BJMG 14/2 (2011) 49-64
10.2478/v10034-011-0047-7

ORIGINAL ARTICLE

METADATA MANAGEMENT AND SEMANTICS IN MICROARRAY REPOSITORIES

Kocabaş F[1,2,*], Can T[3], Baykal N[1]

***Corresponding Author:** Fahri Kocabaş, NATO HQ C3S, Blvd Leopold III B, 1110 Brussels, Belgium; Tel.: +32-2-707-5533; Fax: +32-2-707-5834; E-mail: FK:fahri@ii.metu.edu.tr; f.kocabas@hq.nato.int

ABSTRACT

The number of microarray and other high-throughput experiments on primary repositories keeps increasing as do the size and complexity of the results in response to biomedical investigations. Initiatives have been started on standardization of content, object model, exchange format and ontology. However, there are backlogs and inability to exchange data between microarray repositories, which indicate that there is a great need for a standard format and data management.

We have introduced a metadata framework that includes a metadata card and semantic nets that make experimental results visible, understandable and usable. These are encoded in syntax encoding schemes and represented in RDF (Resource Description Frameword), can be integrated with other metadata cards and semantic nets, and can be exchanged, shared and queried. We demonstrated the performance and potential benefits through a case study on a selected microarray repository. We concluded that the backlogs can be reduced and that exchange of information and asking of knowledge discovery questions can become possible with the use of this metadata framework.

Key words: Knowledge discovery; Metadata card; Metadata registry; Microarray; Semantic net

[1] Middle East Technical University, Informatics Institute, Department of Health Informatics, 06531 Ankara, Turkey
[2] NATO HQ C3S, Information Services Branch, Blvd Leopold III B1110, Brussels, Belgium
[3] Department of Computer Engineering, Middle East Technical University, 06531 Ankara, Turkey

INTRODUCTION

The amount of data from experiments on microarray repositories becomes unmanageable as the number and content of submissions grow. The annotations and metadata additions to microarray records add to their existing content. However, these contextual data are not appropriately structured and do not conform to defined standards. The biomedical community has an interest in the interpretation of results of investigations in which microarrays are used. There are serious backlogs and exchange between the repositories cannot take place.

Several standardization initiatives in the microarray community have progressed. For example, MIAME (Minimum Information About a Microarray Experiment) focuses on content [1]. Others include: minimum dataset checklist, MIBBI (Minimum Information for Biological and Biomedical Investigations); object model, MAGE OM (Microarray Gene Expression Object Model); exchange platform, MAGE-ML (Microarray Gene Expression Mark-up Language); ontology, MGED (Microarray Gene Expression Data) Ontology [2]. These initiatives and their developments have been presented in review articles [3]. The three primary microarray repositories are: NCBI GEO (National Center for Biotechnology Information Gene Expression Omnibus) [4], EBI (European Bioinformatics Institute) ArrayExpress [5], and CIBEX (Center for Information Biology Gene Expression Database) [6].

Microarray repositories not only host the exper-

imental data but also present tools for querying and analyzing microarray records. Public-domain software has been developed on the BioConductor platform [7], such as GEOmetadb [8], to extend the functionality of the GEO repository, and to implement MAGE OM such as Sequence Analysis and Management System (SAMS) [9]. However, it is difficult for laboratories with less bioinformatics support to implement these applications. Thus, exchange and common understanding of data among disparate repositories continues to be an issue, despite the fact that mediating software is available [10]. The MINiML (MIAME Notation in Mark-up Language) and MAGE-TAB (Microarray Gene Expression Tabular) that have been developed to provide solutions to these problems [11] lack standard syntax and semantics. The solution is standard-related and can be provided with data management discipline using architectural frameworks.

The GEO repository has been selected for this study. We detected the following flawed and ambiguous entries on GEO records. (1) Inconsistent, incomplete, and incorrect entries for the same information element. For example, there are seven different spellings (United States of America, United States, USA, US, U.S., U.S.A., U.S.A) in address data for the country name 'USA'. There are city names in the country field. There are different patterns for the names of the same person, organization and date. (2) Three different versions of MINiML files for the same Series record that have different content are ***i)*** MINiML format for HTML Series record, ***ii)*** MINiML_family link within the HTML Series record, and ***iii)*** programmatically extracted Series data for the whole database. For example, one of the contributors is missing in Series Record GSE362 at "i." The Summary, PubMed ID, and Overall Design information fields are not available at "iii." (3) Related experiments (super Series and sub Series records) are not visible. A super Series record includes individually submitted subset records, all of which belong to one experiment. Since some Series records about an experiment are submitted separately without stating if they are related, it is difficult to trace records for such an experiment. For example, Vijay G. Sankaran submitted three Series records (GSE13283, GSE13284, and GSE13285) on 5 December 2008, which did not seem to be part of a single experiment. However, they prove to be connected to a single experiment so that GSE13285 is a super Series record, which includes subset Series GSE13283 and GSE13284. (4) The MIAME guideline (1), that the summary part of a microarray experiment record and the abstract in its publication should be the same, is not followed. For example, GSE3570 and GSE15808 have different summary information than the abstracts of their publications. This is a data integrity issue. GSE5546 was submitted to GEO in 2006 and has no citation information yet but its related publication was published in 2008 (PMID18271932.)

Some areas that have room for improvement in GEO data management are as follows: the microarray repositories are not connected. Thus, the records that are on different repositories are not visible. The MIAME is a content standard that lists the minimum content without format guidance. The type, content, format, and availability of data and metadata on different repositories are at varying degrees. Therefore, the regular exchange of data as it occurs among DNA repositories does not happen. There is an initiative by the ArrayExpress staff to import GEO records (approximately 10% of GEO records) on a weekly basis. However, they are not synchronized and if the records in GEO are updated, this will not automatically be reflected in the corresponding ArrayExpress entry [12].

The metadata about the records are not structured in accordance with the DC (Dublin Core) metadata standard [13]. There are entry anomalies, inconsistent terminology and even incorrect entries within metadata, *e.g.,* in contact information (names, organizations, country names, date) or in the summary. This can be handled with a structured data entry that is based on controlled vocabulary and ontology. Mandating patterns could also be included in a relevant schema file as tested in OpenSDE projects [14]. The experimenter could enter more of the experimental findings including metadata on contributors, experiment settings, biomaterials, data analyses, and especially on the result/summary section if there was a structured format.

The quality and state of the record is not clearly labeled at submission and throughout its lifetime. The quality metrics (values such as "verified" and "citation >10") and states (values such as "incomplete" or "retired") can add important meaning to the records. For example, some experiments are published in a high-citation publication, are performed by respected scientists, verified with RT-PCR (real-time polymerase chain reaction), and repeated with success. However, a record may be identified as a poor study if it is contradicted by experiments of high quality. There are also

comparability issues between different platforms as pointed out by the MAQC (MicroArray Quality Control) project [15].

