AFRL-RI-RS-TR-2017-002

SEMMAT: FEDERATED SEMANTIC SERVICES PLATFORM FOR OPEN MATERIALS SCIENCE AND ENGINEERING

WRIGHT STATE UNIVERSITY

JANUARY 2017

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2017-002 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
ALBERT YU, Captain, USAF
Work Unit Manager

/ S /
JULIE BRICHACEK
Chief, Information Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JAN 2017	FINAL TECHNICAL REPORT	JUL 2013 – JUN 2016

4. TITLE AND SUBTITLE

SemMat: FEDERATED SEMANTIC SERVICES PLATFORM FOR OPEN MATERIALS SCIENCE AND ENGINEERING

5a. CONTRACT NUMBER
FA8750-13-1-0244

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
62102F

6. AUTHOR(S)

Amit Sheth, Krishaprasad Thirunarayan, Nishita Jaykumar, PavanKalayn Yallamelli, Sarasi Lalithsena, Vinh Nguyen

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
R157

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Wright State University
3640 Colonel Glenn Hwy
Dayton OH 45435-0001

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RISA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2017-002

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

An open source framework was developed to enable crowd-sourcing and curation of controlled vocabularies. The framework was then applied to the materials and manufacturing domain to create several hundred terms and definitions extracted from various structured sources. The domain model and the curation platform supports preserving important metadata information including provenance. Additionally, a novel Singleton property approach was implemented to enable representation of relevant information efficiently and in a semantically clean manner. A visualization tool, iExplore, was developed to provide the capability to visually search and explore links between materials and manufacturing domain concepts. Tools and techniques were developed to enable identification of materials entities in unstructured data sources and documents (such as PDF documents).

15. SUBJECT TERMS

Semantic Technology, Materials Development

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			ALBERT YU, Captain, USAF
U	U	U	SAR	26	19b. TELEPHONE NUMBER (*Include area code*) (315) 330-7116

TABLE OF CONTENTS

LIST OF FIGURES

1.0 SUMMARY

One of the goals of the White House's Materials Genome Initiative (MGI) is to develop solutions that provide broad access to scientific data. This allows materials scientists to exchange and integrate each other's data for better outcomes. Kno.e.sis Center, in collaboration with Materials and Manufacturing Directorate and the Information Directorate, Air Force Research Laboratory (AFRL), identified the following two important tasks to remedy the data heterogeneity challenge to promote data integration: (1) creating the semantic infrastructure to curate vocabularies and domain models to standardize and represent materials data, and (2) leverage these vocabularies to process unstructured documents and use the annotated data to improve document search.

Standardized vocabularies are widely used as the shared language to solve the data heterogeneity issues. Vocabulary development and evolution is an iterative process that requires community agreement and ongoing curation for wider adoption. For this purpose, we developed a crowdsourcing platform (MatVocab) by adopting and progressively adapting the existing Semantic MediaWiki (SMW) platform. This approach enables materials scientists across the globe to participate in the vocabulary curation activity. It is critical that provenance metadata be faithfully preserved in order to enable reliable data integration from disparate sources. In fact, this is particularly important for a crowd sourced data set, where the quality of different authors and sources may be non-uniform. Thus, the design of MatVocab pays particular attention to supporting capabilities that keep track of the provenance information. We initially populated our vocabulary from existing structured data sources such as the glossaries of ASM Handbook Composites Volume 21 (ASM-21) [1], Composite Materials Handbook (MIL HDBK-17) [2] and the Metallic Materials and Elements for Aerospace Structures handbook (MIL-HDBK-5) [13].

Further, we show how to search occurrences of the curated vocabulary instances in unstructured documents. Specifically, for this purpose, we have developed an annotation tool that spots the entities in a PDF document using terms in a given vocabulary. Currently our tool provides the concept driven search over documents. These annotations can later be exploited for more advanced semantic querying of the documents.

2.0 INTRODUCTION

The aim of this project was to provide easy access to highly distributed and heterogeneous materials and biomaterials data for researchers to share and exchange for various purposes including new materials discovery and deployment. A key component of this project was to introduce better data management practices to materials and process community. In this project, we leveraged the strengths of semantic web technologies, which have been used successfully in other disciplines such as Bioinformatics, Life Sciences and Health Care at the Kno.e.sis Center.

We have gathered information from domain experts, handbooks and web resources to establish a common vocabulary for the materials manufacturing and design domain. Data representation was further enriched by capturing provenance information. Our open source framework is designed to engage the community to curate, use, and explore the vocabulary which will greatly improve its coverage, reliability, and application. Specifically, we provide tools to query and browse the data that will allow easy access to the data to novice users. Furthermore, we develop techniques to spot the entities in unstructured documents and tools to search the documents with the vocabulary.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

In this research and development effort, we mainly considered two tasks to apply informatics to materials domain. The first relates to creating a semantic infrastructure for the materials data by building vocabularies and domain models to represent materials data. This provides a data exchange scheme for materials science, which also includes provenance information to promote flexible data access and integration. The second relates to semantic search on structured and unstructured materials and processing data annotated using standardized vocabularies and domain models that we developed in the first task.

- Developing and curating vocabularies for broader materials domain
    - Develop vocabularies and domain models to represent materials data
    - Develop a crowdsourcing platform to curate the vocabularies
    - Incorporate provenance into the domain models
    - Convert legacy data into triples
- Indexing and semantic search of materials documents and data for documents
    - Identify data sources
    - Spot entities and relationships in unstructured documents
    - Efficient indexing of data
    - Semantic search over data

3.1 Development of Vocabularies or Domain Models to Represent Materials Data

A vocabulary defines domain terms and characterizes their relationships. Vocabularies help to establish a common agreement among the community about the interpretation of terms, organize the available knowledge, and integrate the data. Medical professionals heavily use vocabularies such as SNOWMED CT, ICD and MeSH to represent knowledge about symptoms, diseases and treatments. For the materials domain, we have developed the MatVocab vocabulary to establish a common agreement about the definition of the terms. Next we describe the semantic model we developed for the vocabulary.

