KnowledgeWiki: An OpenSource Tool for Creating Community-Curated Vocabulary, with a Use Case in Materials Science

Nishita Jaykumar
Kno.e.sis Center
Wright State University
Dayton Ohio, USA
nishita@knoesis.org

PavanKalyan Yallamelli
Kno.e.sis Center
Wright State University
Dayton, OH, USA
pavany@knoesis.org

Vinh Nguyen
Kno.e.sis Center
Wright State University
Dayton, OH, USA
vinh@knoesis.org

Sarasi Lalithsena
Kno.e.sis Center
Wright State University
Dayton, OH, USA
sarasi@knoesis.org

Krishnaprasad Thirunarayan
Kno.e.sis Center
Wright State University
Dayton, OH, USA
tkprasad@knoesis.org

Amit Sheth
Kno.e.sis Center
Wright State University
Dayton, OH, USA
amit@knoesis.org

Clare Paul
Air Force Research Laboratory
Wright-Patterson AFB
Dayton, OH, USA
clare.paul@us.af.mil

ABSTRACT

Resource Description Framework (RDF) datasets can be created by transforming structured databases, extracting the triples from semi-structured and unstructured sources, crowd-sourcing, or by integrating the existing datasets. The reliability and quality of these datasets can be improved by the participation of domain experts via a special purpose tool or a crowd-sourced application. Wikidata and Semantic MediaWiki are platforms which facilitate this kind of crowd-sourced data curation.

We present our system, KnowledgeWiki, which is built upon the existing Semantic MediaWiki. We develop a novel extension by adopting the singleton property data model in our KnowledgeWiki. This extension allows various kinds of metadata about the RDF triples to be created in the Wiki. We combine this extension with other extensions such as *semantic forms* to provide a user-friendly, Wiki-like interface for domain experts with no prior technical expertise to easily curate data. We also present our new enhancement to Semantic Mediawiki, which facilitates importing existing RDF datasets into the wiki-based curating platform based on the singleton property approach, that preserves the provenance of individual triples. We also describe how it is being used by the materials science community to create and curate consolidated vocabularies.

Keywords

Linked Data application, Semantic MediaWiki, Singleton property, KnowledgeWiki, Wikidata, Open source, Semantic Web, Provenance metadata, Materials Science

WWW2016 Workshop: Linked Data on the Web (LDOW2016)

1. INTRODUCTION

The White House's Materials Genome Initiative (MGI) seeks to substantially improve the process of materials discovery and development, and shorten the time to deployment. One of the main goals of MGI is to develop solutions which provide broader access to scientific data. This allows materials scientists to integrate each other's data and facilitate communication among scientists working in different stages of the materials development continuum.

A key challenge in data integration is dealing with the heterogeneity of data in the Semantic Web community [5]. Standardized vocabularies are widely used as a component of a toolkit to solve data heterogeneity issues. They play the role of a shared language, which facilitates easy communication and information exchange among people within the community.

While there exist disparate sets of vocabularies developed for the materials domain, there is no easy mechanism to curate these vocabularies by the domain scientists spreading all over the world. The lack of such a mechanism prevents the wider adoption of these vocabularies. Further, existing vocabularies lack the support to capture provenance metadata. Provenance metadata is crucial for data integration from disparate sources in order to determine the trustworthiness of the data and also to give proper credit to the creators of the data. Provenance metadata would increase interoperability, discoverability, reliability as well as reproducibility for scientific discourse and evidence-based knowledge discovery [7].

Crowd-sourcing is a cost-effective and reliable approach to easily distribute a task among a potentially large group of contributors. The Semantic Web community has used crowd-sourcing techniques for knowledge acquisition tasks, including vocabulary development [15]. Semantic Mediawiki [10] is one such platform that can be used for crowd-sourced vocabulary curation. However, it lacks built-in support to capture the provenance metadata of RDF triples.

To the best of our knowledge, Wikidata is the only crowd-sourced knowledge acquisition platform which supports incorporating such provenance metadata [6]. It is being used to create and curate structured online database for Wikipedia. Wikidata allows editors to annotate attribute-value pairs using qualifiers and references as a way to support metadata. For this purpose, Wikidata uses an auxiliary node in a way which is similar to a blank node as discussed in [6]. However, the RDF representation of this data model is not intuitive and does not map to the standard RDF triples.

In this work, we try to address the aforementioned challenges with our tool, KnowledgeWiki. We adopt the singleton property template approach developed by Nguyen et al. A *singleton property* is defined as a unique property instance representing a newly established relationship between two existing entities in one particular context [13]. This approach offers a concise representation for RDF statements about statements, with a formal semantics for an accurate interpretation across applications, tools, and datasets. Recent studies [7, 8] reported that the singleton property approach offered the most concise representation on a triple level. Therefore, we adopted the singleton property in our data model for the extension.

In our KnowledgeWiki, by using Semantic MediaWiki as the curation platform and embedding the singleton property approach into Semantic MediaWiki as an extension, we intend to capture metadata of RDF triples such, as provenance information. We combine this extension with other extensions such as semantic forms to provide a user-friendly interface for data collection. The goal of KnowledgeWiki is to provide a wiki-based platform for assisting in the creation and curation of vocabularies in the materials science domain. Our main contributions in this work are three-fold:

1. **A Singleton Property Template Extension to Semantic Mediawiki to capture the metadata of RDF triples**
   We incorporate the singleton property approach for RDF data representation into SMW. SMW takes a simple straightforward approach to represent triples. However, as mentioned earlier provenance information is essential, but missing. The singleton property template data model, which is an improvement over the standard reification approach, is suitable for representing metadata about the data in materials science [7].

2. **An algorithm to identify Singleton Property Templates from RDF datasets and to support importing RDF datasets into the wiki**
   We propose an algorithm to automatically identify the regular and singleton property templates for every entity from a given RDF dataset. All the templates related to a single entity will be presented in the same wiki page. This page allows the domain experts to curate the content.