Microarray records, related publications, and relevant data fed into databases such as gene and biological pathways should be consistent. The microarray repository should be the reference for other platforms. The semantics is not addressed in the design of microarray repositories. Thus, understandability and usability is weak, and life cycle management to include version and change management is not available.

More automation would be addressing slow curation work and the increasing number of backlogs. For example, GEO is experiencing a significant backlog in curated Dataset (GEO Data Set: GDS), creation and most of the submitted Series records (GEO Series: GSE)do not have a corresponding Dataset. Analysis tools operate on GDS records. At present, there are about 2721 GDS records and 22677 Series records (two GSE in one GDS on average). There are more than 15,000 GSE records yet to be curated. This amounts to an 80% backlog. Also, 20% of submitted Series records have not yet been published due to ongoing curation work. The number of GDS records has been unchanged since last year.

Here we report on a framework, MAdmf (Microarray Discovery Metadata Framework), which addresses these issues and its application to a case study.

MATERIALS AND METHODS

The Solution – MAdmf (Microarray Discovery Metadata Framework). The GEO repository is one of the main submission areas and a primary information resource for biomedical inquiries. There are three records (Platform, Sample, and Series) that are supplied by submitters on GEO. A GEO Series (GSExxx) record summarizes an experiment by linking a group of related samples. The GEO curator reassembles this data (one or more GSE records) into a GEO Dataset (GDSxxx), which represents samples processed using the same platform [4]. The GEO provides an XML file (MINiML) for each submitted record. Our focus has been on the MINiML file which includes both data (such as summary, platform, and sample data) and metadata (such as title, description and contact information) in this study. The MINiML file should serve as metadata card, but it is not named and designed as such.

We propose a framework, MAdmf, which includes a format for metadata in microarray results to address listed issues. The metadata card, semantic net and metadata registry are the key elements of this framework. The metadata card is an index card for storing basic data elements about specific domain information. The metadata card would provide the reader with information to assist him/her in making a decision as to whether the record(s) might suit his/her needs. SemNet is a small data model to represent domain-specific information. The metadata cards and SemNets are encoded in RDF/XML (a language for metadata and knowledge representation format). Syntax encoding schemes are used in SemNets. The metadata registry is a shareable repository for metadata and its related SemNet(s). The framework has four components as depicted in Table 1.

First, we provide a metadata card (Madmc, Microarray Discovery Metadata Card) to include common exchange elements in a standard format in accordance with metadata standards. Thus, discoverability, semantic interoperability, and integration operations

Table 1. MAdmf (microarray discovery metadata card framework).

Component	What It Does
MAdmc (microarray discovery metadata card)	Supports the MINiML file
Semantic layer (semantic nets)	Details domain-specific topics, fortifies the intended meaning; discloses otherwise hidden data
Query layer (optional)	SPARQL queries
MAdmr (microarray discovery metadata registry)	Main files for MAdmf[a] are stored at this ebXML-based shared space

[a] The content of MAdmf is as follows: MAdmc.xml: Microarray discovery metadata card; MAdmc.xsd: schema file for Madmc; Experimenter.rdf: SemNet (FOAF/RDF file) for experimenters; Result.rdf: SemNet (RuleML Datalog/RDF file) for result/summary section; MAdmc.rq: Query file in SPARQL to run on SemNets.

are supported. The format and structure of MAdmc is the extension of MINiML [16] and based on DC, and Metadata Registry Standard [17]. Second, SemNets are developed for experimenters and results for related experiments. Third, Queries in SPARQL (Simple Protocol and RDF Query Language) [18] format, have been developed for information access and discovery operations. Finally, these products (MAdmc, SemNets, and associated queries) are stored in a common reference area for further use. They can also be exchanged among microarray repositories. Such an exchange or share may reduce the need for multiple submissions and undesired redundancy where raw data resides at its original place.

The metadata card and its associated SemNet(s) may hold frequently accessed data patterns as well as previously hidden or unavailable content in a structured format. Thus, much more automated processing can be involved. They can be queried without a need for a dedicated application. It is because they are represented in RDF/XML that is extendable, integrable, and queryable. The proposed framework is about organizing and structuring the microarray metadata in its syntax and semantics. The user may perform complex queries and backlogs can be reduced with the use of such machine processable metadata cards and their related SemNet(s). Microarray analysis has already evolved into microarray informatics. We believe that such architectural solutions are needed in the microarray domain. The goal to reach shared semantics and common understanding can be realized by applying data management principles over structured and semantically enriched data.

There are two main contributions of this study with the proposition of such a metadata framework. The experimenter could submit more contextual data. And, machine interpretable content is promoted that would support curation and analysis work. The expressive power gained is twofold. The producer is tempted to include more of the experimental findings and the implicit or previously unavailable data becomes discoverable by consumers who get the intended meaning.

The life cycle management of the records is important. The experimentation and its publication together with some updates on specific databases constitute the first part of the activities in the lifetime of the record. The biomedical community has been successful in this part. However, the important part, which has largely been overlooked, follows this first part and ends when the record is deleted. This second part involves in validation, modification and knowledge discovery (for example, developing research hypotheses in meta-analysis) operations. The weakness lies here as highlighted in several publications [19]. This study is performed on this part to make the results visible, understandable and usable.

MAdmf will require additional resources but such an effort will pay off in data-centric operations. We enforced data management by organizing and structuring data that would improve the quality of microarray data analysis. Data management must be built into the process from the beginning to support information system development. It is a knowledge-interoperable development that allows domain experts to build or contribute to a separate data layer which can then be incorporated into knowledge-based design [20]. For example, the domain expert may create a SemNet to include the information "P53 gene related experiments which finds relevance on arsenite and apoptosis on breast cancer as verified by RT-PCR, published in peer-reviewed journal, with citation >10, curated into GDS record and inputted to a specialized repository (such as GO or pathway database, Reactome [21]) in the last decade," provided that metadata cards contain it.

We used the tools from W3C resources in the development of these products. Respective concepts and techniques are borrowed from semantic web (SemWeb), data management, structured reporting, electronic business management, configuration management, and metadata standards. We state that shareable metadata cards which are semantically powered by semantic nets can be a solution. The framework presented in this study can be used in any high throughput repositories as well as third party platforms.