MatVocab can be accessed via http://wiki.knoesis.org/index.php/MaterialWays.

3.1.1 Semantic Model for the Vocabulary

We identified a list of term definition elements with the help of domain experts. These elements capture different aspects of the term and provide a comprehensive description. The elements currently used to fully define a term are:

- Definition Text
- Definition on Other Websites
- Name
- Abbreviation
- Synonym
- Unit
- Image
- Video
- Sound Recording
- Equation
- Code Snippet
- Link to Source Code
- Related Information

Creating a semantic model for the vocabulary terms primarily requires identifying the properties (semantic property name between the Term and the Element) and classes (semantic class for each Element). By adhering to the reusability principle of the Semantic Web, we assessed the properties and classes from existing vocabularies such as SKOS [3], Dublin Core [4], PROV [5], FOAF [6], MathML [7] and QUDT [8] for reuse suitability. A complete list of vocabularies can be found in Appendix A. We identified and analyzed 106 candidate classes and properties with the help of our domain expert and agreed on the above classes and properties to be used in our vocabulary model. We used RDF representation for our semantic model. Each term may have multiple occurrences of each definition element. For example, a term can have any number of textual definitions and each textual definition can be from a different source. This approach was

chosen, in-part, to enable the community to collectively view candidate elements of the definition and winnow them down to those that would ultimately be used to define the term.

The vocabulary was initially populated with the terms extracted from ASM-21and MIL-HDBK-17. Currently the vocabulary consists of several hundred terms, and can be found on the MatVocab wiki page.

3.1.2 Incorporation of Provenance into the Domain Models

Provenance information helps to capture the relevant metadata associated with each term. For example, ASM-21 and MIL-HDBK-17 each provide definitions for the term "Creep."

In cases such as these, it's important to include source and license details with each *Definition Text.* This was made possible through the use of a semantic model which incorporated the Singleton Property [9] approach.

3.1.3 Singleton Property Approach to Capture Provenance Information.

The singleton property approach is a mechanism to add metadata to RDF triples. It uses a property instance to refer to the entire triple succinctly and enables metadata to be associated with triples

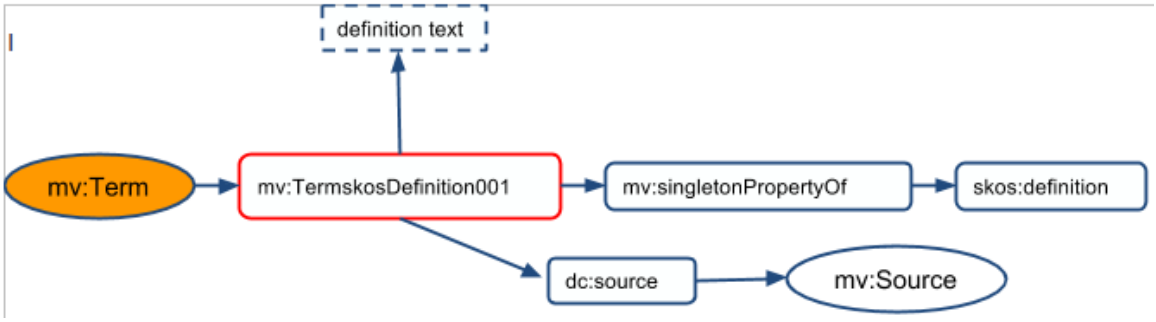

Figure 1. Schema View of the Singleton Property Usage for Definition Text to Include the Source Information.

Figure 1 depicts the schematic view of the usage of singleton property to represent the source information with the definition text element. The singleton property instance is being used to attach the meta-triple for the source information. Out of all the elements, seven elements contain provenance information. Table 2 describes the provenance information associated with each element. More details on the modeling of selected elements can be found in Appendix B.

Table 1. Element Information and the Meta-information Associated with Each Element

Element	Property Name	Meta Information	
		Meta Element	**Meta Property Name**
Definition Text	**skos:definition**	source	dcterms:source
		source category	mv:sourceType
		source website	mv:sourceURL
		rights	dcterms:rights
		creator	dcterms:creator
		creator category	dcterms:creator
Image	**mv:image**	source	dcterms:source
		source category	mv:sourceType
		source website	mv:sourceURL
		rights	dcterms:rights
		creator	dcterms:creator
Moving Image	**mv:movingImage**	source	dcterms:source
		source category	mv:sourceType
		source website	mv:sourceURL
		rights	dcterms:rights
		creator	dcterms:creator
Sound	**mo:recording_of**	source	dcterms:source
		source category	mv:sourceType
		source website	mv:sourceURL
		rights	dcterms:rights
		creator	dcterms:creator
Equation	**xhv:math**	source	dcterms:source
		source category	mv:sourceType

		source website	mv:sourceURL
		creator	dcterms:creator
Code Snippet	**mv:codeSnippet**	programming language	schema:programmingLanguage
		source	dcterms:source
		source category	mv:sourceType
		source website	mv:sourceURL
		license agreement	dcterms:license
		created by	dcterms:creator
		creator category	mv:creatorType
Source Code	**Link to Source Code**	Link to Source Code	mv:sourceURL
		Description	rdfs:comment
		Programming Language	schema:programmingLanguage

3.1.4 A Crowdsourcing Platform to Curate the Vocabularies

While the MatVocab vocabulary was initially populated with a bulk up-load using ASM-21 and MIL-HDBK-5, given that the vocabulary is being created and edited through community agreement, it is important to have proper mechanisms to allow the geographically dispersed materials community to curate the MatVocab vocabulary.

Wikis have been used as a tool to organize and share knowledge in communities and organization in a user friendly manner. Wikipedia, one of largest publicly available knowledge sources, is a great example of what is possible using wikis. The SMW is a free and open source extension to MediaWiki, which is the application on which Wikipedia is based. While traditional wiki supports only textual context, SMW allows semantic annotation of data. It allows users to create statement about a given entity while insulating the users from the details about the underlying semantic modelling and data representation. Figure 2 shows an example of such a statement which states the population of the *Berlin*.