3. **Demonstration of the use of this extension in the materials science domain**
   We import three vocabularies extracted from the ASM Handbook Volume 21, MIL-HDBK-5, and MIL-HDBK-17 from the materials science domain. We adopted the singleton property template extension for representing the provenance of the vocabularies. The three vocabularies have been curated in our KnowledgeWiki. For representing the provenance information, such as source and license, we reuse the existing vocabularies such as SKOS[1], Dublin Core[2], and QUDT (Quantities, Units, Dimensions and Data Types)[3].

The remainder of the paper is organized as follows: Section 2 discusses the context for the research, Section 3 discusses related work, Section 4 outlines our approach, Section 5 discusses our materials science use case. Finally, Section 6 and Section 7 discusses the future work and conclusion of our work respectively.

2. CONTEXT FOR THIS RESEARCH

The availability of a crowd-sourcing tool is crucial to achieve the requirements proposed by MGI. If made widely available, disparate sources of materials data also could be inventoried to identify gaps in data and to limit redundancy in research efforts. According to MGI, "To benefit from broadly accessible materials data, a culture of data sharing must accompany the construction of a modern materials data infrastructure that includes the software, hardware, and data standards necessary to enable discovery, access, and use of materials science and engineering data. The system should be available to house, search, and curate materials data generated by the community. The initiative also states that the community-developed standards should provide the format, metadata, data types and criteria necessary for interoperability and seamless data integration" [19].

The Air Force Research Laboratory (AFRL) proposed to work towards the goals of the Materials Genome Initiative, mainly in the context of bringing the materials science community together to collaboratively create and curate a consolidated vocabulary for materials science. For this purpose, AFRL and its partners provided us with initial legacy data. This data is comprised of materials science dictionary terms from three vocabularies: ASM Handbook Volume 21 [9], MIL-HDBK-5 [14], and MIL-HDBK-17 [12]. The materials science community was in need of a system that could: (1) consolidate these vocabularies, (2) represent each dictionary term and its metadata such as, provenance appropriately, and (3) provide the community broader and easier access to materials science terms and definition.

These requirements which were presented to us by AFRL are fulfilled via our KnowledgeWiki tool. We present its development for fulfilling the requirements along with some preliminary results in the rest of the paper.

3. RELATED WORK

The Semantic Web community leveraged crowd-sourcing techniques for the purpose of data curation [15]. Semantic wikis, which enhance traditional wikis with structured knowledge representation, have been used to facilitate collaborative development of ontologies and knowledge bases [3]. Semantic MediaWiki [10] is a widely known semantic wiki and it has been used to capture semantic data in variety

[1] https://www.w3.org/2009/08/skos-reference/skos.html
[2] https://www.w3.org/TR/prov-dc/
[3] http://qudt.org/

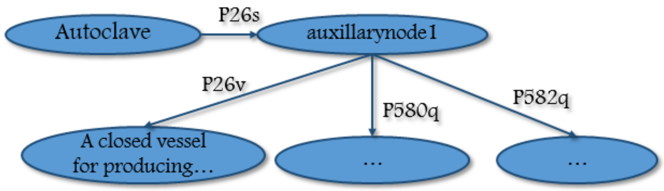


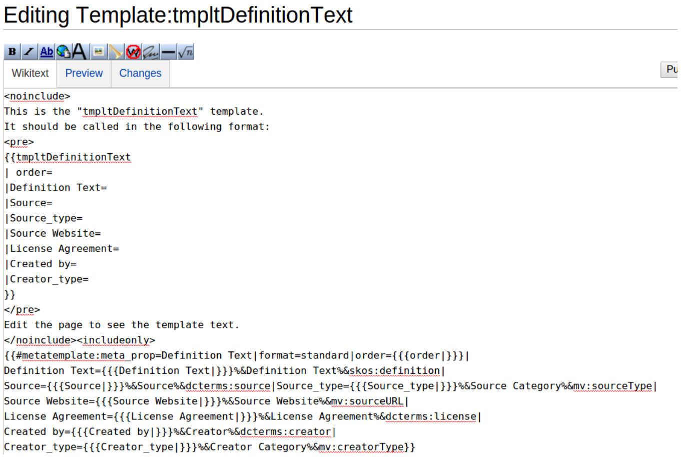
Figure 1: RDF graph representation from Wikidata for the definition of a term with source and license as provenance metadata information. P26 is a property definition text, P580 is a property source and P582 is a property license.

of areas including healthcare [2], energy, media, and air navigation. Semantic Mediawiki is currently being used in over 300 public wikis around the globe. Some of the other examples of Semantic wikis include: OntoWiki [1], IkeWiki [16], SweetWiki [4], and Acewiki [11]. More details on different kinds of wikis can be found in [3]. However, none of these wikis have a simple way to incorporate metadata into their data models, a challenge our work addresses.

As mentioned earlier Wikidata is the only crowd-sourced application which provides the capability to incorporate metadata. Wikidata [18] is a prominent community-oriented effort to create and curate an online, structured knowledge base that can be used by every language version of Wikipedia. For example, the definition of a term can be further accompanied by its source and license information. In order to represent this in RDF, Wikidata uses auxiliary node in a way that is similar to a blank node. Figure 1 shows an example of the modeling with an auxiliary node. Here, property P26 (Definition Text) is broken into two properties to use the auxiliary node to link to the context information. As they mentioned in their paper, this led to Wikidata properties not directly corresponding to properties in RDF.

The widely-known techniques for incorporating metadata into the RDF data model are: (1) reification, (2) n-ary relation, (3) the singleton property, and (4) a named graph. In the recent work by Hernández et al., [8] the authors compared these techniques for reifying RDF triples, with the goal of representing Wikidata as RDF, which would allow legacy Semantic Web languages, techniques, and tools to be used for Wikidata. They reported that the singleton property approach offered the most concise representation on the triple level. Similarly, this approach also offers the most compact dataset according to the experimental comparison in the PubChem dataset [7]. Therefore, we adopted the singleton property approach [13], which was originally developed by our group, to represent the metadata about the RDF triples.