MAdmc (Microarray Discovery Metadata Card). MAdmc is a metadata card for a microarray experiment. The metadata card is a stable concept and used for resource discovery. In our framework, it not only facilitates the visibility but also the usability and common understanding. With that goal in mind, we extended the structure, organization, and syntax of the MINiML file to produce MAdmc. The overall syntax of MAdmc is said to be a format layout for the content. We propose the standardization of metadata in the MINiML file by including DC elements and by introducing the metadata card concept. The metadata card has administrative, descriptive, structural and semantic elements. Dublin core is a standard (ISO 15386) for

cross-domain resource description. The use of DC elements in metadata definition also promotes structured entry. Thus, it becomes easy to find and understand information resources. The MINiML seems to serve this purpose but its structure and content is not appropriate to support this function. Structuring the records and making structured entry for data elements within the records are closely related and complementing paradigms. The structured entry for the values is enforced by selecting a value from a controlled vocabulary or entering a value dictated by a pattern in the schema file.

Microarray records pose more meaning when analyzed in a batch and placed in a biological context. Since the experimental settings, samples, methods, tools, and format widely differ; it is a challenging task for microarray repositories to offer such an analysis in an efficient manner. We introduced the layers into the organization of metadata elements and employed data and syntax encoding schemes. Repeatability and structural relationships between elements were defined. For example, the title may be repeated (alternative title). Or, the use of an element can depend on a condition of another one. Life cycle management concept was introduced with the use of versioning and modification status information. The life cycle management covers the period from the submission until the retirement, thus bringing up the living record concept. It is implemented based on the relation element which may include the values 'is version of,' 'replaces,' or 'part of.' Thus, this becomes a part of the microarray data rather than the software code. The human or automated users can modify, annotate, and verify a record several times throughout its lifetime.

We developed an XML application (MAdmc program) so that the user selects the elements from the MINiML document and add new ones from the DC Metadata Set and attributes from the Metadata Registry standard to create the MAdmc. The DC Metadata Set includes 15 information elements. In MAdmc, we added four new information elements (three in Security, one in Format Specification layer) and detailed each element with the introduction of four attributes including an obligation category. We then organized them into four layers as shown in Table 2.

The detail of metadata card definition is given in MAdmc.xsd file, Figure 1. The user can reference this schema file to create his/her own instance document (metadata card). The experimenter or curator can create the MAdmc file by using the MINiML file and the MAdmc program, as explained in the Case Study section.

The structure of MAdmc can also be extended by employing associations among the tags. The associations can be represented in EBNF (Extended Backur Naur Form) syntax and defined in the schema file, as was the case for the structured messaging system at NATO (North Atlantic Treaty Organization). For example, an element may occur several times; information elements such as the title, location, organization may have alternate contents; information elements are labelled with one of the categories such as 'Mandatory,' 'Optional' or 'Conditional,' requirement and prohibition of use on a condition (*e.g.,* mutual exclusivity) may be enforced. The rules are encoded in Xpath expressions [22]. Although it is an optional extension, this topic could be visited upon recognition of the metadata concept. The layers (segmentation), repeat,

Table 2. MAdmc elements (2a) and obligation categories (2b) for elements.

***2a)* Layers**	**Elements**	**Attributes (ISO 11179)**
Security	Policy Classification Category	
Resource Description	Title Identifier Creator Publisher Contributor Date Rights Language Type Source Relation	Definition Comment Obligation category Max. occurrence
Format Specification	Format Version	
Content Description	Subject Description Coverage	
***2b)* Obligation**	**Definition**	
Mandatory **(M)**	An element must be supplied with a value to comply with MAdmc	
Conditional **(C)**	The usage of an element is dependent upon a particular condition	
Optional **(O)**	An element may be supplied with a value but it is not a requirement	

```
<?xml version="1.0"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" xmlns="urn:edu:metu:MAdmc:1:0"
    targetNamespace="urn:edu:metu:MAdmc:1:0" elementFormDefault="unqualified" xml:lang="en-GB">

    <!--    This schema specifies the Dublin Code metadata element set portion of the
            Microarray Discovery Metadata Card (MAdmc). It specifies a set of information
            fields that are to be used to describe all items belonging to the microarray
            experiments in XSD.     -->
    <!--
    Root element of MAdmc - Discovery Metadata Specification
    -->
    <xs:element name="MAdmc_DC">
            <xs:complexType>
                    <xs:sequence>
                            <xs:element ref="Security"/>
                            <xs:element ref="ResourceDescription"/>
                            <xs:element ref="FormatDescription"/>
                            <xs:element ref="ContentDescription"/>
                    </xs:sequence>
            </xs:complexType>
    </xs:element>
    <!--
    Definition of MAdmc Layer Elements
    -->
    <xs:element name="Security">
            <xs:annotation>
                    <xs:documentation>The security layer elements enable the description of security
                     classification and other security-related fields. These fields provide for the
                     specification of security-related attributes of the associated data assets and
                     may be used to support access control.</xs:documentation>
            </xs:annotation>
```

Figure 1. MAdmc.xsd (schema file for microarray discovery metadata card).

and structural constraints in the mark-up tags can be designed to enhance the structure and meaning in the metadata card.

Semantic Nets – Micro Formats. Different parts of the metadata card can be detailed with SemNets. Such work is analogous to the one performed by domain experts on data layer in knowledge-based systems. The SemNets can be generated for each GEO record, or a group of related records or the whole repository, depending on the contextual requirements. The SemNets accompany their related metadata cards and they can all be integrated into a related RDF store. The RDF store can be coupled with any platform and can then be used for ontology development, database modeling, and for any semantic task.

Data and syntax encoding schemes are used for information elements such as experimenters, address, description and summary. The data encoding schemes could be Controlled Vocabularies [*e.g.,* Code lists (ISO 3166-Country codes), Classifications (ICD), Subject headings (MeSH)] or formal notations such as ISO 3601(Date Time Group), ISO 639 (Language), or use of a specific name space. Friend of a Friend (FOAF) and Rule Mark-up Language (RuleML) syntaxes are used for encoding relevant data into SemNet. The FOAF is a SemWeb language that describes relationships among people in RDF by forming ontology on its own [23]. RuleML is a mark-up language for publishing and sharing rule bases. It is based on a deductive reasoning engine and its statements can be embedded in knowledge-based systems [24]. The experimenter and the summary parts are extended with SemNets in accordance with relevant syntax to add meaning and to build semantic expressiveness in this study. The experimenters are modeled by using FOAF syntax, and the result part is modeled by using RuleML data log syntax. Online tools in the public-domain, as suggested by W3C, are used in the development of the SemNets.

The human concept in the microarray record should be structured. There are types such as human, automated; categories such as scheduled, unscheduled; status such as novel, experienced; roles such as producer, consumer; actors such as submitter, contact, contributor, author of publication, publisher, curator, funding agency representative, government official, meta-analyst, verifier, system developer, reviewer, *etc*. Such a detailed definition may hold valuable information for a potential consumer. Data sets are at different maturity levels in terms of structure and content. One's data may be labeled as metadata or information by someone else. And today's information may become data in the future in its lifetime. An experimenter may need to make a search for the human element to make some decisions for experiment design. There are mature formats such as hcard [25], vcard [26], or W3C's PIM (Personal Information Management) [27] to include this information into the FOAF model to form a coalition of complementing vocabularies.