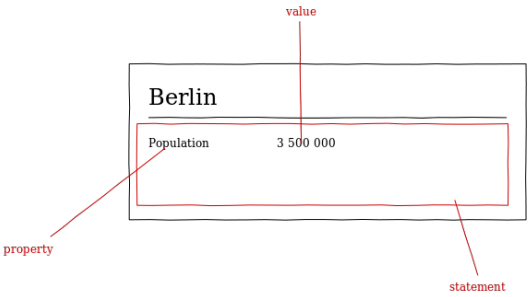

Figure 2. A Statement about the Berlin Population in Semantic MediaWiki

3.1.5 Curation of MatVocab via Semantic MediaWiki

We have adopted SMW platform [10] for developing the collaborative environment for materials scientists to curate the vocabulary. This requires two main extensions to the current SMW platform to facilitate the modelling we described above: (1) support for the singleton property approach to capture the metadata, and (2) support for adding typed information (class information) for the modelling elements.

SMW provides a form to add, edit and query the data via Semantic Forms extension. Figure 3 shows the form that is used to add a term and add or modify applicable elements. Separate tabs have been created for each element, and Figure 3 shows the details of the *Definition Text* element.

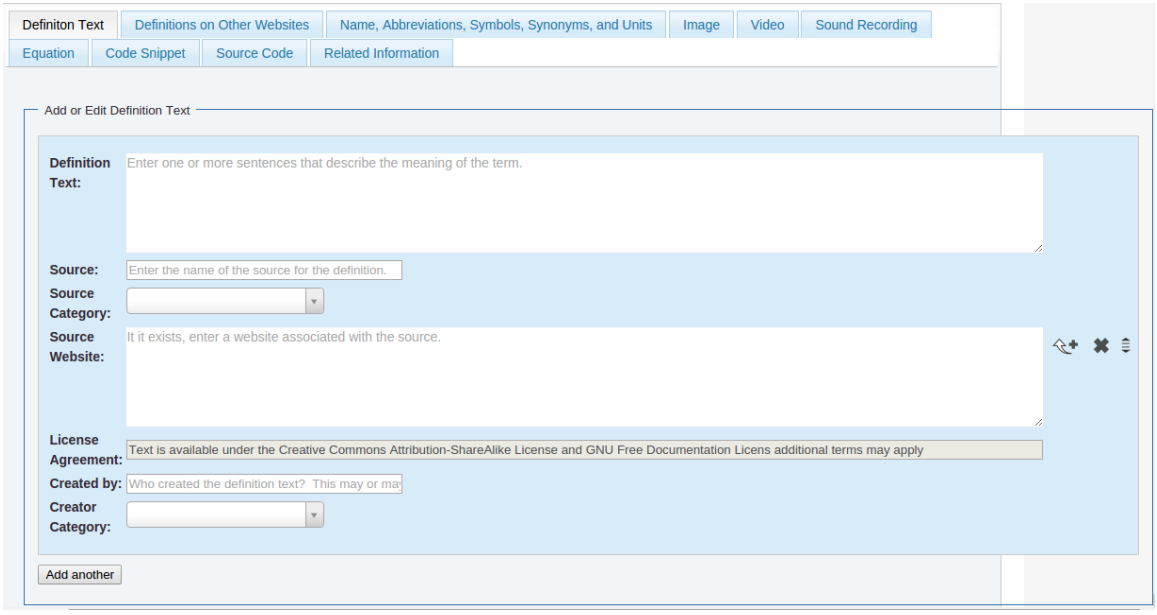

Figure 3. A Screenshot of the MatVocab Form Used by Domain Experts to Add a Term and Associated Elements to the Vocabulary

The Semantic Forms extension enforces the use of Templates in creating semantic data. Templates are a popular way of handling semantic annotations in SMW and for storing data via SMW. Templates define the allowable properties (e.g., skos:definition and mv:image) and the data types (e.g., text, .jpg) of the value. A regular template in SMW does not support adding metadata. So, we developed our own new extension for template, called "Singleton Template", for this purpose. Singleton Template uses the base features of the SMW template, and additionally enables the support for adding metadata as well. Singleton template distinguishes different usage of properties.

A typical property is termed a regular property, used to create a statement. In order to attach meta information about a statement, we create a singleton property instance of the regular property. A property that has a singletonProperty derived from it is termed a generic property. For example, let's assume we want to attach the provenance information such as source and license information associated with a definition text. We proceed as follows:

a. Create a singleton property instance of the *skos:definition* property:

mv:Property/ABasis_Definition_Text_01	*rdf:singletonPropertyOf*	***skos:definition***

ABasis_Definition_Text_01 is the singleton property of the generic property *skos:definition*, and both properties are regular properties.

b. Use the singleton property instance to link the term to its definition text:

mv:ABasis	*mv:Property/ABasis_Definition_Text_01*	*"A statistically-based..."*

c. Define the source of this definition text:

mv:Property/ABasis_Definition_Text_01	*dcterms:source*	*mv:URI_01*
mv:URI_01	*rdfs:label*	*"MIL-HDBK-17F-1F, 17 June 2002"*

In addition to the regular properties, singleton template allows one to create singleton property instances and regular properties associated with singleton property instance. Figure 4 depicts the singleton template for the *Definition Text* element.

The capabilities of the MatVocab wiki are described in the section entitled "Platform and Tools Developed or Extended." This allows material scientists worldwide to contribute to and help curate the MatVocab vocabulary.

