The tissue microarray OWL schema: An open-source tool for sharing tissue microar

Hyunseok P. Kang,[1] Charles D. Borromeo,[2] Jules J. Berman,[3] and Michael J. Becich[2]

[1]Department of Pathology, Roswell Park Cancer Institute, Elm and Carlton St, Buffalo 14263, NY, USA,
[2]Department of Biomedical Informatics, University of Pittsburgh, 5150 Centre Avenue, Pittsburgh, PA 15232, USA
[3]No Affliations
Michael J. Becich: becich@pitt.edu
*Corresponding author

▶ This article has been cited by other articles in PMC.

Abstract

Background:

Tissue microarrays (TMAs) are enormously useful tools for translational research, but incompatibilities in database various researchers and institutions prevent the efficient sharing of data that could help realize their full potential. Description Framework (RDF) provides a flexible method to represent knowledge in triples, which take the form Su Object. All data resources are described using Uniform Resource Identifiers (URIs), which are global in scope. We (Web Ontology Language) schema that expands upon the TMA data exchange specification to address this issue sharing and integration.

Methods:

A minimal OWL schema was designed containing only concepts specific to TMA experiments. More general data incorporated from predefined ontologies such as the NCI thesaurus. URIs were assigned using the Linked Data fo

Results:

We present examples of files utilizing the schema and conversion of XML data (similar to the TMA DES) to OWL.

Conclusion:

By utilizing predefined ontologies and global unique identifiers, this OWL schema provides a solution to the limitati represents concepts defined in a localized setting. This will help increase the utilization of tissue resources, facilita translational research efforts.

Keywords: Ontology, OWL, tissue microarray

INTRODUCTION

Tissue microarrays (TMAs) are collections of hundreds of tissue cores arrayed into a single paraffin histology bloc containing between 100 and 1000 core tissue samples. Each TMA block can be sectioned and mounted onto glas

hundreds of nearly identical slides. By allowing researchers to simultaneously measure a marker in hundreds of s[...] a single slide, they conserve an enormous amount of time, money, and reagents.[1] Most importantly, they amplif[...] of limited tissue resources which are irreplaceable, enabling high-throughput controlled studies on large cohorts of [...] advantage of TMA experiments is that specimens from different donor tissue blocks are treated to identical incuba[...] temperatures, and washing conditions, standardizing the experiment and making it much easier to compare marke[...] different core sections.

As more and more studies are performed using high-throughput technologies such as DNA microarrays, TMA tech[...] proven to be a valuable tool for high-throughput validation of marker genes identified in these experiments.[2,3]

Because a single paraffin TMA block can be sectioned into nearly identical glass slides and dispensed to many dif[...] this technology also assists in collaboration and sharing of resources. However, this is also accompanied by increa[...] different laboratories may use different experimental protocols and instruments and capture data using different da[...] formats, and structures. There have been some efforts to address this by developing Common Data Elements (CD[...] tissue resources.[4,5] Although integrating the TMA data from different laboratories would dramatically increase th[...] experimental results and reduce redundant testing, it is difficult to put this into practice due to incompatibilities in d[...] laboratories using the different database or information systems that are available.[6] Compounding this is the fact [...] Laboratory Information Systems (LISs) are not prepared for this type of data.[7] Together, these factors prevent th[...] value of this technology from being realized. Expanding upon the TMA data exchange specification previously des[...] *al*,[8] we have designed an OWL (Web Ontology Language) schema that will help researchers share experimenta[...] TMA experiments.

METHODS

Before discussing the features of OWL, an introduction to Resource Description Framework (RDF) is necessary. C[...] RDF concepts. As such, an understanding of RDF is essential for describing OWL. RDF provides a flexible metho[...] knowledge by deconstructing it into small pieces called triples. Triples, also known as statements, take the form S[...] Object and can be regarded as being similar to simple sentences. For example, the statement "tissue core B15 is [...] RP2007-189" can be broken down into (tissue core B15) (Subject) (is derived from) (Predicate) (block RP2009-18[...] the subjects, predicates, and objects are names for resources which represent some entity, such as a person, wel[...] block, etc. These names are usually Uniform Resource Identifiers (URIs) and are global in scope, meaning that th[...] the same entity. The most well-known format for URIs is the URL, although it can be anything that the creators of [...] such as an International Standard Book Number (ISBN). Objects can also be literals such as numbers or text strin[...] (block RP2009-258) (Subject) (number of cores) (Predicate) (100) (Object). Subjects and objects can be instance[...] while predicates are RDF properties. RDF classes, properties, and the relationships between them are defined in [...] documents.