The summary information has been a frequently accessed area. This portion of the microarray record should also have a machine understandable structure and content. For that reason, we employed an encoding process for the statements to create a SemNet. We included free text statements, the encoded format, and annotations which are all in RDF notation. More data are stored in the RDF format to create linked data today. The RDF files can be integrated into a persistent RDF store to form connected graphs.

The properties and relationships of information resources are described within RDF graphs for SemNets [experimenter net (in FOAF) and result net (in RuleML Datalog)] in our study. These are associated to each or a group of related MAdmc record(s) in accordance with which specific knowledge is represented. Thus, Experimenter and Result SemNets can be packed with metadata cards while ontology use is in place. SemNets are data models that are easy to create for specific domain information, which can support both ontology development and database design. Ontology extensions can subsequently be built from these SemNets. For example, describing a person in ontology may eventually converge to a FOAF model. A new vocabulary and ontology extension can be generated from the RDF resources. The RDF triples for information objects may become instances for existing Ontology Web Language (OWL) classes or they may trigger the creation of new classes for specific concepts. It is obvious that ontology terms should be used as the tokens in a SemNet. Ontology is used for annotation, but we encode data and metadata with syntax systems in SemNets.

There is a proliferation of ontologies, and there are interoperability problems among them. Ontology for Biomedical Investigations (OBI) standardization initiative focuses on upper ontology development, whereas lower level ontology remains in the realm of domain-specific ontology such as MGED Ontology. Ontology is a conceptual model that may not map to physical data sources, whereas a SemNet does. Semantic net can serve as a basis for bottom up ontology development. Ontology is monotonic where new statements should not falsify previous conclusions [28]. Regarding microarray experiments, there are conflicting results as well as supporting ones and SemNets may include such non monotic statements.

Queries. Some frequently asked queries can be materialized in SPARQL within the framework and be posted to a shared registry; SPARQL is similar to Structured Query Language (SQL) and is de-facto standard as RDF Query language. The answers for specific queries for which the results are difficult to obtain at the moment such as the following can then become possible when MAdmf is employed: ***1)*** list submitters who have worked on breast cancer over Tamoxifen effect on humans within X organization for which the records have been curated to GDS; ***2)*** list breast cancer records that have been published in SCI journals with citation numbers >10 and verified and have been included in special databases; ***3)*** list all facts and hypotheses from records related to the P53 gene between 2000 and 2009; ***4)*** list the versions, states (*modified, retired, etc.*), type (*comparative, collaborative, validation, etc.*) and modification details of BRCA1 and BRCA2 related records; ***5)*** list super GSE records and their child records that are related to experimentation on gene ATM that finds relevance on apoptosis on breast cancer by submitters from USA in the last decade. The metadata card and SemNets can hold data to answer these questions in a knowledge representation format. One sample query and its result are demonstrated within the Case Study section.

MAdmr (Microarray Discovery Metadata Registry). Madmr will be the key element to enforce a data strategy by facilitating visibility, usability and understandability of data assets. The submission package to this ebXML (Electronic Business using XML) based shared space may include MAdmc, SemNet, Schema file, Query file, and a Guidance document, Figure 2. MAdmr can be either GEO or another repository. A federated system of microarray repositories can also assume a metadata registry role to host microarray discovery data.

Different users (such as submitter, reviewer, or web services program) can subscribe to such a registry. And producer(s) can make modifications and create new versions throughout the lifetime of the microarray records before retirement on metadata registry.

The Case Study. The GEO records (Series, Platform, and Sample) and contact data have been downloaded and stored in OpenOffice BASE Database and examined with a domain specialist in terms of structure and semantics. We accessed 677 Breast Cancer experiment results (677 GSE records, 89 GDS records) in more than 22,000 Series records for the case study. We developed the metadata card by using our MAdmc program, Figure 3.

Then, two sets of SemNets have been created per record(s) using RDF Editor Protégé [29], online W3C XML Schema Validation [30] and RDF Validation tools [31]. SemNets (RDF graphs) in Protégé are queried by using SPARQL. First SemNet was for experimenters in FOAF/ RDF (was not included for brevity), and the second one was about the result section, Tables 3 and 4. Note that the examples about these SemNets are given for proof of concept only. Two encoded statements by using RuleML Datalog (casual first order logic) are given in Table 3.

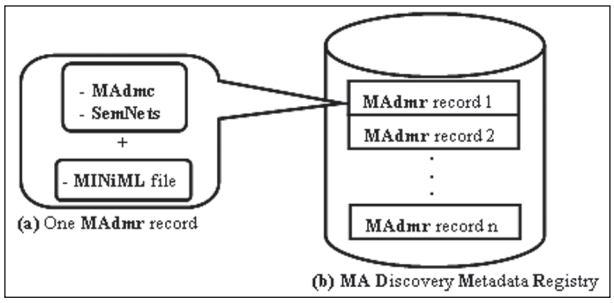

Figure 2. The MAdmr content.

We show an entry level encoding in Table 3 to give an insight. The encoding could have gone further with deeper mark-ups as demonstrated in Table 3, a.2. The statements could have been further categorized such as experimental, statistical, and computational or its status could be labeled as verified, challenged, withdrawn, or modified. The goal is to highlight the elements of MAdmf. Thus, we do not claim to present the optimal representation. We here demonstrate that the results can be formatted in a syntax encoding scheme like RuleML Datalog. This structured set of statements can then be shared and processed by automated means.

The individual statements for each of these 677 breast cancer GEO records can form a semantic net that is associated to the relevant MAdmc. There may also be global statements about meaningful findings for a specific sub-group of records or whole breast cancer records. SemNets can be in different representations such as triple notation, and graph diagram as well as XML/RDF format. We include three elements in this encoding of the SemNet: the original statements, the encoded format, and annotations. The annotation part of this package provides contextual information and may include if: ***1)*** there is a related publication?; ***2)*** the results are posted somewhere else such as GO or a pathway database?; ***3)*** there are other versions?; ***4)*** it is a fact or hypothesis?; ***5)*** it is verified or challenged?

Relevant name space declarations like "MAdmc" can be included into a MAdmc schema file to support the additional definitions, Table 4. A sample Result SemNet is given in RDF/XML format in Table 4, and its graphical output from RDF Validator is given in Figure 4.