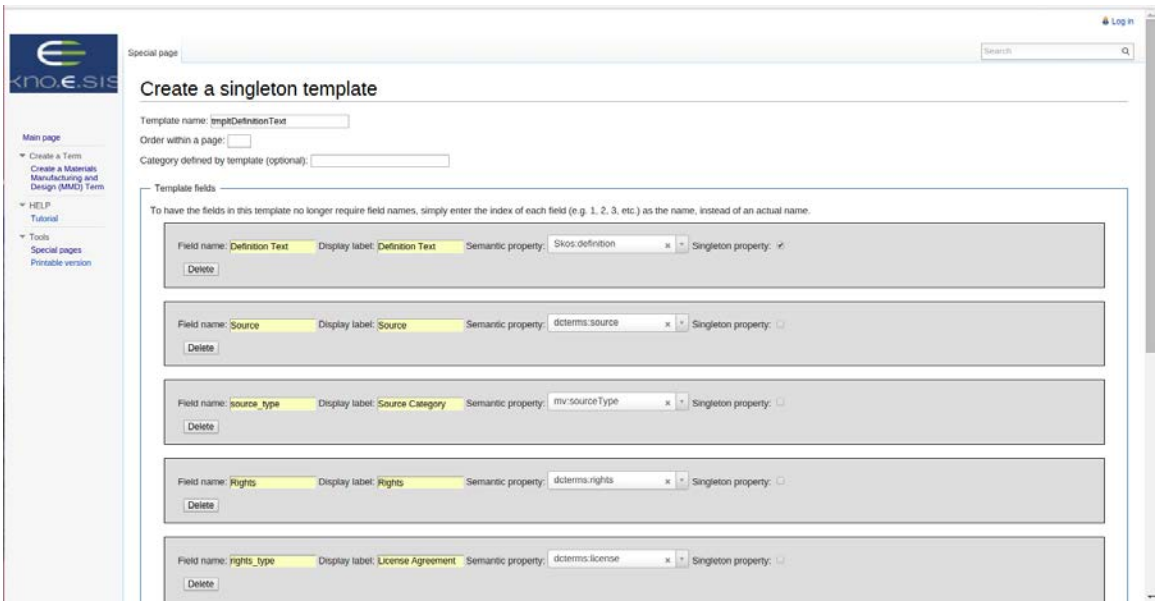

Figure 4. Screenshot of the Singleton Property Template of Definition Text

3.1.6 Converted Legacy Data into Triples

Here, we focused on converting the data from structured data sources into RDF triples using the glossaries of ASM-21 and MIL-HDBK-17 provided in CSV format. Initially, a program was developed to convert the CSV format into the vocabulary model we described above. Later, we integrated this functionality with the MatVocab wiki architecture so that users can upload any CSV file into the MatVocab wiki and automatically convert them into the RDF format for storage in the Virtuoso database store. This also allows anyone (e.g, on behalf of a subcommunity) to bulk upload set of terms to the MatVocab vocabulary rather than add each term individually.

Bulk upload functionality adds terms provided in a predefined, structured form to MatVocab. We extended the SMW Import CSV feature for this purpose. As specified by the sponsor, this functionality requires admin access. We restrict the format of the input CSV file in such a way that it adheres to our Semantic Model. More specifically, we only allow the properties supported by our Semantic Forms as illustrated below. In the CSV file, header row specifies the properties and other rows specify the values for each term.

3.2 Indexing and Semantic Search

3.2.1 Data Sources

Data was primarily sourced from the structural and bio-materials domains.

For structural materials data, we reviewed and used MIL-HDBK-5J [11] and MIL-HDBK-17. Furthermore, we used the ASM-21 glossary for additional vocabularies and definitions. ASM

permitted us to include their glossary into our MatVocab vocabulary. In addition to the structured data we mentioned earlier, we also used a corpus of 140 documents about composite materials provided by our domain expert.

Based on the suggestions given by domain experts in bio-materials, the following sources were accessed:

- PDB for initial ontology construction
- PUBMED articles (2009-2013) related to Gold Binding Peptide (including 1,414,637 papers for Binding, 5,525 for Gold Binding, 1,530 for Gold Binding Peptide, and 37 for Gold Binding Peptide from the year 2013) for our test set
- SciFinder publications (67 publications) for Gold Binding Peptide (as our initial test set for the search engine)

3.2.2 Entities and Relationships in Unstructured Documents

A flexible and robust annotation tool was developed that finds occurrences of materials vocabulary in a document. These annotations can be used later to support semantic querying of the documents. We experimented with PDF documents involving materials requirements provided to us by domain experts. Specifically, these technical reports were downloaded from The Defense Technical Information Center (DTIC) using the search phrase “polymeric matrix composites”. Each technical report is on an average 50 pages long and is mostly scanned photocopies. We extracted the text (excluding images and tables) from the PDF files using Apache PDFBox and used Lucene to index and search the textual description of these documents. Specifically, the PDFTextStripper module of PDFBox extracts the text from these scanned documents.

Users can select the terms from the vocabulary and have the selected terms in a document spotted and highlighted as depicted in Figure 5. Specifically, we used PDFClown to highlight search results directly on the PDF document. Note that, in general, the task of annotating and highlighting phrases directly on the PDF document is non-trivial. For example, the available tools fail to properly isolate text, tables, and images, or handle papers in 2-column format because they incorrectly join lines from adjacent columns.

Figure 5. Annotated Document Using Composites and Metals Vocabulary.

3.2.3 Efficient Indexing

For the annotation tools developed for the broader materials domain, it requires us to create indices since we are dealing with large number of documents. We used Lucene index for this purpose. We maintain a set of 140 documents about composites. Each term in this document collection is indexed with its position of occurrence in the paper. The index is stored on the server. When a user queries the client side, we are able to perform quick search and retrieve relevant results.

For finding entities and relationships in the biomaterials context, we have indexed the whole Medline article abstracts up to June 2013 as one of the valuable resources we use in this project. The index is built on top of the Lucene indexing engine and the index size is 56GB covering 21 million abstracts.

3.2.4 Semantic Search and Visualization

While we can generate high quality data via the proposed crowdsourcing platform, it is important to have a means to search and explore the data. During the course of this project, three approaches for searching both structured and unstructured data were developed.

We adapted an in-house developed tool iExplore [12] to browse and visualize the RDF data generated from MatVocab framework. Users can start browsing using a keyword given to the system and the system will show the most related entity as a node in a graph for the given keyword. For example, Figure 7 depicts the visualization for the term “ABasis”. Users can further browse data by expanding this entity using its relationship to other entities. More details about the tool can be found later in this document.

An instance of a Virtuoso data store was used to store the RDF and includes a SPARQL endpoint. SPARQL queries can be used to explore the data store.