OWL extends the expressivity of the RDF schema by describing more complicated relationships. OWL allows a sc[...] connect two concepts in an inverse relationship. For example, the predicate "has child" in the simple relationship ([...] (Subject) (has child) (Predicate) (child Child1) has a natural inverse relationship "has father". The inverse relations[...] reversing the subject and object: (child Child1) (has father) (father Dad1). Although RDF allows one to describe bo[...] separately, it is OWL that allows the two to be described as inverses of one another. OWL also allows for automat[...] data since it is more semantically stringent than XML or RDF.

The goal of this effort was to design a minimal OWL schema that defined classes and properties specific to descri[...] experiments performed on them, adhering to the spirit of the guiding principles set forth by Berman *et al*.[8] This re[...] of data elements described and simplifies the document. In many cases, more general concepts are described in [...]

which users can utilize. For example, instead of defining classes such as "#pathology_report" or "#organism" in ou suggest that users generate instances of classes defined in external ontologies, such as the NCI Thesaurus (http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl) which includes these classes. In addition to simplifying the d provides a mechanism to integrate data from these OWL documents and other documents utilizing those ontologie

TMAs represent data obtained from a wide variety of diseases and experimental conditions. However, there are s concepts unifying the artifacts generated as part of the TMA manufacturing process. For the purposes of this exer the central issue of relating TMA concepts (i.e. a core to a block, a core to a slide, etc.) to one another in a seman manner. We avoided attempting to describe concepts in huge topics such as assays, diagnosis, and anatomy, ma have their own large-scale vocabulary or ontology projects.[9] Although TMAs are ultimately used for disease rese these topics are far too broad to include in a singular OWL scheme, in addition to not being limited to TMAs. There scope of the schema focused on concepts central to TMAs. We also wanted to present a focused and pragmatic c others can follow to create their own instances of OWL files from their data. This is not intended to be an exhaustiv

There are several methods available for extracting TMA data from data sources (i.e. database, files, etc.). In gene approach should follow these steps:

Collect the data to be represented in the OWL file (e.g. SQL query);

Store the collected data in a well-defined structure (e.g. an XML file);

Convert the well-defined structure into a set of OWL files (e.g. an XSLT transformation).

The actual methods employed to implement these steps will depend largely on the starting data source format. Th involves collecting the data in a well-defined XML file. Next, convert the XML file to an OWL file using XSLT. An X provides the best method for converting data into OWL. Whereas the conversion process is fairly straightforward, creating identifiers for the objects within the TMA OWL file is more involved.

TMAs have a long life and are reused thousands of times. In addition, the samples from a TMA are typically dispe institutions/laboratories for various experiments. The dispersion of samples may occur over the course of several essential to have a centralized identifier to describe the TMA. Unfortunately, these centralized identifiers are a sou debate within the RDF community. For openness and ease of use, we have elected to use the Linked Data[10] for from the Linking Open Data (LOD) project. The LOD attempts to link data from a wide variety of data sources inclu social networking, and life sciences. To use the Linked Data format, one must define a set of URIs for the data. Or that the URIs are resolvable and supply useful information about the data represented by the URI. The Linked Dat impose any specific format, but does supply some suggestions that are given below:

Define URIs in an HTTP namespace under one's control. Do not define them in someone else's namespace.

Keep URIs stable and persistent. Changing URIs will break any already established links, so it is advisable to dev thought to them at an early stage.