There may be a different level of encoding for each record based on the availability of relevant information. We recommend entry level encoding at the be-

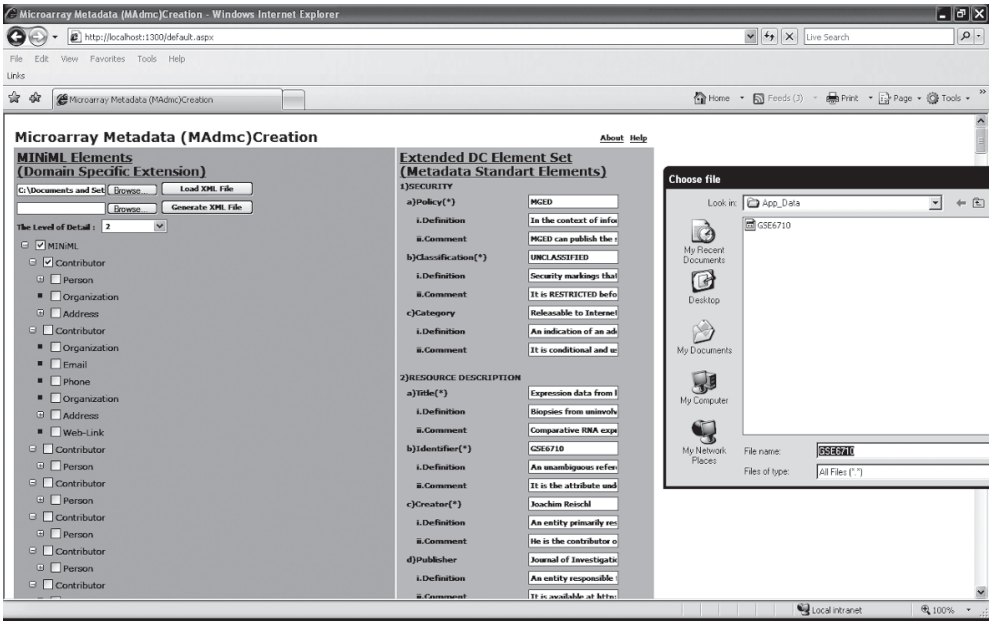

Figure 3. MAdmc program. An application that reads the MINiML file, accepts values for additional fields and creates the metadata card (MAdmc.xml).

Table 3. Statements from GEO records encoded in the RuleML Datalog.

a) **A Fact From GSE12848** ***MicroRNA silences anti-proliferative genes***	**Free text**
<Atom> <Rel>silence</Rel> <Ind>MicroRNA</Ind> <Ind>anti-proliferative gene</Ind> </Atom>	Encoded text (a.1) Condensed encoding
<rulebase> <fact> <Atom> <opr><Rel>silence</Rel></opr> <arg index=”1”><Ind>MicroRNA</Ind></arg> <arg index=”2”><Ind>anti-proliferative gene</Ind></arg> </Atom> </fact> </rulebase>	Encoded text (a.2) Expanded form of encoding for the fact in (a.1)

b) **A Rule from GSE5483** ***RT-PCR confirms the induction of early growth response1 (Egr1) and Stratifin (Sfn) by estradiol-progesterone (EP) and RT-PCR shows that P53 is independent***	**Free text**
<And> <Atom> <Rel>confirmed by</Rel> <Ind> induction of Egr1 and Sfn by EP</Ind> <Var id=1>RT-PCR</Var> </Atom> <Atom> <Rel>show</Rel> <Var id=1>RT-PCR</Var> <Ind>be P53 independent</Ind> </Atom> </And>	Encoded text

ginning, and as acceptance and experience grows, the encoding may be more sophisticated. There are platforms such as jDREW [32] on RuleML Data log in that direction. We not only encode and represent the free-text result section but also open the way for triggering derivations from an already stored rule base. In fact, this is the job of a rule-based system. We demonstrate the capability. Rules can extend the OWL as included in the Semantic Web architecture. In that regard, for example SWRL (semantic web rule language) combines RuleML (Horn-like rules) with OWL (axioms) [33]. And the RIF (rule interchange format) mechanism allows different representations to be grouped for further use [34]. The metadata card and SemNets can also be queried using the online SPARQL tool [35]. The query file in Figure 5 can be attached to the related SemNet file.

RESULTS AND DISCUSSION

There is a rising volume of microarray data. The challenge is if we can provide meaning as well as structure and syntax to this information space for automated means.

The summary part of the records on microarray repositories and related publications are not synchronized, not appropriately structured. They are in free-text format. The statements are usually incomplete and ambiguous, thus not easily comparable with others in similar studies. The results should be visible, un-

Table 4. This is a Result SemNet of GEO Series record, GSE12848 (P53 gene related breast cancer record)