Existing work, including Wikidata, lacks support for importing an existing RDF data set with metadata. Even though the SMW extension RDFIO[4] provides the capability to import arbitrary RDF triples into the wiki, it does not have the capability to handle an RDF data set with metadata information. The proposed RDF import extension automatically identifies the RDF subgraph structure of the dataset and imports the dataset while preserving the provenance of individual triples.


Editing Template:tmpltDefinitionText

```
<noinclude>
This is the "tmpltDefinitionText" template.
It should be called in the following format:
<pre>
{{tmpltDefinitionText
| order=
|Definition Text=
|Source=
|Source_type=
|Source Website=
|License Agreement=
|Created by=
|Creator_type=
}}
</pre>
Edit the page to see the template text.
</noinclude><includeonly>
{{#metatemplate:meta_prop=Definition Text|format=standard|order={{{order|}}}|
Definition Text={{{Definition Text|}}}%&Definition Text%&skos:definition|
Source={{{Source|}}}%&Source%&dcterms:source|Source_type={{{Source_type|}}}%&Source Category%&mv:sourceType|
Source Website={{{Source Website|}}}%&Source Website%&mv:sourceURL|
License Agreement={{{License Agreement|}}}%&License Agreement%&dcterms:license|
Created by={{{Created by|}}}%&Creator%&dcterms:creator|
Creator_type={{{Creator_type|}}}%&Creator Category%&mv:creatorType}}
```

Figure 2: A sample of the singleton property template definition in KnowldgeWiki

4. APPROACH

In this section, we first describe the overall architecture of the system. Next, we describe how we developed the new extension with singleton property templates for representing the provenance metadata of the triples. Then, we describe the algorithm for identifying the singleton property templates for any given RDF dataset. The overall architecture of KnowledgeWiki is shown in Figure 3.

4.1 Overall Architecture

The following sections describes the architecture overview of our system (see Figure 3). We first explain how the data is collected via the existing semantic forms and how the singleton property template is integrated into the SMW. Next, we describe our new data representation module for SMW. We also describe how each entity is processed and, how the triples are created and, how each entity is represented on a wiki page. Finally, we discuss how more complex operations, such as CSV data import and RDF import/export are performed.

Data Collection. Templates are an integral part of Semantic MediaWiki and the simplest way to give input. Each template is defined as a set of field and value pairs. Each field in the template is mapped to a separate semantic property. Semantic properties in SMW are used to express binary relationships between semantic entities. The field values can be set to a number of predefined datatypes, such as Text, URL, Page, Boolean, Number, etc.

For example, a *regular template* which provides the text definition for a given term is defined as follows:

```
{{Definition Text
| Definition = }}
```

For the term *Autoclave*, the page would be titled *Autoclave* and this page would contain the regular template "Definition Text" along with the property name and its value as follows:

```
{{Definition Text
| Definition = A closed vessel for producing ..}}
```

[4] https://www.mediawiki.org/wiki/Extension:RDFIO

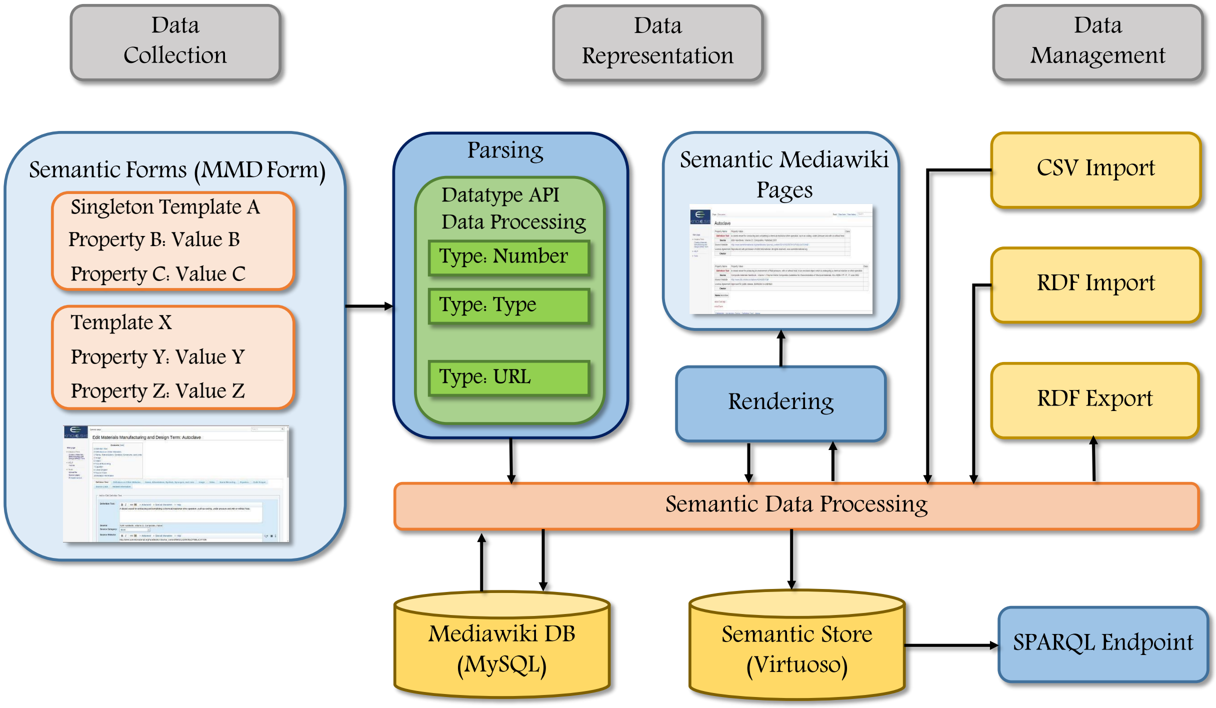


Figure 3: KnowledgeWiki's Overall Architecture

Templates allow users to specify annotations in the wiki without the need of having to learn any kind of syntax. For example, the field "Definition" is mapped to the property *mv:definition*[5], and this property holds the value "A closed vessel..." as asserted in the following triple:

```
Autoclave mv:definition "A closed vessel..."
```

This is the standard triple format; however, if we want to describe the metadata about this triple, SMW does not support it. For example, this triple was taken from the ASM handbook glossary, but the existing template does not allow for such an annotation to be represented.

