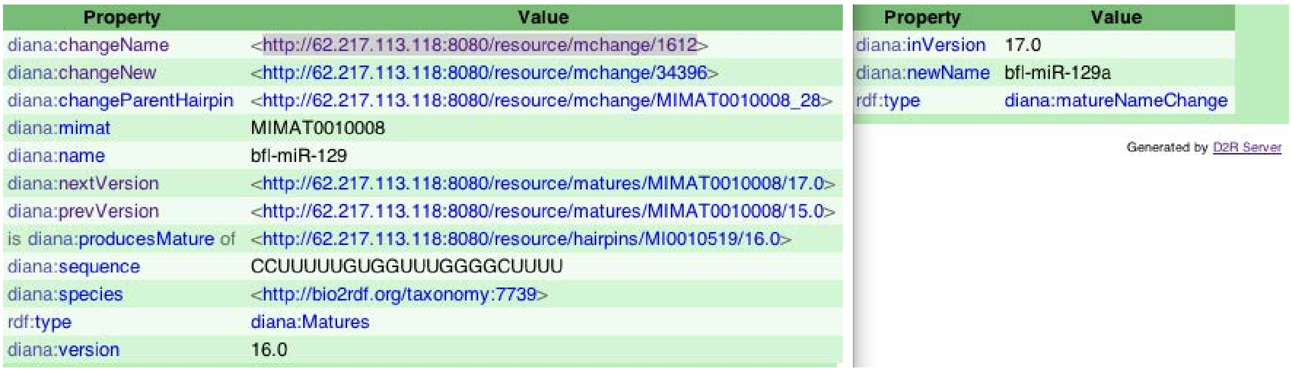
```
SELECT ?h ?s WHERE {
  ?h rdf:type diana:Hairpin.
  ?h diana:sequence ?s.
  ?h diana:chromosome "X".
  ?h diana:label "now". }
```

Note that the use of label property with value `now` indicates that we are interested in results based on the current (i.e., up-to-date) version of miRNA LOD. If we remove that property, we will get back results for resource descriptions for all available versions. E.g., for hairpin `MI0000027`, we will get all instances of its sequence based on version 11.0, 12.0, 13.0, etc, of miRBase.

6.2.3 Browsing and querying historic miRNA data.

Out-of-date resource descriptions are retrieved using the following URI pattern: `URI/{version number}`. E.g., URI `http://.../resource/hairpins/MI0000044/8.0` returns the RDF description of hairpin `MI0000044` in version 8.0 of miRBase. To pose the previous SPARQL query on that version of miRBase, one should replace `?h diana:label "now".` with `?h diana:version "8.0".` Note that we provide properties (`diana:nextVersion`, `diana:prevVersion`) to move to the next and the previous version of a resource description.

To be able to provide the property values and URIs which are valid at a certain version, we exploit the version infor-