In addition to search RDF data, we provide the capabilities to perform concept driven search of the documents. This allows users to search the documents with the terms in the vocabulary as given in Figure 9. More details on this tool can be found in the deliverable section below.

3.3 Developed or Extended Platform and Tools

We discuss the tools/information available from Matvocab. Key capabilities are described through examples and high-level implementation details.

3.3.1 MatVocab: SMW for Curating Materials Vocabulary

MatVocab is the primary deliverable of this project and consists of vocabulary terms for the materials manufacturing and design domain and is intended to be curated by domain experts.

Capability: Add or Modify Terms of the MatVocab Vocabulary. Users can add terms to the MatVocab vocabulary via user friendly interfaces as given in Figure 6. If the term already exists, it will navigate users to the existing page of the term where it can be viewed or modified. Otherwise they can create a new page for the term. Then, users will be presented with the form to add the relevant information as depicted in Figure 3.

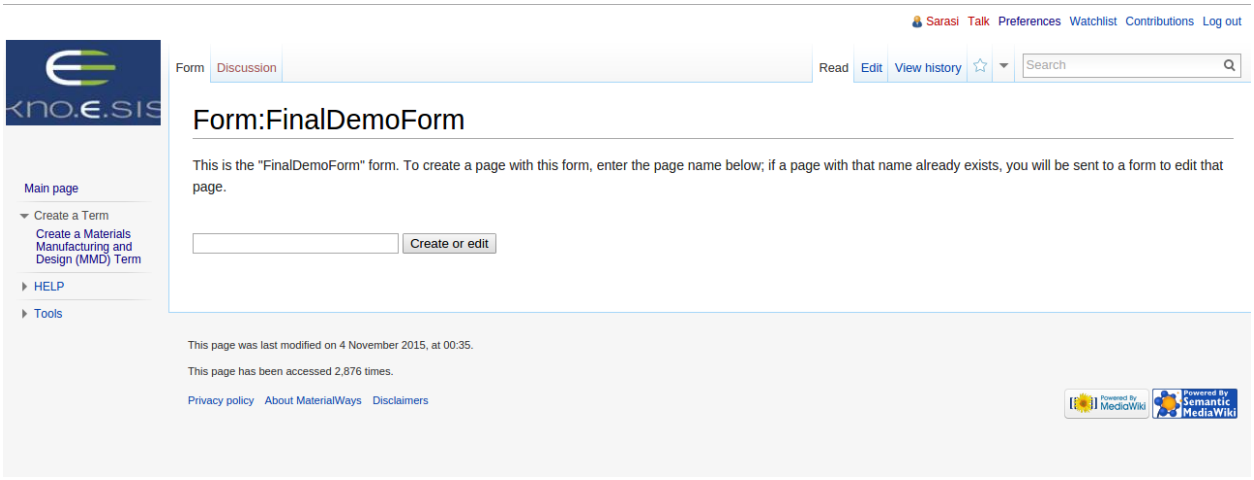

Figure 6. Add/Modify Terms in MatVocab.

Capability: Bulk Upload of Data. Users can add a set of terms together using the bulk upload capability.

Capability: SPARQL Endpoint to Access the Data. Users familiar with the SPARQL query language can query the vocabulary data using the SPARQL endpoint.

Capability: Export MatVocab Data. Users can export the MatVocab data using the export capability.

Capability: Import RDF Data. The current SMW does not allow users to import an existing RDF data set for curation. However, the MatVocab framework allows the upload an existing RDF data set.

Capability: Provide the Framework to Create Any Vocabulary. While MatVocab is hosted at Wright State University to collect and share the terms for materials manufacturing and design community, the framework is generic and available to the broader community to create vocabularies in other domains. We bundled our software and created instructions on how to deploy the system.

3.3.2 iExplore: Visualizing Semantic Web Data.

MatVocab generates RDF triples from various sources (MIL-HDBK-5, MIL-HDBK-17). The RDF triples are stored in a data store and require an understanding of SPARQL to retrieve query results. iExplore, an interactive exploration tool, was developed to visualize the graphs of RDF triples.

Capability: Search for terms and visualize the RDF triples. iExplore allows the user to visualize a set of triples related to a resource in the directed graph form. Starting with a term, a directed subgraph of RDF triples related to the term can be explored in both forward and backward direction. Figure 7 visualizes a search on ABasis.