To overcome this, we modify the data model at the template level. We developed a special purpose template called "singleton property template", that is generic and allows for any kind of data annotation such as provenance, access control, or spatio-temporal information. We took the built-in template and enhanced it to support metadata information. We refer to this extension as the singleton property template[6]. This enhancement, when used in semantic forms for data collection from domain experts, will allow users to specify assertions about the entity within the existing infrastructure. Below is an example of how this information will be represented using our singleton property template:

```
{{Definition Text
| Definition = A closed vessel for producing..
| Source = ASM handbook}}
```

Templates are used to create forms. A semantic form is a collection of related templates. This singleton property template is included in our "Materials Manufacturing and Design Form" depicted in Figure 5. We provide the details on the development of this template in Section 4.2. We use the semantic forms provided by the SMW to allow users to add the terms and/or modify the element details.

One of the goals of our system is to promote collaboration and facilitate experts to edit data easily using our tool. By using the semantic forms, domain experts can easily edit and create new data on the wiki. In order to foster collaboration, making this wiki-based form easy to use is crucial. For this purpose, SMW also provides an "Edit with Form" option, which allows users to edit each page via user-friendly forms as depicted in Figure 5.

KnowledgeWiki provides the following features: (1) support for the singleton property approach to capture the metadata, and (2) support for adding typed information of the modeling elements.

Data Representation. Here we discuss (1) how we use the singleton property templates to represent data about an entity in the semantic store and also (2) how each entity is represented in a wiki page.

The first task is accomplished by the *parsing* phase. Once the data is collected from users via the form, the Datatype API is responsible for type checking and for mapping the imported properties within the wiki for appropriate representation. The definition of a sample singleton property template is shown in Figure 2. Within each template definition, there exists a hidden wiki snippet with a magic word. Each magic word is associated with a set of parameters corresponding to the field-value pairs of the template. The magic word and the associated parameters are processed for creating triples in our extension. The triples generated

[5] mv: Materials Vocabulary - MatVocab
[6] http://matvocab.org/wiki-dev/index.php/Special:CreateMetaTemplate

are inserted into the semantic store, and can be queried via SPARQL. The semantic store can be chosen from a variety of well-known engines: 4store, Blazegraph, Sesame, Virtuoso, etc. In the work by Hernández et al., they report that singleton properties worked best with the Virtuoso triple store and, hence, for our KnowledgeWiki we chose Virtuoso. MySQL DB stores most of the schema-level information such as, template, property and, form names and Virtuoso has the semantic data as triples.

The second task of representing entities on a wiki page is accomplished by the *rendering* phase. In addition to the creation of triples, a page is created for each entity on the wiki. For this page, our extension renders the data from the templates associated with each entity. Finally, a wiki page is created for each entity with the term name as the page title within the main namespace.

Data Management. This phase involves the semantic content management. This phase performs complex operations that are conducted within the wiki (such as semantic data processing). The wide range of capabilities that KnowledgeWiki provides, allow for various tasks such as legacy data processing, RDF data import, RDF data export, and so on.

4.2 Singleton Property Template Extension

As mentioned earlier in Section 4.1, templates are an integral part and the simplest way of including markup in the wiki. We introduced the singleton property template in Section 4.1. and in this section we describe the implementation details. The singleton property template is an extension that we implemented for representing the metadata of RDF triples such as provenance. Since we developed a new template for the purpose of disambiguation, we name the traditional SMW template as a *regular template* and our new template as a *singleton property template.* Every singleton property template contains a set of field-value pairs:

```
{{Singleton Property Template name
| property_1 = value_1
| property_2 = value_2
| property_3 = value_3
}}
```

Each field is mapped to a semantic property. Within each template, a property selected for instantiating singleton properties and is specified in the MagicWord parameters associated with this template. A sample definition for the singleton property template is shown in Figure 2. We introduce the MagicWord "#singletontemplate" for processing the singleton property template, for example:

#singletontemplate:singleton_prop=property 1.

When creating the singleton property template, the user gets to select the property to instantiate the singleton property. Each template can contain only one singleton property. This singleton property will bear the metadata of the RDF triples and the semantic forms can contain more than one singleton or regular template. In this case, the singleton property will be created for the property "property 1." It is associated with other meta properties such as "property 2" and "property 3." The singleton property template above will be mapped to this set of triples as seen in Table 1.

Table 1: The set of singleton triples generated for the page SomePageTitle with the singleton property template described in Section 4.2

SomePageTitle	singletonProperty#n	value_1
singletonProperty#n	singletonPropertyOf	property_1
singletonProperty#n	property_2	value_2
singletonProperty#n	property_3	value_3

Table 2: The set of singleton triples generated for the page Autoclave with the singleton property template described in the example

Autoclave	hasDefinition#1	"A closed vessel.."
hasDefinition#1	singletonPropertyOf	skos:definition
hasDefinition#1	dcterms:source	"ASM Handbook"
hasDefinition#1	dcterms:license	"Reproduced.."

For instance, for the term *Autoclave*, we have a term definition for it. This definition has other metadata associated with it, such as the source of the definition, which is provenance information. We have the rights or the license information associated with this definition. As discussed earlier, the regular template doesn't have support for this kind of representation. With the singleton property template, we can represent the metadata requirement in this example. Particularly, we define the singleton property template for the term *Autoclave* as follows:

```
{{Definition Text
| Definition = A closed vessel..
| Source = ASM handbook Volume 21: Composites.
| Rights = Reproduced with permission of ASM
International. All rights reserved.
www.asminternational.org}}
```

For processing this singleton property template, the MagicWord parameter is defined as:
#singletontemplate:singleton_prop=Definition. This singleton property "Definition" will bear the metadata of the RDF triples, such as "Source" and "Rights." The singleton property template above will be mapped to this set of triples as seen in Table 2.