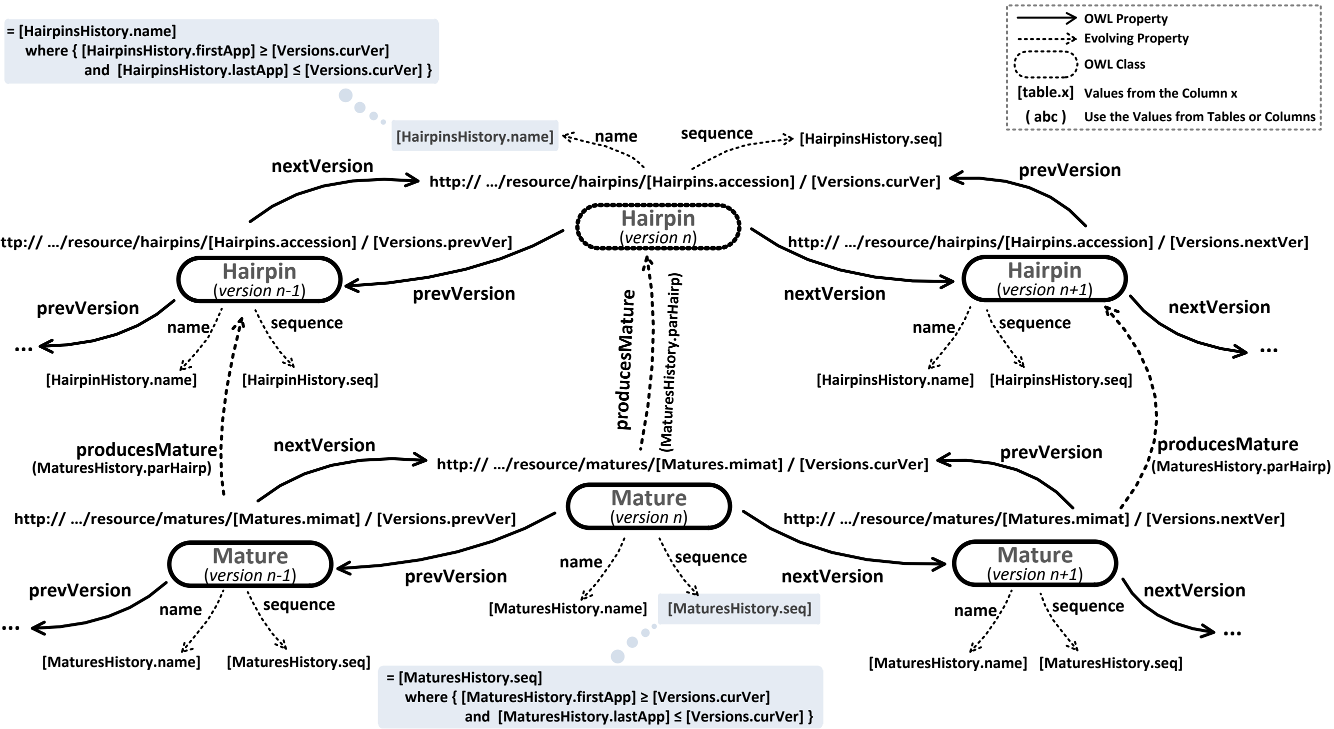


Property	Value
diana:changeName	<http://62.217.113.118:8080/resource/mchange/1612>
diana:changeNew	<http://62.217.113.118:8080/resource/mchange/34396>
diana:changeParentHairpin	<http://62.217.113.118:8080/resource/mchange/MIMAT0010008_28>
diana:mimat	MIMAT0010008
diana:name	bfl-miR-129
diana:nextVersion	<http://62.217.113.118:8080/resource/matures/MIMAT0010008/17.0>
diana:prevVersion	<http://62.217.113.118:8080/resource/matures/MIMAT0010008/15.0>
is diana:producesMature of	<http://62.217.113.118:8080/resource/hairpins/MI0010519/16.0>
diana:sequence	CCUUUUUGUGGUUUGGGGCUUUU
diana:species	<http://bio2rdf.org/taxonomy:7739>
rdf:type	diana:Matures
diana:version	16.0

Property	Value
diana:inVersion	17.0
diana:newName	bfl-miR-129a
rdf:type	diana:matureNameChange

Generated by D2R Server

Figure 5: Resource description of mature MIMAT0010008 at version 16.0



= [HairpinsHistory.name]
where { [HairpinsHistory.firstApp] ≥ [Versions.curVer]
and [HairpinsHistory.lastApp] ≤ [Versions.curVer] }
OWL Property
Evolving Property
OWL Class
[table.x] Values from the Column x
(abc) Use the Values from Tables or Columns
[HairpinsHistory.name]
name
sequence
[HairpinsHistory.seq]
nextVersion
http:// .../resource/hairpins/[Hairpins.accession] / [Versions.curVer]
prevVersion
http:// .../resource/hairpins/[Hairpins.accession] / [Versions.prevVer]
Hairpin
(version n)
http:// .../resource/hairpins/[Hairpins.accession] / [Versions.nextVer]
Hairpin
(version n-1)
prevVersion
nextVersion
Hairpin
(version n+1)
prevVersion
nextVersion
...
name
sequence
[HairpinHistory.name]
[HairpinHistory.seq]
producesMature
(MaturesHistory.parHairp)
name
sequence
[HairpinsHistory.name]
[HairpinsHistory.seq]
...
producesMature
(MaturesHistory.parHairp)
nextVersion
http:// .../resource/matures/[Matures.mimat] / [Versions.curVer]
prevVersion
producesMature
(MaturesHistory.parHairp)
http:// .../resource/matures/[Matures.mimat] / [Versions.prevVer]
Mature
(version n)
http:// .../resource/matures/[Matures.mimat] / [Versions.nextVer]
Mature
(version n-1)
prevVersion
name
sequence
nextVersion
Mature
(version n+1)
prevVersion
nextVersion
...
name
sequence
[MaturesHistory.name]
[MaturesHistory.seq]
[MaturesHistory.name]
[MaturesHistory.seq]
name
sequence
[MaturesHistory.name]
[MaturesHistory.seq]
...
= [MaturesHistory.seq]
where { [MaturesHistory.firstApp] ≥ [Versions.curVer]
and [MaturesHistory.lastApp] ≤ [Versions.curVer] }