In general, one needs to use a primary key inside a URI to make sure that each URI is unique. Whenever possible meaningful inside your domain. For example, when dealing with books, making the ISBN number part of the URI i the primary key of an internal database table. If one is representing a TMA ordered from National Mesothelioma V use the NMVB identifier in the URI.

For example, assume one's domain is *www.institutionXYZ.org* and within the organization there is an RDF data fil TMAs that can be found at http://www.institutionXYZ.org/tma/rdf. The URI of a TMA with identifier 123456 would th

http://www.institutionXYZ.org/tma/rdf/123456

RESULTS

The OWL schema document can be found at *http://bioportal.bioontology.org/ontologies/42764* and is also provide (tma_minimal.owl) with this article. The classes and properties defined in this document are listed in Table 1. As s paired properties describing the bidirectional relationship between tma, block, slide, and core. The "top-down" prop (includes_block, block_includes_core) should be used in OWL documents that include the entire hierarchical struc the "bottom-up" properties (included_in_tma, cut_from_block) can be used when a resource refers to a parent blo in a different document. repository_product and experimental_component are abstract parent classes which are n that the various classes and properties are used according to OWL syntax rules.

Figure 1 shows an example of a valid OWL file utilizing this schema. In this example, the first and second lines sp document is XML, and that it is an RDF resource, respectively. Lines 3, 4, and 5 assign shorthand prefixes for the documents that will be referred to. When one of these prefixes is followed by a colon, it can be expanded into the namespace document. Line 6 specifies the base URI of this document. Resource identifiers that are simply prece assumed to belong to this base URI and expanded accordingly. Sample expansions are listed in Table 2.

There are some shorthand conventions in RDF/XML that one needs to be aware of to read it correctly. Each block by opening and closing Class names ('<tma:tma.... </tma:tma>' – these are separated by lines in this case for rea instance (the URI of which is declared in "rdf:about") of that class. The opening and closing tags also encompass object pairs that share the same subject (the resource following "rdf:about"). Hence, the first block represents the in Table 3. The first triple states that "type" of the resource "tma/000001" is the "tma" class (creating an instance o process), the second and third describe the "title" and "creator", and the fourth states that it includes the resource "http://www.institutionXYZ.org/block/RP2008-325". In this case, the TMA and block URIs simulate the existence of respective addresses, while those for slide, core_in_block, and core_on_slide are within the example file itself. Th line illustrates how to use the NCI ontology to assign a diagnosis to a resource. The clinical_annotation property w similar fashion, most likely by pointing to a separate resource that includes these data, or generating a blank node necessary property/value pairs. The latter option requires a separate schema for the clinical data elements, which project specific but could be updated and merged with other schema as the field matures.

In the supplementary data, we have included example files (which can be viewed using any text viewer) for using (tma_rdf.xsl) to convert data in an XML file (mvbtma1.xml) from the NMVB project[11] to RDF (mvbtma1.rdf). Thes simulate the existence of RDF files that catalog the various identifiers. These, and the example described in Figure exhaustive explorations, but are rather intended to demonstrate some of the ways in which the schema can be util

As mentioned previously, TMAs are a long-term experimental resource. As such, they potentially represent the sa several experiments in a wide variety of assays. High-throughput assays like genomic microarrays are commonly researchers to explore a wide range of biological questions and samples included in TMAs are frequently used in experiments. To further demonstrate the utility of the TMA OWL file, we present an excerpt of the TMA OWL conn slide to its microarray experiment results. The results are represented using the MGED Ontology,[12] an OWL ont provide well-defined microarray results. In this case, the MGED Ontology is applied to an existing dataset from the (https://cabig.nci.nih.gov/tools/caArray).[13] The MGED OWL data are presented here for illustrative purposes but derived from the caArray data.

Figure 2 demonstrates how to connect data from an example TMA OWL file to a sample MGED OWL file. In the T relationship is established between the core_on_slide (#slide/example_slide1/D4) and the experiment conducted i

file (http://www.the_url_here.org/microarray_data/1015897590474029.owl#experiment_tumorCel01). This same a
used to connect the TMA OWL file to an RDF or OWL file containing instances of other types of assay results.