The advantage of our approach is that the singleton property template extension was seamlessly incorporated into the existing extensions. We will demonstrate the use of the singleton property extension for curating materials science vocabularies in Section 5.

4.3 Identifying Templates and Singleton Property Templates from RDF datasets

As we extend Semantic MediaWiki to develop KnowledgeWiki for curating the vocabularies, we identify its potential as a curation platform for any given RDF dataset. To exploit the full potential of this crowd-sourcing platform for the curation of RDF datasets, importing the RDF dataset into the Semantic MediaWiki platform is required. Hence, we developed an extension to import an existing RDF dataset into Semantic MediaWiki.

We developed an algorithm to create pages in Semantic MediaWiki for each entity given the RDF dataset as an input. It also has the capability to handle the provenance information included in the dataset by reusing the singleton property template.

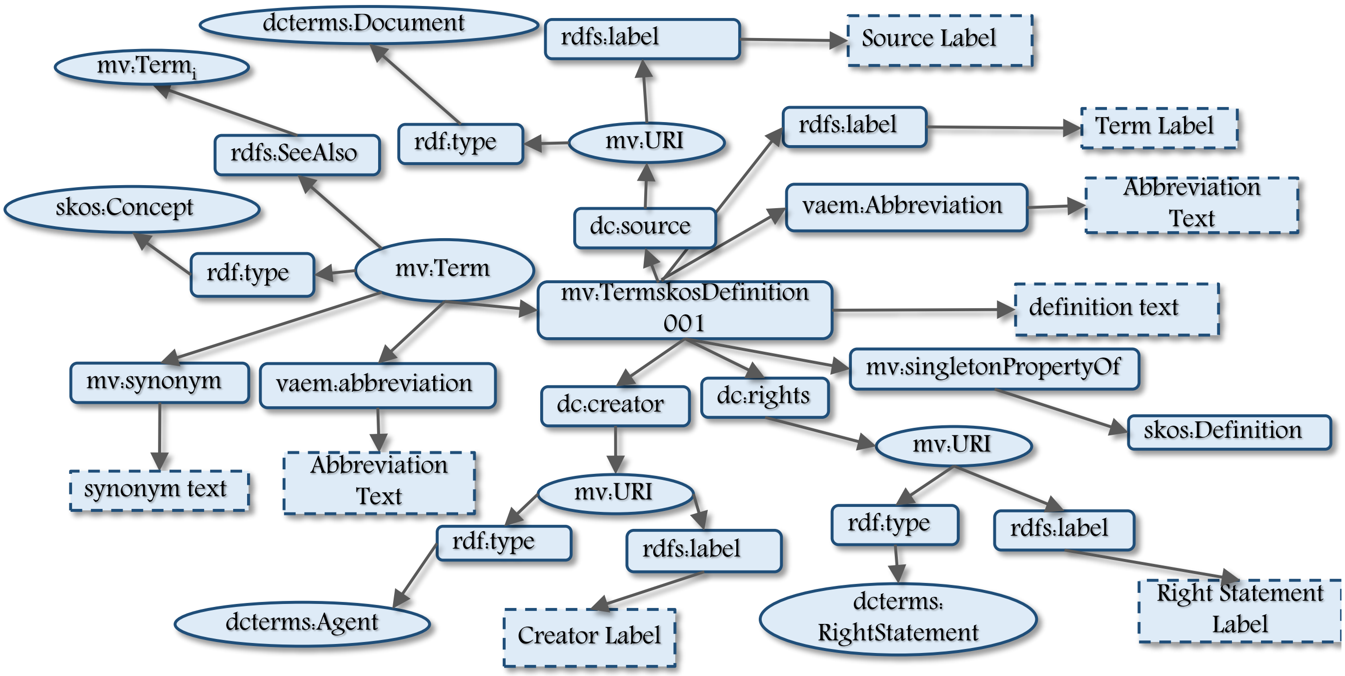


Figure 4: KnowledgeWiki data model representation for Materials Science term along with its metadata information

For the purpose of curation, there exists no mechanism to bring existing RDF datasets into the wiki for curation. This is essential for wider usage and acceptability in the semantic web community. To this end, we have implemented an enhancement in our KnowledgeWiki where users can upload existing RDF datasets into the wiki and open them up to the community to add or edit existing information[7]. The user is only required to provide the named graph and the SPARQL endpoint address to KnowledgeWiki. Our algorithm first automatically identifies the RDF subgraph structures corresponding to the regular and singleton property templates associated with each entity. Finally, we create a wiki page for each entity along with the data obtained from the associated templates. In Section 4.1, we states that in order to use the forms the properties and the templates have to be created in advance. However, this algorithm identifies and creates the set of semantic properties from a given RDF dataset on the fly. Properties and templates that are being used in any given form should be created in advance and must exist prior to using them within the semantic forms.

Automatic creation of the properties and templates is a very useful feature since it is not feasible to create all the necessary properties and templates in advance. For a dataset like Yago2S, which contains over 2.8 million entities with 33 distinct regular properties and 83 distinct generic properties, it becomes tedious to create all these different properties in advance. With this approach, we can automatically identify the set of properties in a given RDF dataset and create properties on the fly.

Algorithm 1 Property-Template Approach algorithm

```
procedure PT–Approach
    identify a list of regular properties
    identify a list of generic properties
    create one wiki page per property
    for each property, check the count of datatypes it has (using group by query) do
        if it has only datatype, map that datatype to the SMW datatype (create the [[has type: type]])
        else create an empty property page
        if the object is URI then the datatype is Page
    end for
    create a list of regular templates, the name of the template is taken from the name of the property
    generate the regular template tag.
    create a list of singleton property template, the name of the template is taken from the generic property
    generate the meta-template tag/code for each template
    identify the list of entities
    for each entity do
        identify the list of regular and singleton property template associated with this entity
        create a wiki page for the content obtained from the templates
    end for
end procedure
```

[7] http://matvocab.org/wiki-dev/index.php/Special:Importdataset

Table 3: Sample CSV file segment

Title	DefinitionText	Source	Rights
A-basis	The A mechanical property value is the value...	ASM Handbook	Reproduced with permission...
A-stage	An early stage in the preparation of certain...	ASM Handbook	Reproduced with permission...
ADK	Notation used for the k-sample Anderson-Darling..	Composite Materials Handbook	MIL-HDBK-17F-1F, 17 June
Aliquot	A small representative portion...	Composite Materials Handbook	MIL-HDBK-17F-1F, 17

Algorithm. The stepwise implementation of the algorithm is defined here for importing an RDF dataset into the wiki by identifying the regular and singleton property templates in the dataset.