Figure 6: Publishing miRNA data as LOD data: RDFS to database mappings (historic data).

mation present in the history tables HairpinsHistory and MaturesHistory (see Tables 2 and 3). Figure 6 shows an overview of the schema adopted and part of the mappings used to manage changes and versions. For example, for given version `curVer`, to find the valid value for the `name` property of a hairpin for that version, we should focus on name values that remain unchanged for a time period that starts before `curVer` and ends after `curVer`. This is ensured by using the following join condition (as shown in Figure 6):
`HaipinsHistory.firstApp≥Versions.curVer` AND
`HaipinsHistory.lastApp≤Versions.curVer`

Consider for example hairpin `MI0000001`. It was created in version 1.0, using the name `cel-let-7L` which remained unchanged till version 1.5. However, in version 2.0, the name was changed to `cel-let-7` and remained unchanged till version 17.0. Asking for hairpin's name in version 1.3 requires that we retrieve the name value from the record of Table 2 having versions in columns `firstApp` and `lastApp` earlier than 1.3 and later than 1.3, respectively. A similar approach is used for mature miRNAs (see e.g. the join condition to identify the sequence of a mature miRNA given a version).

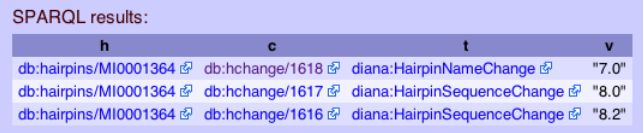
Each hairpin or mature resource description includes properties that capture the changes which those resources are affected by. For each change, we track its effect and its cause. Figure 5 shows the description of mature MIMAT0010008 at version 16. The following SPARQL query retrieves 10 hairpins that where deleted or replaced in version in version 1.3. of miRBase, and the URIs of the change operations:

```
SELECT ?h ?d ?c WHERE {
 ?h rdf:type diana:Hairpin.
 {{?h diana:changeDelete ?d.} UNION
 {?h diana:changeForward ?c.}}
 ?h diana:version "1.3". } LIMIT 10
```

We can also retrieve historical info about change occurrences. The following SPARQL query returns name or sequence changes that happened on hairpin `MI0001364`:

```
SELECT ?h ?c ?v WHERE {
 ?h rdf:type diana:Hairpin.
 ?h diana:accession "MI0001364".
 {{?h diana:changeName ?c.
   ?c diana:inVersion ?v.} UNION
  {?h diana:changeSequence ?c.
   ?c diana:inVersion ?v.}}}
```

Resuls are shown in Figure 7. Compare the results with those in Table 2. Note that, in our system, values 16 (first appearance of NAME change), 18 (first appearance of SEQ change) and 20 (first appearance of a second SEQ change) represent mirBase versions 7.0, 8.0 and 8.2, respectively.



SPARQL results:

h	c	t	v
db:hairpins/MI0001364	db:hchange/1618	diana:HairpinNameChange	"7.0"
db:hairpins/MI0001364	db:hchange/1617	diana:HairpinSequenceChange	"8.0"
db:hairpins/MI0001364	db:hchange/1616	diana:HairpinSequenceChange	"8.2"

Figure 7: Example of retrieving change occurrences.

7. CONCLUSION AND FURTHER WORK

In this work we presented a case study of publishing genomic and experimental data related to microRNA biomolecules as Linked Open Data. Legacy data from two well-known microRNA databases with experimental data and observations, as well as change and version information about microRNA entities, are fused and exported as LOD. The miRNA LOD server assists biologists to explore biological entities, and navigate between previous/next versions of the resources, and also provides a SPARQL endpoint for applications to query historical miRNA data and track changes.