DISCUSSION

Various efforts have been made to define data exchange standards for TMA data. One obstacle in doing this is the
and vast diversity of clinical and histopathologic data that can accompany TMAs. In order to address these issues
described an XML TMA data exchange specification that focused on a generic, extensible format that was easy to
have been some implementations of this, in one case building upon the original specification to design a system m
the needs of a specific collaborative tissue resource.[14–16] In addition, Lee *et al*, designed a more elaborate syst
Microarray-Object Model (TMA-OM). They then implemented it in a web-based database application called Xperan
as TMA-TAB, a spreadsheet-based data exchange format integrated with the database.[18] This data model, while
comprehensive, is also very complex, which creates a barrier to entry and limits users to the applications that are
these can be useful as standalone programs and forms, they are relatively inflexible and cannot utilize externally c
vocabularies, which also restricts interoperability with other applications based on these standards.

The limitation of XML in the context of data sharing is that although the extensible tags provide a format for repres
humans can understand, there must be some agreement on these tags before they are useful for sharing data. In
requirement for standards still exists; this decision has been left up to the users. OWL, on the other hand, is a form
preexisting standards and definitions. In general, XML files represent concepts defined by their authors in a localiz
institution, laboratory, a particular application, etc). As a result, the precision of these definitions depends on the a
invested by the authors. In many cases, ambiguity in XML arises because XML is not designed to stringently defin
primarily intended to exchange information. On the other hand, RDF is designed to reduce ambiguity while definin
concepts within a domain. In XML, a tag (or property) is simply a string that denotes a data point:
<average_count>3.5</average_count>. In this example, <average_count> is the XML property and "3.5"is the dat
XML will not provide an adequate definition describing the concepts. In this example, average_count represents a
concept with a calculated value of 3.5. Without a clear definition of "average_count", another person utilizing the X
to know how the "average_count" was calculated. The average_count could represent a simple mean, a weighted
from an assay's software package. (Another example of this ambiguity exists in the financial world. Most financial
price/earnings ratios yet they all have varying methods for calculating the ratio.[19]) The lab creating the XML file
what the average_count represents, but the precise definition is lost on a third party.

OWL is designed to reduce ambiguity by providing facilities for defining concepts within files and for defining conc
use in files. In addition, OWL can import vocabularies and concepts from RDF structures. There are several gener
structures available to provide the basis for new OWL or RDF (RDF/OWL) files:

- Dublin Core Metadata Initiative (dcmi)
- Friend of a Friend (foaf)
- Simple Knowledge Organization System (skos)

These RDF structures provide support for representing bibliographic information, basic labels, information about p
information about concepts. In addition, there are several organizations creating RDF (or OBO/OWL) structures or
topics including diseases, taxonomy, phenotypes, etc. These structures are collected by different repositories sucl

- National Center for Biomedical Ontology
- Open Biomedical Ontologies

Centralizing the RDF/OWL structures allows the organizations to define concepts and refine these definitions to re These repositories also provide versioning capabilities to support refinements and corrections. By centralizing RDI an organization can create permanent definitions for concepts. This encourages adoption of the standardized cond existing users and new adopters. Unlike an assay representing a snapshot of data at a particular point in time, TM permanent real-life entity. Consequently, a data structure that defines concept definitions in a permanent and univ RDF/OWL) is preferable to a more transient and localized data structure (e.g. XML). Global unique identifiers and definitions allow true portability, freeing up users to exchange data without prior agreement on data elements, etc. suitable schema, one institution or investigator may make their data available in an OWL document that other indiv utilize. This makes OWL an ideal candidate for designing a framework for the sharing of clinical and test data on T above, such a data sharing framework will create substantial value by greatly facilitating collaborative studies utiliz

Footnotes

REFERENCES

1. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, et al. Tissue microarrays for high-throughput molecular profiling *Med.* 1998;4:844–7. [PubMed: 9662379]

2. Han H, Bearss DJ, Browne LW, Calaluce R, Nagle RB, Von Hoff DD. Identification of differentially expressed genes in pancreatic cancer cel *Cancer Res.* 2002;62:2890–6. [PubMed: 12019169]