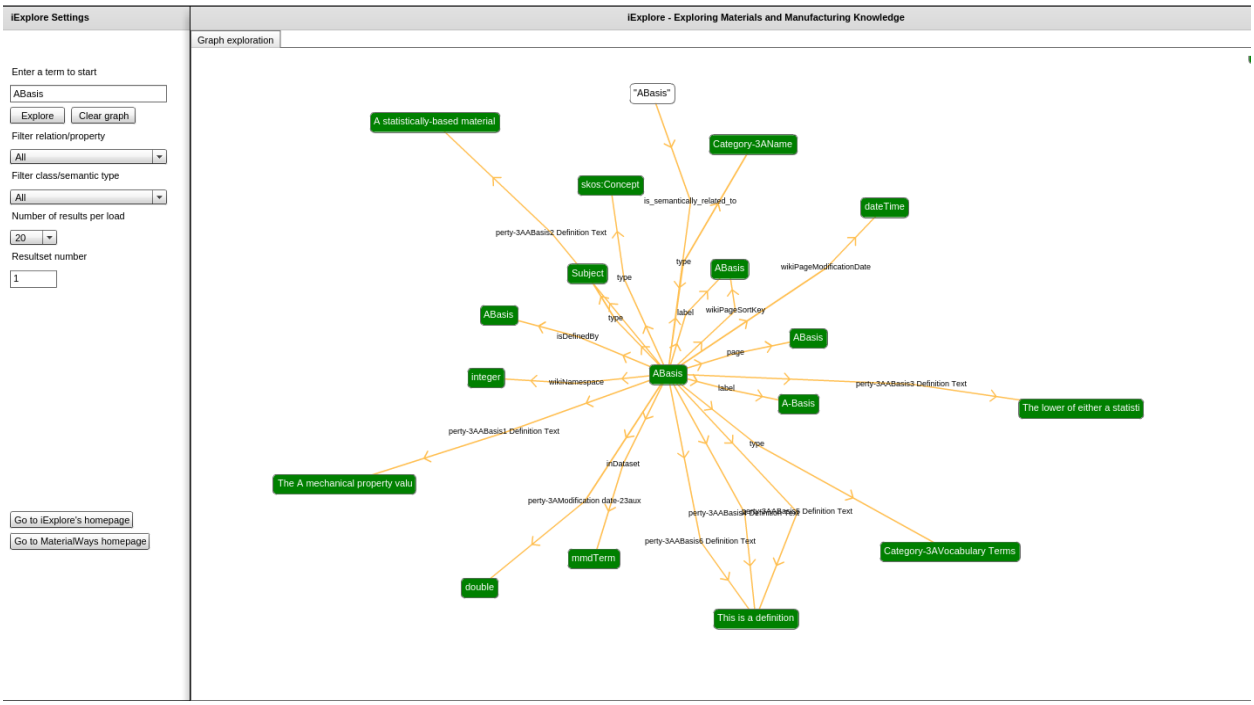

Figure 7. A Sample Graph for "ABasis"

To stay focused on certain terms of interest, one may also collapse a subgraph of incoming or outgoing terms. By combining the two operations (*expand* and *collapse*) in two directions (*forward* and *backward*), a user can construct a summarized graph of interest.

3.3.3 Semantic Annotation Tool

A semantic annotation tool was developed for the Materials Science community for finding relevant entities in materials science documents.

Capability: Search Documents for Terms in a Curated Vocabulary. The materials science domain experts provided us with seed documents which were subsequently loaded in the system. These documents were then indexed using Lucene. Users can search this seed data set using the terms in the vocabulary. There are three ways to provide the search terms.

- Select a terms/phrase from the default vocabulary in the system
- Provide a csv file which contains a list of terms/phrases - Here, users can add a list of terms which do not occur in the controlled vocabulary to be searched.
- Provide a single keyword in the search bar

The tool is able to perform both conjunctive and disjunctive search in the case of multiple terms

Figure 8 depicts a screenshot of the main page of the annotation tool where users can provide the input.

Capability: Find the Selected Terms in the Relevant Documents. Returned search results (documents) are highlighted with the user's input term(s). Users can download the original file with the annotation of the selected terms.

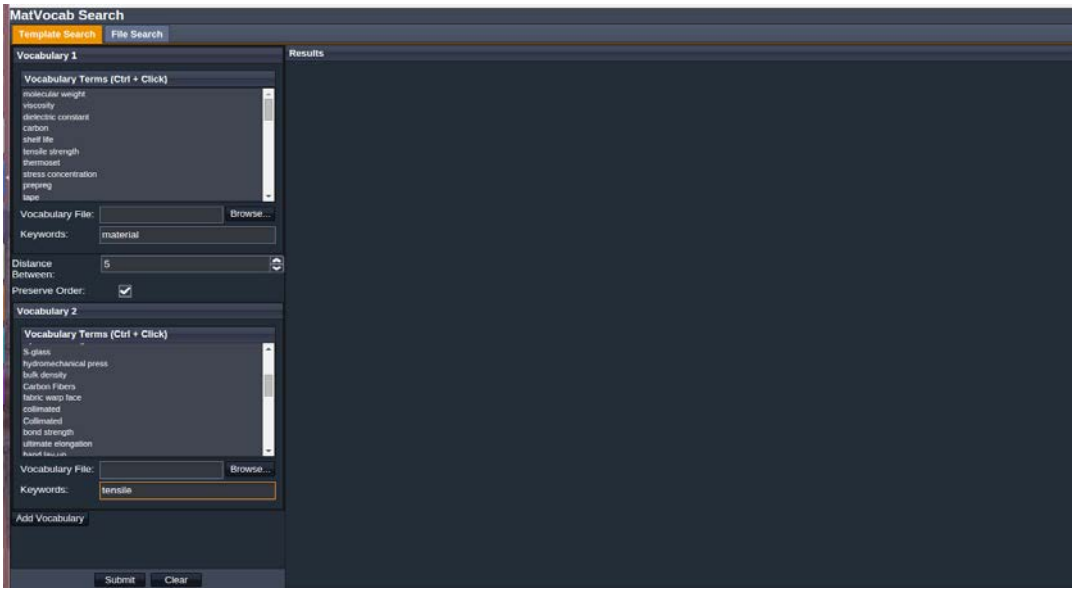

Figure 8. A screenshot of the Main Page of the Annotation Tool.

Capability: Upload the Documents to the File System. This semantic annotation tool was developed in Extjs, a JavaScript application framework for building interactive cross platform web applications. Annotation of PDF documents are performed using the Lucene Highlighting API along with the PDFClown API.

4.0 RESULTS AND DISCUSSION

As discussed in the deliverable section, we have developed the following tools among others during the project.

MatVocab – An extended Semantic MediaWiki for curating materials vocabulary: Kno.e.sis is hosting the MatVocab wiki for curating the materials vocabulary being developed. Domain experts can add terms to the vocabulary, edit information associated with each term, and upload a collection of terms simultaneously using bulk import facility. Currently, the MatVocab vocabulary contains several hundred terms. We have used a novel technique based on singleton property to represent the metadata information very efficiently and in a semantically clean manner. Even though the current vocabulary provides a flat list of terms, in the future, these terms can be further organized and enriched using different relationships such as class-subclass-instance, partonomy or based on any domain specific characteristics.

MatVocab software package: Wiki platform being used to develop the vocabulary is open source and this will allow any interested organization to use our software package to develop their vocabulary, or build upon the current system.

Annotation tools: We have developed and experimented with tools that annotate PDF documents with the vocabulary terms. The tool allows concept driven search over the documents. In future, this tool can be used for more flexible and advanced semantic querying exploiting richer vocabulary.

iExplore Tool: Our visualization tool iExplore provides the capability to search and browse the curated vocabulary terms.