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:foaf="http://xmlns.com/foaf/0.1/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:MAdmc="http://www.ii.metu.edu.tr/MAdmc#">
     <!-- about (title and description) Breast Cancer Records (677) in April 2011 -->
    <rdf:Description
        rdf:about="http://www.ncbi.nlm.nih.gov/geo/query/browse.cgi?view=series">
        <dc:title>Breast Cancer Records</dc:title>
        <dc:description>The Result of a P53 related breast cancer Series record is captured in this SemNet</
dc:description>
        <dc:source>You can access GSE in this link</dc:source>
    </rdf:Description>
    <rdf:Description
        rdf:nodeID="GSE12848">
        <dc:identifier>GSE_12848</dc:identifier>
        <dc:title>p53-repressed miRNAs are involved with E2F in a Feed Forward Loop Promoting Proliferation</
dc:title>
        <MAdmc:silence>anti-proliferative genes</MAdmc:silence>
        <MAdmc:category> category="fact" status="modified" verified="RT-PCR" MicroRNAs silence anti-prolifera-
tive genes</MAdmc:category>
        <MAdmc:RuleMLDatalog>
        <Atom>
          <Rel>silence</Rel>
          <Ind>MicroRNA</Ind>
          <Ind>anti-proliferative gene</Ind>
        </Atom>
        </MAdmc:RuleMLDatalog>
        <MAdmc:ruleset>
1:MicroRNAs silence anti-proliferative genes.
2:MicroRNAs are novel key players in the mammalian cellular proliferation network.
3:Expression of microRNAs is down-regulated in senescent cells and in breast cancers harboring wild-type p53.
4:MicroRNAs are repressed by p53 in an E2F1-mediated manner.
5:MicroRNAs silence anti-proliferative genes, which themselves are E2F1 targets.
6:MicroRNAs and transcriptional regulators appear to cooperate in the framework of a multi-gene transcriptional and
post-transcriptional feed-forward loop.
        </MAdmc:ruleset>
        <MAdmc:similar>GSE5483</MAdmc:similar>
        <MAdmc:Publication>Publication= PMID=19034270 SCI=11 Impact factor=12.125SpecialDB=http://www.
uniprot.org/uniprot/Q8TCJ2BiologicalPathway=http://www.reactome.org/</MAdmc:Publication>
        <MAdmc:summary_alternate_abstract>Normal cell growth is governed by a complicated biological system, featuring multiple levels of
control, often deregulated in cancers. The role of microRNAs in the control of gene expression is now increasingly appreciated, yet their involve-
ment in controlling cell proliferation is still not well understood. Here we investigated the mammalian cell proliferation control network consisting of
transcription regulators, E2F and p53, their targets, and a family of 14 microRNAs. Indicative of their significance, expression of these microRNAs
is down-regulated in senescent cells and in breast cancers harboring wild-type p53. These microRNAs are repressed by p53 in an E2F1-mediated
manner. Furthermore, we show that these microRNAs silence anti-proliferative genes, which themselves are E2F1 targets. Thus, microRNAs and
transcriptional regulators appear to cooperate in the framework of a multi-gene transcriptional and post-transcriptional feed-forward loop. Finally, we
show that, similarly to p53 inactivation, overexpression of representative microRNAs promotes proliferation and delays senescence, manifesting the
detrimental phenotypic consequence of perturbations in this circuit. Together these findings position microRNAs as novel key players in the mam-
malian cellular proliferation network.</MAdmc:summary_alternate_abstract>
        </rdf:Description>
</rdf:
```

derstandable, and usable throughout their life cycles. This is an information management principle. Once we structure (MAdmc) and encode the contextual data (SemNet), not only certain operations such as discovery and exchange become feasible, but also hidden and previously unavailable facts may be extracted from

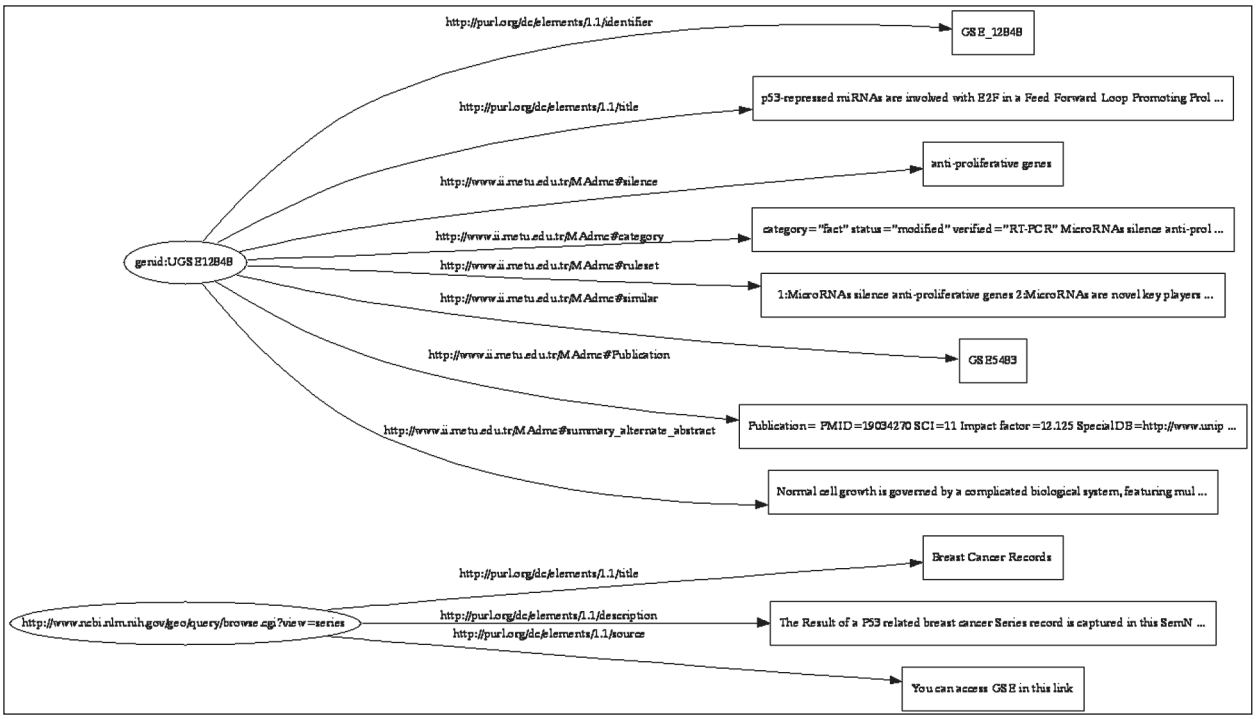

Figure 4. The graph output for the SemNet in Table 4 as validated by the RDF Validator.

such structured and encoded data sets. The structured entry paradigm can also be enforced in addition to annotation *via* ontology within a SemNet.

If one searches MAdmr (MAdmc and SemNets), it will be more efficient than a search on GEO for domain specific information at present. It is something like sorting data before an efficient search. It is the process of linking data for which the resources-properties-relationships are identified. MAdmf brings about an overhead, but future benefits will justify this start-up cost.

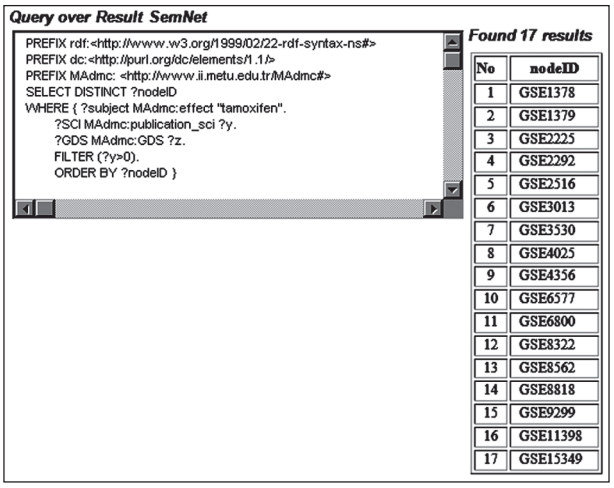

Figure 5. A sample SPARQL query on Result SemNet (online "SPARQLer RDF Query Tool" used at http://www.sparrl.org/query.html)

Describing data in a structured manner can be better done in a database, but microarray information space includes several microarray repositories, experimenter web sites, publications, and specialized databases. Practically, they cannot all be stored in a database or easily be federated. If all parties could have agreed to use MAGE-OM object model and MAGE-ML exchange platform, there would have been no format, exchange and integration issues. But, this is unlikely and there will always be different implementations that bring about exchange and interoperability problems. Note that metadata cards and semantic nets can also be used in a MAGE-OM/MAGE-ML based repository.

We can say that the microarray domain includes semi-structured data that can be best managed with SemWeb technology. SemWeb emphasizes the use of metadata standards and connected data to support data centric operations. The proposed framework, MAdmf follows SemWeb paradigm. The microarray community should adopt such a data centric approach because the operations are data intensive. Data management is the vehicle for data centric initiatives, and an IT system is as weak as its data management. A data layer is built separately than the business logic layer in future-proof applications. MAdmf is related to the data layer. It promotes the data standardization on microarray re-

positories. Any modelling or application development effort can then follow its use.