3. Mousses S, Bubendorf L, Wagner U, Hostetter G, Kononen J, Cornelison R, et al. Clinical validation of candidate genes associated with pro in the CWR22 model system using tissue microarrays. *Cancer Res.* 2002;62:1256–60. [PubMed: 11888886]

4. Mohanty SK, Mistry AT, Amin W, Parwani AV, Pople AK, Schmandt L, et al. The development and deployment of Common Data Elements f translational research in cancer - an emerging standard based approach for the Mesothelioma Virtual Tissue Bank. *BMC Cancer.* 2008;8:91. [[PubMed: 18397527]

5. Patel AA, Kajdacsy-Balla A, Berman JJ, Bosland M, Datta MW, Dhir R, et al. The development of common data elements for a multi-institut bank: The Cooperative Prostate Cancer Tissue Resource (CPCTR) experience. *BMC Cancer.* 2005;5:108. [PMCID: PMC1236914] [PubMed:

6. Kajdacsy-Balla A, Geynisman JM, Macias V, Setty S, Nanaji NM, Berman JJ, et al. Practical aspects of planning, building, and interpreting t Cooperative Prostate Cancer Tissue Resource experience. *J Mol Histol.* 2007;38:113–21. [PubMed: 17318343]

7. Becich MJ. The role of the pathologist as tissue refiner and data miner: The impact of functional genomics on the modern pathology laborato pathology informatics and bioinformatics. *Mol Diagn.* 2000;5:287–99. [PubMed: 11172493]

8. Berman JJ, Edgerton ME, Friedman BA. The tissue microarray data exchange specification: A community-based, open source tool for shari *BMC BMC Med Inform Decis Mak.* 2003;3:5.

9. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: Ontologies and integrated data resources at the click of a mouse. 2009;37:W170–3. [PMCID: PMC2703982] [PubMed: 19483092]

10. How to Publish Linked Data on the Web. Available from: http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ [last cited on 2010 N

11. Amin W, Parwani AV, Schmandt L, Mohanty SK, Farhat G, Pople AK, et al. National Mesothelioma Virtual Bank: A standard based biospec resource to enhance translational research. *BMC Cancer.* 2008;8:236. [PMCID: PMC2533341] [PubMed: 18700971]

12. Whetzel PL, Parkinson H, Causton HC, Fan L, Fostel J, Fragoso G, et al. The MGED Ontology: A resource for semantics-based descriptio experiments. *Bioinformatics.* 2006;22:866–73. [PubMed: 16428806]

13. Yu YP, Landsittel D, Jing L, Nelson J, Ren B, Liu L, et al. Gene expression alterations in prostate cancer predicting tumor aggression and malignancy. *J Clin Oncol.* 2004;22:2790–9. [PubMed: 15254046]

14. Berman JJ, Datta M, Kajdacsy-Balla A, Melamed J, Orenstein J, Dobbin K, et al. The tissue microarray data exchange specification: Impler Cooperative Prostate Cancer Tissue Resource. *BMC Bioinformatics.* 2004;5:19. [PMCID: PMC373442] [PubMed: 15040818]

15. Nohle DG, Ayers LW. The tissue microarray data exchange specification: A document type definition to validate and enhance XML data. *B Mak.* 2005;5:12. [PMCID: PMC1183103] [PubMed: 15871741]

16. Nohle DG, Hackman BA, Ayers LW. The tissue micro-array data exchange specification: A web based experience browsing imported data *Mak.* 2005;5:25. [PMCID: PMC1208890] [PubMed: 16086837]

17. Lee HW, Park YR, Sim J, Park RW, Kim WH, Kim JH. The tissue microarray object model: A data model for storage, analysis, and exchan experimental data. *Arch Pathol Lab Med.* 2006;130:1004–13. [PubMed: 16831026]

18. Song YS, Lee HW, Park YR, Kim DK, Sim J, Kang HP, et al. TMA-TAB: A spreadsheet-based document for exchange of tissue microarray microarray-object model. *J Biomed Inform.* 2009 In press.