5.0 CONCLUSION

In this project, we have developed an open source MatVocab framework which is a crowd sourced platform to curate the vocabularies. We adopted the MatVocab crowd sourced platform for creating and curating a common vocabulary for the materials manufacturing and design domain. MatVocab vocabulary consists of several hundred terms extracted from various structured sources. Our domain model and the curation platform supports preserving important metadata information including provenance. We have used the novel Singleton property approach to represent the relevant information efficiently and in a semantically clean manner. Our visualization tool iExplore provides the capability to search and browse the vocabulary

terms. We have also developed tools and techniques to search and spot the vocabulary terms (denoting materials entities) in unstructured data sources and documents (such as PDF documents).

6.0 REFERENCES

[1] Handbook, A.S.M. "Volume 21: Composites." ASM International, January (2005).

[2] Department of Defense Handbook, Composite Materials Handbook Volume 1. Polymer Matrix Composites Guidelines for Characterization of Structural Materials, Volume 1 of 5 2002.

[3] Miles, Alistair, and José R. Pérez-Agüera. "SKOS: Simple knowledge organization for the web." Cataloging & Classification Quarterly 43, no. 3-4 (2007): 69-83.

[4] Dublin Core. Available at: http://dublincore.org/ (Accessed 01/02/2015)

[5] Moreau, Luc, and Paolo Missier. "Prov-dm: The prov data model." (2013).

[6] Brickley, Dan, and Libby Miller. "FOAF vocabulary specification 0.98."Namespace Document 9 (2012).

[7] MathML. Available at: http://www.w3.org/Math/ (Accessed 01/02/2015)

[8] Hodgson, Ralph, and Paul J. Keller. "QUDT-quantities, units, dimensions and data types in OWL and XML." Online (September 2011) http://www. qudt. org(2011).

[9] Vinh Nguyen, Olivier Bodenreider, Amit Sheth. Don't like RDF Reification? Making Statements about Statements using Singleton Property. In Proceedings of the 23rd International Conference on World Wide Web. ACM, 2014.

[10] Krötzsch, Markus, Denny Vrandečić, and Max Völkel. "SMW." The Semantic Web-ISWC 2006. Springer Berlin Heidelberg, 2006. 935-942.

[11] Department of Defense Handbook, Metallic Materials and Elements for Aerospace Vehicle Structures, 2003.

[12] Nguyen, Vinh, et al. "The knowledge-driven exploration of integrated biomedical knowledge sources facilitates the generation of new hypotheses." (2011).

APPENDIX A - Vocabularies Used in the Semantic Model

Table A-1: List of vocabularies being assessed for semantic modelling

Name	Abbreviation	Namespace URI
Simple Knowledge Organization System	skos:	http://www.w3.org/2004/02/skos/core#
DCMI Metadata Terms	dcterms:	http://purl.org/dc/terms/
W3C PROVenance Interchange	prov:	http://www.w3.org/ns/prov#
Friend of a Friend	foaf:	http://xmlns.com/foaf/0.1/
Vocabulary for Attaching Essential Metadata	vaem:	http://www.linkedmodel.org/1.2/schema/vaem#
Vocabulary Of Attribution and Governance	voag:	http://voag.linkedmodel.org/1.0/owl/schema/voag
Quantities, Units, Dimensions and Types	qudt:	http://qudt.org/1.1/vocab
Vocabulary of a Friend	vaof:	http://purl.org/vocommons/voaf#
DCMI Type Vocabulary	dctype:	http://purl.org/dc/dcmitype/
Mathematical Markup Language	mathml:	http://www.w3.org/1998/Math/MathML