We examined the MINiML file and introduced an extended format for a metadata card in this study. We created domain-specific SemNets and offered their posting to an ebXML based metadata registry, which provides a shared information space. Thus, in the proposed framework: ***1)*** the producer can add structured data and the consumer can get the conveyed meaning (*what has been received is limited to what has been understood*), ***2)*** due to the possibility for more automation, backlog is reduced in curation work *(from submitted records to GEO Series or GEO Series to GEO Datasets or GEO Datasets to Array Express records*), ***3)*** ambiguity and redundancy is reduced with standard format and additional semantics, ***4)*** data centric approach is adopted, and the quality and expressiveness of data are promoted where a separate data layer from business logic is maintained, ***5)*** consumers reach data otherwise unavailable (*new entries in descriptive information and semantic layer*), ***6)*** life cycle management (*lifetime modification and living data set*) concept is introduced, ***7)*** visibility, understandability and usability are enforced, ***8)*** users can use W3C and the public-domain tools to extract data, ***9)*** the controlled vocabularies (*Countries, Date/Time Group, Names*) are used not only to annotate but also to encode the metadata and data, ***10)*** the produced metadata card and its associated SemNet(s) are extendable, integrable, queryable and exchangeable, ***11)*** microarray records and subsequent entries (*publication, specialized databases*) can be synchronized.

The extension on the MINiML file has three aspects. First, content is detailed in summary and experimenters. Second, format is materialized through the employment of data and syntax encoding schemes. The organization and structure is improved with the introduction of layers, additional metadata elements and attributes. Third, the process is extended with the new concepts such as life cycle management, metadata registry use, and structured entry. In this manner, the MINiML file has been transformed into a metadata card and its semantics is extended with SemNets. Then, they can be used in any similar data center.

The people, experiment, and result data are linked as the proposed framework provides such a foundation. Thus, for example, a meta-analyst can get a consolidated summary of the result part of all breast cancer data sets by using a SPARQL query. The originator, the curator, the developers and other experimenters may benefit from this framework. We give the specification and present key products in a case study where a proof of concept is introduced.

The MAGE-ML and MINiML seem to be alternative structures but they are not in reality. The MINiML is an intermediary data structure, whereas a MAGE-ML application can be developed onto. The creation of MAdmc and SemNet includes two different and complementary contributions to support MINiML towards a format and exchange standard. They do not replace any existing work. However, if adopted, they can be a focus for discovery, integration and exchange. The SemNets can be created for other parts of microarray record, in addition to the experimenter and summary data. Note also that this study can easily be adapted to other microarray repositories or high throughput repositories.

There is up to a 3% monthly increase in records at GEO in recent years. There is a backlog of up to 20% in Series records for varying reasons. There is also a serious backlog of 80% in Dataset transformation (GSE to GDS) tasks performed by GEO curators. This is likely to increase because the amount of data and its complexity are on the rise (Table 5).

An RDF-enabled database that provides both reasoning and ontology modeling capabilities, may consume metadata card and SemNets. Another one could be a semantic platform that connects heterogeneous data contained in microarray repositories and related publications. One can combine people, location, or-

Table 5. Data composition as of May 6, 2011.

GEO Repository	Public	Unreleased	Total	Backlog
Platforms (GPL)	8,713	494	9,207	~6.0%
Samples (GSM)	557,206	121,682	678,888	~18.0%
Series (GSE)	22,677	4,224	26,901	~16.0%
Datasets (GDS)	2,721	–	Number of experiments (Series records/2)	~80.0%

ganization, and date information with experimentation results across microarray information space to formulate complex inquiries over SemNets and metadata cards. Moreover, the development of knowledge interoperable systems with a separate data layer can be facilitated with such a mode of operation on data. Equally, rule based systems can make use of the summary portion of a microarray record that is structured and encoded.

Standardization studies like this one, which promote machine understandability and semantic interoperability, are required. This study not only brings metadata card and semantic net concepts within a format standard approach but also introduces the importance of the life cycle management, data management and structured entry concepts. Such a study will be beneficial, especially for producers, curators, future experimenters and system developers, whether they employ manual or automated means. The experimental data, encoded formats, and program, can be requested from the corresponding author.

CONCLUSIONS

Microarray informatics has been an active research direction, especially in architectural and computational aspects. The conduct of microarray experimentation is only the first part of the process. The second part, which is often poorly handled, is to organize, present, exchange, understand and use the interpreted experimental evidence. Thus, gaps and inconsistencies as well as ambiguities in the microarray knowledge base such as candidate theories, scientific disagreements, and open questions can be managed and resolved. To obtain new insights and knowledge, the data generated by high throughput experiments need to be transformed into meaningful executive summaries. We propose metadata card and semantic net to represent such summaries. Testing the hypotheses based on these summaries may become an interesting task for computational biology.

This study covers the improvement in the structure, syntax, and semantics of the metadata of microarray experiment result data sets. We demonstrate that the introduction of metadata cards can support discovery and exchange operations. SemNets could be a vehicle to represent the meaning in the microarray domain. Since domain experts created the SemNets, previously unknown details can be revealed. The proposed framework, MAdmf, does not replace but complements the existing products in the microarray domain. MAdmf can be used in microarray repositories, other high throughput repositories, and third-party platforms. The driving philosophy behind MAdmf comes from data management, knowledge engineering, semantic web and structured messaging paradigms.