19. Why do I often see different PE ratios on different sites and in different publications? Conscious Investing Knowledge Base. Available from http://www.stablegrowthcompanies.com/knowledgemanager/questions//Why+do+I+often+see+different+PE+ratios+on+different+sites+and+in-[last cited on 2010 Mar12]

Figures and Tables

Table 1

RDF classes and properties

Resource type	Resource
Class	repository_product
	experimental_componen
	tma
	block
	slide
	core_in_block
	core_on_slide
Object property	includes_block
	included_in_tma
	block_includes_core
	included_in_block
	slide_includes_core
	affixed_to_slide
	includes_slide
	cut_from_block
	derived_from_core
	donor_block
Datatype property	status
	level
	size
	thickness
	core_spacing
	location
	diagnosis
	clinical_annotation
	drill_site
	protocol
	report_link
	control
	result
	repository
	assay

Table 2

Sample expansions

Shorthand	Expanded
#slide058	http://www.the_url_here.org/tma_example1.rdf#slide058
tma:block	http://bioontology.org/ontologies/tma-minimal#block
dc:title	http://purl.org/dc/elements/1.1#title

```
<?xml version="1.0"?>
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1#"
        xmlns:tma="http://bioontology.org/ontologies/tma-minimal#"
        xml:base="http://www.the_url_here.org/tma_example1.owl#">

<tma:tma rdf:about="http://www.institutionXYZ.org/tma/000001">
   <dc:title>All-Purpose Tissue Array</dc:title>
   <dc:creator>Michael J. Becich</dc:creator>
   <tma:includes_block rdf:resource="http://www.institutionXYZ.org/block/RP2008-325"/>
</tma:tma>

<tma:block rdf:about="http://www.institutionXYZ.org/block/RP2008-325">
   <tma:includes_slide rdf:resource="#slide058"/>
   <tma:block_includes_core rdf:resource="#coreD4"/>
</tma:block>

<tma:slide rdf:about="#slide058">
   <tma:slide_includes_core rdf:resource="#slide058-D4"/>
</tma:slide>

<tma:core_in_block rdf:about="#coreD4">
   <tma:donor_block rdf:resource="http://www.institutionXYZ.org/block/RP2007-228"/>
   <tma:repository>Generic Tissue Bank</tma:repository>
   <tma:drill_site>78,90</tma:drill_site>
   <tma:diagnosis rdf:about="#diagnosis1" rdf:type="http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#Prostate_Carcinoma"/>
</tma:core_in_block>

<tma:core_on_slide rdf:about="#slide058-D4">
   <tma:location>D4</tma:location>
   <tma:derived_from_core rdf:resource="#coreD4"/>
</tma:core_on_slide>

</rdf:RDF>
```

Figure 1

Example OWL file; this example uses the TMA OWL schema to describe the basic elements of a simplified TMA

Table 3

List of triples

Subject	Predicate	Object
tma/000001	owl:Class	tma:tma
tma/000001	dc:title	"All-Purpose Tissue Array"
tma/000001	dc:creator	Michael J. Becich
tma/000001	tma:includes_block	block/RP2008-325
block/RP2008-325	owl:Class	tma:block
block/RP2008-325	tma:includes_slide	#slide058
block/RP2008-325	tma:block_includes_core	#coreD4
#slide058	owl:Class	tma:slide
#slide058	tma:slide_includes_core	#slide058-D4
#coreD4	owl:Class	tma:core_in_block
#coreD4	tma:donor_block	block/RP2007-228
#coreD4	tma:repository	"Generic Tissue Bank"
#coreD4	tma:drill_site	"78,90"
#coreD4	tma:diagnosis	#diagnosis1
#diagnosis1	rdf:type	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#Prostate_Carcinoma
#slide058-D4	owl:Class	tma:core_on_slide
#slide058-D4	tma:location	"D4"
#slide058-D4	tma:derived_from_core	#coreD4

Figure 2

MGED example; this example illustrates one way to connect a file describing a TMA to a data file using a different

---