APPENDIX B - Definition Elements Models

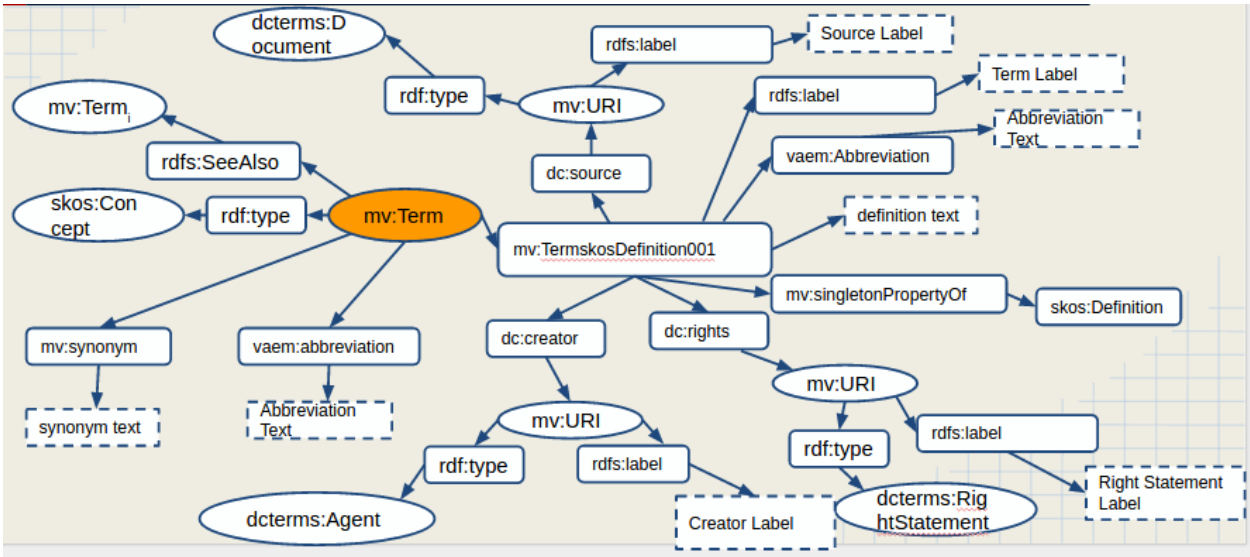

Figure B-1. Definition Text, Abbreviation and Synonym Model

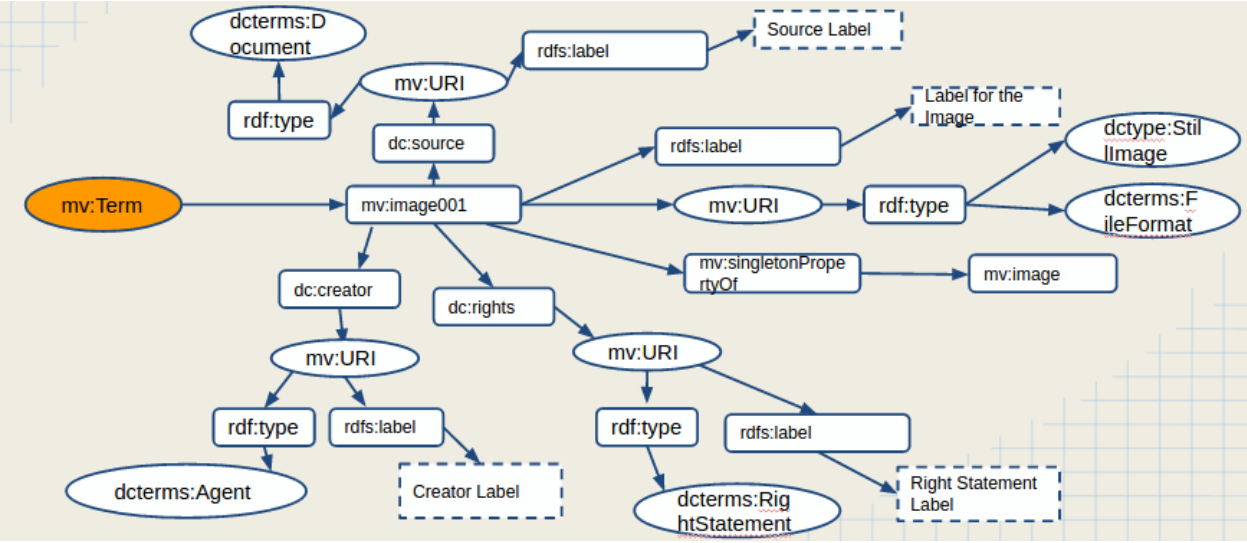

Figure B-2. Image Model

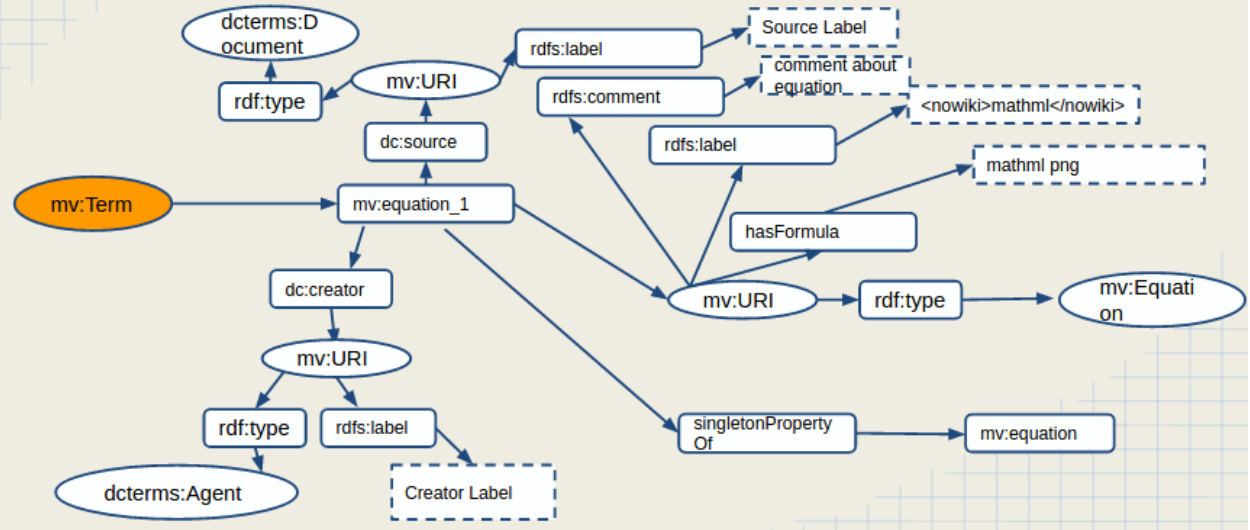

Figure B-3. Equation Model

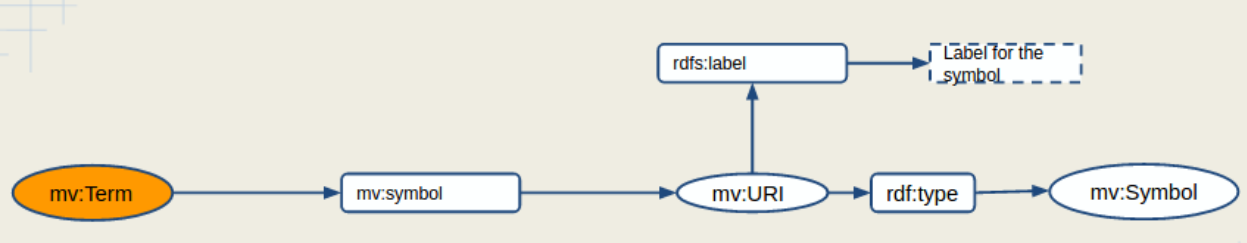

Figure B-4. Symbol Model

LIST OF ACRONYMS

AFRL	Air Force Research Laboratory
CSV	Comma Seperated Variable
DTIC	Defense Technical Information Center
FOAF	Friend of a Friend Ontology
ICD	International Classification of Disease
MathML	Mathematical Markup Language
MatVocab	Materials Vocabulary
MeSH	Medical Subject Headings
MGI	Materials Genome Initiative
PDB	Protein Data Bank
PROV	The Provenance Ontology
PUBMED	U.S. National Library of Medicine Website
QUDT	Quantities, Units, Dimensions and Data Types Ontology
RDF	Resource Description Framework

SKOS	Simple Knowledge Organization System
SMW	Semantic MediaWiki
SNOMED	Systematized Nomenclature of Human Medicine
SPARQL	SPARQL Protocol and RDF Query Language