Here we define three kinds of properties: regular properties, generic properties, and singleton properties. A typical property is termed as *regular property*. A property that has a singleton property derived from it, is termed a *generic property*. *Singleton properties* can be viewed as instances of generic properties whose extensions contain a set of entity pairs. If **SP rdf:singletonPropertyOf P**, then **SP** is the singleton property of the generic property **P** and both these properties are regular properties.

Next, we identify all the distinct regular and generic properties and create a page for each property. During property creation task, our algorithm checks the datatype of each of the property for the appropriate datatype association. For example, for the property *mv:sourceURL*, which is used to specify the resource link of the source of the definition, we create a property of the type URL. We create a property page with the title *mv:sourceURL* and the content of this page contains `[[Has type::URL]]`.

Then, we create one template per property. The template title is the name of the property, and the wiki page holding the template information also contains the MagicWord for generating the necessary triples.

We also create a list of singleton property templates. Since a singleton property is an instance of a generic property, we name the singleton templates with the generic property name. We add the MagicWord and its associated parameters "#singletontemplate:singleton_prop=" within each template page for processing singleton property templates.

Once all the necessary properties and templates are created, for each entity in the dataset, a wiki page is created to represent the entity by adding all the required templates with its values in the content of the wiki page.

We implemented the algorithm in our KnowledgeWiki for importing any given RDF graph, provided via SPARQL endpoints. Our SPARQL endpoint is public and available for querying. This feature is available at our wiki[8]. This work is ongoing and we are planning to evaluate this algorithm with different datasets.

5. USE CASE FOR MATERIALS SCIENCE

In this work, we were particularly interested in understanding how our KnowledgeWiki could be utilized in materials science community. We would like to explore this question quantitatively and qualitatively with respect to the materials science domain. The architecture of our KnowledgeWiki was discussed in Section 4.1. Here we describe a specific use case of the system for materials science.

Schema-level description of vocabularies. We were provided with three vocabularies by the community, including ASM Handbook Volume 21, MIL-HDBK-17, and MIL-HDBK-5. The data provided to us was legacy data present in excel spreadsheets. We worked with domain scientists and identified the following requirements: (1) consolidate these disparate sets of vocabularies to have one common vocabulary describing the materials science domain, (2) develop a method to represent each term and its provenance metadata appropriately, and (3) make this wiki-based tool open for community authoring.

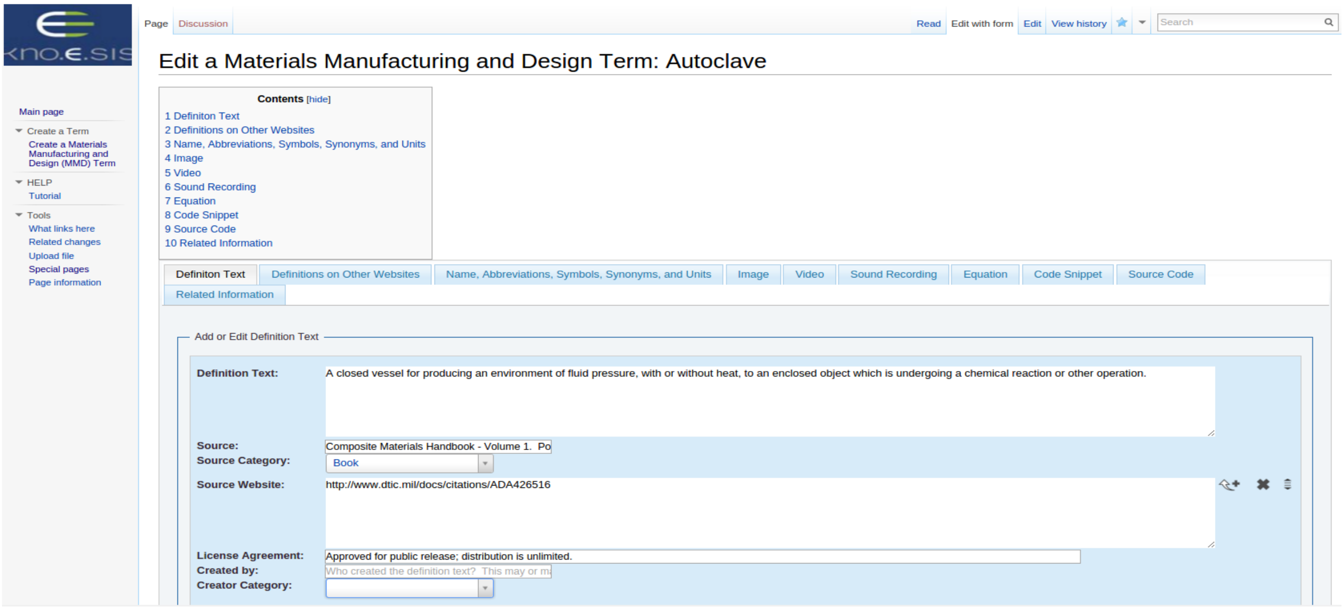
Singleton Property Templates with semantic interface for crowd-sourcing. To address the requirements described above, we developed a new data model for representation. Our data model captures elements such as definition text, image, sound and other elements to represent each term. We defined eleven templates to model and represent the materials science vocabulary, including six singleton property templates and five regular templates. The singleton property templates are Definition Text, Image, Video, Sound Recording, Equation, and Code Snippet. The regular templates are Name Abbreviation Synonym, Symbol, Other Website Definition, Unit, and References.