As future work, we will work more on the database and RDF schema design, as well as on RDF-to-database mappings to provide SPARQL-friendly querying capabilities. Our aim is to simplify the syntax of queries that are related to change tracking. We also plan to expand the LOD set with resources available from gene databanks, and also to implement materialized approaches (i.e., using a native RDF store).

8. REFERENCES

[1] E. Antezana, W. Blondé, M. Egaña, A. Rutherford, R. Stevens, B. D. Baets, V. Mironov, and M. Kuiper. Biogateway: a semantic systems biology tool for the life sciences. *BMC Bioinformatics*, 10(S-10), 2009.

[2] D. Ayers and M. Völkel. Cool uris for the semantic web. http://www.w3.org/TR/cooluris, 2008.

[3] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. volume 41, pages 706–716, 2008.

[4] T. Berners-Lee. Linked data - design issues. http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[5] C. Bizer and R. Cyganiak. D2r server – publishing relational databases on the semantic web. In *Proceedings of the ISWC'06 Conference*, 2006.

[6] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[7] B. Chen, X. Dong, D. Jiao, H. Wang, Q. Zhu, Y. Ding, and D. J. Wild. Chem2bio2rdf: a semantic framework for linking and data mining chemogenomic and systems chemical biology data. *BMC Bioinformatics*, 11:255, 2010.

[8] K.-H. Cheung, K. Y. Yip, A. K. Smith, R. de Knikker, A. Masiar, and M. Gerstein. Yeasthub: a semantic web use case for integrating data in life sciences domain. In *ISMB (Supplement of Bioinformatics)*, 2005.

[9] G. Correndo, M. Salvadores, I. Millard, and N. Shadbolt. Linked timelines: Temporal representation and management in linked data. In *Proceedings of the COLD'10 Workshop*, 2010.

[10] C. T. O. Council. Designing uri sets for the uk public sector. http://www.cabinetoffice.gov.uk/sites/default/files/-resources/designing-URI-sets-uk-public-sector.pdf, 2009.

[11] H. V. de Sompel, R. Sanderson, M. L. Nelson, L. Balakireva, H. Shankar, and S. Ainsworth. An http-based versioning mechanism for linked data. In *Proceedings of the LDOW'10 Workshop*, 2010.

[12] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space.* Morgan & Claypool, 2011.

[13] A. Jentzsch, J. Zhao, O. Hassanzadeh, K. H. Cheung, M. Samwald, and B. Andersson. Linking open drug data. In *I-SEMANTICS*, 2009.

[14] M. Maragkakis, M. Reczko, V. A. Simossis, P. Alexiou, G. L. Papadopoulos, T. Dalamagas, G. Giannopoulos, G. Goumas, E. Koukis, K. Kourtis, T. Vergoulis, N. Koziris, T. Sellis, P. Tsanakas, and A. G. Hatzigeorgiou. Diana-microt web server: elucidating microrna functions through target prediction. *Nucleic Acids Research*, 37(suppl 2):W273–W276, 2009.

[15] E. K. Neumann and D. Quan. Biodash: A semantic web dashboard for drug development. In *Pacific Symposium on Biocomputing*, 2006.

[16] N. Popitsch and B. Haslhofer. Dsnotify: handling broken links in the web of data. In *Proceedings of the WWW'10 Conference*, pages 761–770, 2010.

[17] A. K. Smith, K.-H. Cheung, K. Y. Yip, M. H. Schultz, and M. Gerstein. Linkhub: a semantic web system that facilitates cross-database queries and information retrieval in proteomics. *BMC Bioinformatics*, 8(S-3), 2007.

[18] J. Umbrich, M. Hausenblas, A. Hogan, A. Polleres, and S. Decker. Towards dataset dynamics: Change frequency of linked open data sources. In *Proceedings of the LDOW'10 Workshop*, 2010.

[19] J. Umbrich, B. Villazo'n-Terrazas, and M. Hausenblas. Dataset dynamics compendium: A comparative study. In *First International Workshop on Consuming Linked Data (COLD2010)*, 2010.

[20] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *Proceedings of the ISWC'09 Conference*, 2009.