We believe that once such standardization efforts become adopted, the required tools and detailed guidance will follow. The following topics need further investigation. The set up of a metadata registry and guidance for how to submit a package to the metadata registry; the life cycle management of records; structured data entry; configuration model to include states (retired, incomplete, or complete) and status in each state (conflicting, derived, or verified); the synchronization mechanism among various repositories over metadata information elements.

ACKNOWLEDGMENTS

The authors thank Dr. Neslihan Aygun Kocabas (Department of Toxicology, Faculty of Pharmacy, Ankara, Turkey; neslihanak@gazi.edu.tr) for her support and her assistance in the interpretation of GEO records.

REFERENCES

1. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001; 29(4):365-371.
2. MGED (Microarray Gene Expression Data) Society [Internet]. [cited 2011 May 3] (http://www.mged.org/).
3. Field D, Sansone SA. A special issue on data standards. OMICS. 2006; 10(2): 84-93.
4. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R. NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 2009. 37(Suppl. 1): D885-D890.
5. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M,

Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A. ArrayExpress update-from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res. 2009; 37(Suppl. 1): D868-D872.

6. Sugawara H, Ogasawara O, Okubo K, Gojobori T, Tateno Y. DDBJ with new system and face. Nucleic Acids Res. 2008; 36(Suppl. 1): D22-D24.
7. Bioconductor, open source software for bioinformatics [Internet]. [cited 2011 May 3] (http://www. bioconductor.org).
8. Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: powerful alternative search engine for the GEO. Bioinformatics. 2008; 24(23): 2798-2800.
9. Bekel T, Henckel K, Küster H, Meyer F, Mittard RV, Neuweger H, Paarmann D, Rupp O, Zakrzewski M, Pühler A, Stoye J, Goesmann A. The Sequence Analysis and Management System – SAMS 2.0: data management and sequence analysis adapted to changing requirements from traditional sanger sequencing to ultrafast sequencing technologies. J Biotechnol. 2009; 140(1-2): 3-12.
10. Te Pas MF, Hulsegge I, Coster A, Pool MH, Heuven HH, Janss LL. Biochemical pathways analysis of microarray results: regulation of myogenesis in pigs. BMC Dev Biol. 2007; 7: 66.
11. Rayner TF, Rocca-Serra P, Spellman PT, Causton HC, Farne A, Holloway E, Irizarry RA, Liu J, Maier DS, Miller M, Petersen K, Quackenbush J, Sherlock G, Stoeckert CJ Jr, White J, Whetzel PL, Wymore F, Parkinson H, Sarkans U, Ball CA, Brazma A. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. BMC Bioinformatics. 2006; 7: 489.
12. ArrayExpress, Functional Genomics Database at European Bioinformatics Institute. [cited 2011 May 3] (http://www.ebi.ac.uk/microarray-as/ae/).
13. DC (Dublin Core) Metadata Initiative [Internet]. [cited 2011 May 3] (http://dublincore.org/).
14. Lors RK, van Ginneken AM, van der Lei J. OpenSDE: a strategy for expressive and flexible structured data entry. Int J Med Inform. 2005; 74(6): 481-490.
15. MAQC Consortium, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Schrf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, Zhang L, Amur S, Bao W, Barbacioru CC, Lucas AB, Bertholet V, Boysen C, Bromley B, Brown D, Brunner A, Canales R, Cao XM, Cebula TA, Chen JJ, Cheng J, Chu TM, Chudin E, Corson J, Corton JC, Croner LJ, Davies C, Davison TS, Delenstarr G, Deng X, Dorris D, Eklund AC, Fan XH, Fang H, Fulmer-Smentek S, Fuscoe JC, Gallagher K, Ge W, Guo L, Guo X, Hager J, Haje PK, Han J, Han T, Harbottle HC, Harris SC, Hatchwell E, Hauser CA, Hester S, Hong H, Hurban P, Jackson SA, Ji H, Knight CR, Kuo WP, LeClerc JE, Levy S, Li QZ, Liu C, Liu Y, Lombardi MJ, Ma Y, Magnuson SR, Maqsodi B, McDaniel T, Mei N, Myklebost O, Ning B, Novoradovskaya N, Orr MS, Osborn TW, Papallo A, Patterson TA, Perkins RG, Peters EH, Peterson R, Philips KL, Pine PS, Pusztai L, Qian F, Ren H, Rosen M, Rosenzweig BA, Samaha RR, Schena M, Schroth GP, Shchegrova S, Smith DD, Staedtler F, Su Z, Sun H, Szallasi Z, Tezak Z, Thierry-Mieg D, Thompson KL, Tikhonova I, Turpaz Y, Vallanat B, Van C, Walker SJ, Wang SJ, Wang Y, Wolfinger R, Wong A, Wu J, Xiao C, Xie Q, Xu J, Yang W, Zhang L, Zhong S, Zong Y, Slikker W Jr. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol. 2006; 24(9): 1151-1161.
16. Barrett T, Troup DB, S, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO: mining tens of millions of expression profiles — database and tools update (MINiML). Nucleic Acids Res. 2007; 35(Suppl. 1): D760-D765.
17. ISO 11179, Metadata Registries (MDR). [cited 2011 May 3] (http://www.iso.org).
18. SPARQL, Query Language for RDF. [cited 2011 May 3] (http://www.w3.org/TR/rdf-sparql-query/).
19. Miller H, Norton CN, Sarkar IN. Genbank and PubMed: How connected are they? BMC Res Notes. 2009; 2: 101.
20. Garde S, Chen R, Leslie H, Beale T, McNicoll I, Heard S. Archetype-based knowledge management for semantic interoperability of electronic health records. Stud Health Technol Inform. 2009; 150: 1007-1011.
21. REACTOME, a curated knowledgebase of biological pathways. [cited 2011 May 3] (http://www.reactome.org).
22. XPath (XML Path Expression Language). [cited 2011 May 3] (http://www.w3.org/TR/XPath).
23. FOAF (Friend of a Friend) Project. [cited 2011 May 3] (http://www.foaf-project.org/).
24. RuleML (Rule Markup Language) Initiative. [cited 2011 May 3] (http://www.ruleml.org/).
25. hcard, Format for representing people, organizations, and places. [cited 2011 May 3] (http://www.w3.org/2006/03/hcard).
26. vcard (Format for electronic business cards). [cited 2011 May 3] (http://www.w3.org/TR/vcard-rdf).
27. W3C PIM (Personal Information Management) Vocabulary. [cited 2011 May 3] (http://www.w3.org/2000/ 10/swap/pim/).
28. Esposito M. An ontological and non-monotonic rule-based approach to label medical images. Third Proceedings of the Third International Institute of Electrical and Electronics Engineers (IEEE) Conference on Signal-Image Technologies and Internet-Based System

(SITIS), Shanghai, People's Republic of China, 16-18 December 2007: 603-611.

29. Protégé, open source ontology editor and knowledge-base framework. [cited 2011 May 3] (http:// protege. stanford.edu/).
30. W3C XML Schema Validator. [cited 2011 May 3] (http://www.w3c.org/2001/03/webdata/xsv).
31. W3C RDF Validation Service. [cited 2011 May 3] (http://www.w3.org/RDF/Validator/).
32. jDREW, A Java Deductive Reasoning Engine for the Web (SPARQL, RuleML support). [cited 2011 May 3] (http://www.jdrew.org/).
33. SWRL (Semantic Web Rule Language). [cited 2011 May 3] (http://www.w3.org/ Submission/SWRL/).
34. W3C Rule Interchange Format. [cited 2011 May 3] (http://www.w3.org/2005/rules/).
35. SPARQLer, an online RDF Query platform on the public domain. [cited 2011 May 3] (http://www.sparql. org/ query.html).