We use existing vocabularies such as *dcterms* (Dublin Core), *skos* (SKOS), and *qudt* (Quantities, Units, Dimensions and Data Types) to define the classes and properties. We also create a new prefix *mv* (MatVocab) for the new integrated materials science vocabulary. Figure 4 shows the schema-level data model that we developed using the singleton property template extension. The property *mv: TermskosDefinition001* is a singleton property instance of the imported property *skos:definition* and is used to specify the definition of a term. Using the singleton property data model, we can specify all the metadata associated with the definition such as provenance information *dcterms:source* and rights information *dcterms:rights*. When creating a singleton property template, the user gets to select the property to instantiate the singleton property. For example, in Figure 4 the "skos:definition" is selected to generate the singleton property.

The singleton property template uses the singleton property instance to refer to the entire triple succinctly, and enable metadata to be associated with triples via the property instance. Given a vocabulary term, our data model allows users to add multiple definitions (with associated details) for any element. For example, the term *Autoclave* can have multiple definitions derived from different vocabularies such as the ASM Handbook Volume 21, MIL-HDBK-5, and MIL-HDBK-17. Singleton property templates can handle such representations easily.

Finally as mentioned earlier, using these templates along with the forms a wiki page is created for each entity with the term name as the page title within the main namespace. Specifically, Table 4 describes a simple example where, **mat-**

[8] http://matvocab.org/wiki-dev/index.php/Special:Importdataset



Edit a Materials Manufacturing and Design Term: Autoclave

Contents [hide]

1 Definiton Text
2 Definitions on Other Websites
3 Name, Abbreviations, Symbols, Synonyms, and Units
4 Image
5 Video
6 Sound Recording
7 Equation
8 Code Snippet
9 Source Code
10 Related Information

Add or Edit Definition Text

Definition Text: A closed vessel for producing an environment of fluid pressure, with or without heat, to an enclosed object which is undergoing a chemical reaction or other operation.

Source: Composite Materials Handbook - Volume 1. Po
Source Category: Book
Source Website: http://www.dtic.mil/docs/citations/ADA426516

License Agreement: Approved for public release; distribution is unlimited.
Created by: Who created the definition text? This may or m
Creator Category:

Figure 5: KnowledgeWiki's semantic form

sup could be the namespace URI coined by a materials supplier coupled with a local part that specifies the materials. Likewise, **testhouse** could be the namespace URI devised by the company testing the materials (perhaps only used on their intranet). The existing prefix URI **rdf** and **dctype** include a number of defined terms (local parts) that are widely used.

Loading vocabularies into KnowledgeWiki. The three vocabularies (ASM Handbook Volume 21, MIL-HDBK-5, and MIL-HDBK-17) were imported into the wiki using the "CSV Import" feature[9]. The implementation described so far, allows users to add terms via semantic forms. However, in the case when data providers have a large number of terms, it is a tedious task for them to go and add each term manually via semantic forms. Therefore, we developed a bulk upload functionality to add a large number of terms to KnowledgeWiki using a predefined structured form. We extended the Semantic MediaWiki Import CSV feature.

We restrict the format of the input CSV file in a way which adheres to our data model. More specifically, we only allow the properties supported by our semantic forms. In the CSV file, the header row specifies the properties and other rows specify the values for each term (Title). A sample CSV file segment is shown in Table 3.

Here the user just needs to upload a CSV file from their file system. Once the file is uploaded, KnowledgeWiki represents each entity on a wiki page and generates the semantic triples. For example as seen in Table 3.

In order to help the materials science community to create a consolidated vocabulary (MatVocab) using our system, we had to address some complex problems when dealing with knowledge integration. For example, more than one definition can be present on a term's page. The community can use KnowledgeWiki as a means to express member opinions and ultimately select specific elements as terms to define. In other cases, what started as a definition for one term could evolve into multiple terms with their own respective definition elements. For example, if "Modulus" were added as a term then, through community discussion that specific term could be spun-off into other terms like TensileModulus, ShearModulus, CompressiveModulus, or BulkModulus.

One unique feature of KnowledgeWiki is the ability to allow for multiple textual definitions along with their respective source and license information. Therefore, whenever an element of a definition is used, the license is presented along with the element.

The three vocabularies were successfully imported into the wiki using our extensions. There is a total of 2,800 entity pages created from the three vocabularies. Each of these terms may have values for 11 templates defined on the wiki. Currently Kno.e.sis is hosting an instance of KnowledgeWiki, called MatVocab, that is being used to create and curate a vocabulary for material scientists.

Open Sourcing. Open sourcing a piece of software is to make its source code available for modification or enhancement by anyone. By making the software open source, it becomes transparent, easy to use, and widely accepted in the community. It allows for communities to freely access and modify the software and customize it per their requirements. The implementation of our KnowledgeWiki is open source and free to use. The software is available on GitHub[10].

6. DISCUSSION AND FUTURE WORK

One of the important features of our wiki is how easily it can be adapted to a new domain, as discussed below.

1) *Creating the data model for the vocabularies from other domains.* In order to create a vocabulary for a new domain, we only need to create the set of templates to reflect the new relationships in the domain. For this, the user has to identify the information that the user wants to capture, and based on requirement, the user can use a *singleton property* template

[9] http://matvocab.org/wiki-dev/index.php/Special:ImportCSV

[10] https://github.com/MaterialWays/semanticwiki/tree/deployment/deployment/README.md

Table 4: Triple representation

Subject	Predicate	Object
matsup:Material_7075T6	mmd:hasTensileUltimateStrength	"510 MPa"
testhouse:TestSpecimen_3	eg:isComprisedOf	matsup:Material_7075-T6
testhouse:TestSpecimen_3	rdf:type	dctype:PhysicalObject

or a *regular template*. For example, if the user wants to capture provenance information for an entity, the user should use the singleton property template. Next, the user needs to create the semantic forms using these templates. These forms are used for creating and curating information on the wiki. Further information on how to set up KnowledgeWiki for a new domain can be found on our wiki page[11]

This work can be applied to other domains such as chemistry (e.g., PubChem) and bioinformatics (e.g., BKR). In fact, the singleton property approach has been discussed by the bio hackathon community and has also been evaluated and compared with other approaches in PubChem. In another work by Tudorache et al. they describe their tool iCAT to curate the International Classification of Diseases (ICD-11) [17]. They report the use of such a tool in the classification of epidemiology and healthcare data management for clinical purposes. Therefore, with the above mentioned features we feel our system is easy to adapt to a new domain. Another aspect is the

2) *Use of our system for creating RDF datasets in general,* this is not only with provenance metadata, but for any kind of meta data about the triples. The semantic extension that we developed for representing meta information is generic and allows for any kind of data annotation such as provenance, access control or spatio-temporal information. Finally,

3)*Data re-usability,* the data and its meta information that is collected or created using our system is ultimately represented as triples in the RDF format. This output created by our system, where the data is represented using the singleton property approach can be queried or reused by other applications and tools. The singleton property triple pattern allows consumers of our data to discover expected access patterns for this data. Currently our data can be queried using the standard SPARQL queries, we have listed a set of simple SPARQL queries on our wiki page[12].

7. CONCLUSION

In this work we have reported about the development of our KnowledgeWiki and how we adopted it for curating vocabularies in the materials science domain. We have showed that by enhancing and extending the existing Semantic MediaWiki we can facilitate communities to create and curate vocabularies. KnowledgeWiki has been designed as an extension to the well known open source platform Semantic MediaWiki (SMW) that adds a mechanism to capture provenance information efficiently by leveraging the well known singleton property approach.

KnowledgeWiki is open source and licensed under GNU General Public License. The content published in KnowledgeWiki is licensed under Creative Commons Attribution-ShareAlike License unless overridden by the data providers as stated in the new licensing statement. We have been able to bring existing materials science data into the wiki successfully for curation and preservation the metadata of the RDF triples. The extension of singleton property templates in KnowledgeWiki is able to accomplish such a goal.

8. ACKNOWLEDGEMENTS

This research was funded by the U.S Air Force Research Laboratory(AFRL), through the contract FA8750-13-1-0244. We would like to thank Clare Paul for his vision of how domain experts would use the KnowledgeWiki as well as his technical contributions in its design and development. We also thank Kalpa Gunaratna, Siva Kumar Cheekula, Swapnil Soni, and Mary Panahiazar who contributed to related aspects of this work.

9. REFERENCES

[1] S. Auer, S. Dietzold, and T. Riechert. OntoWiki–a tool for social, semantic collaboration. In *The Semantic Web-ISWC 2006*, pages 736–749. Springer, 2006.

[2] M. N. K. Boulos. Semantic Wikis: A comprehensible introduction with examples from the health sciences. *Journal of Emerging Technologies in Web Intelligence*, 1(1):94–96, 2009.

[3] F. Bry, S. Schaffert, D. Vrandečić, and K. Weiand. *Semantic wikis: Approaches, applications, and perspectives.* Springer, 2012.

[4] M. Buffa, F. Gandon, G. Ereteo, P. Sander, and C. Faron. SweetWiki: A semantic wiki. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):84–97, 2008.

[5] K. Cheung, J. Drennan, and J. Hunter. Towards an Ontology for Data-driven Discovery of New Materials. In *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*, pages 9–14, 2008.

[6] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić. Introducing Wikidata to the Linked Data Web. In *The Semantic Web–ISWC 2014*, pages 50–65. Springer, 2014.

[7] G. Fu, E. Bolton, N. Q. Rosinach, L. I. Furlong, V. Nguyen, A. Sheth, O. Bodenreider, and M. Dumontier. Exposing Provenance Metadata Using Different RDF Models. *arXiv preprint arXiv:1509.02822*, 2015.

[8] D. Hernández, A. Hogan, and M. Krötzsch. Reifying RDF: What Works Well With Wikidata?

[9] A. International, A. I. A. P. D. Committee, and A. I. H. Committee. *ASM handbook*, volume 13. ASM International, 2005.

[10] M. Krötzsch, D. Vrandečić, and M. Völkel. Semantic mediawiki. In *The Semantic Web-ISWC 2006*, pages 935–942. Springer, 2006.

[11] T. Kuhn. Acewiki: A natural and expressive semantic wiki. *arXiv preprint arXiv:0807.4618*, 2008.

[11] http://wiki.knoesis.org/index.php/KnowledgeWiki

[12] http://wiki.knoesis.org/index.php/KnowledgeWiki

[12] D. Kukich. MIL-HDBK-17. *AMPTIAC Newsletter*, 1(3):2.

[13] V. Nguyen, O. Bodenreider, and A. Sheth. Don't like RDF reification?: making statements about statements using singleton property. In *Proceedings of the 23rd international conference on World wide web*, pages 759–770. International World Wide Web Conferences Steering Committee, 2014.

[14] R. C. Rice. *Metallic Materials Properties Development and Standardization (MMPDS): Chapters 1-4*, volume 1. National Technical Information Service, 2003.

[15] C. Sarasua, E. Simperl, N. Noy, B. Abraham, and J. M. Leimeister. Crowdsourcing and the Semantic Web: A Research Manifesto. *Human Computation*, 2015.

[16] S. Schaffert. IkeWiki: A semantic wiki for collaborative knowledge management. In *Enabling Technologies: Infrastructure for Collaborative Enterprises, 2006. WETICE'06. 15th IEEE International Workshops on*, pages 388–396. IEEE, 2006.

[17] T. Tudorache, J. Vendetti, and N. F. Noy. Web-Protege: A Lightweight OWL Ontology Editor for the Web. In *OWLED*, volume 432, 2008.

[18] D. Vrandečić and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

[19] A. White. The materials genome initiative: One year on. *MRS Bulletin*, 37(08):715–716, 2012.