

# Swoogle's Metadata about the Semantic Web

Lushan Han, Li Ding, Rong Pan, Tim Finin  
May 2006

## Abstract

Semantic Web technology enables us to specify metadata about things in the world; however, the metadata of the Semantic Web on the Web is also important for accessing online Semantic Web data. In this paper, we show how Swoogle collects the metadata from Semantic Web document to build a global picture of the Semantic Web. We also show the database schema that stores the metadata, and use an example Semantic Web document to explain how the metadata has been extracted and stored.

## Keywords

Semantic Web, metadata, Swoogle

## 1 Introduction

Swoogle [Ding et al., 2004] [Ding et al., 2005] is a Semantic Web search engine that discovers, indexes, analyzes and searches Semantic Web documents and Semantic Web terms published on the Web. It aims at enhancing Semantic Web surfing and facilitating Web-scale Semantic Web data access. By March 2006, Swoogle has indexed 1.3 million of Semantic Web documents and 1.4 million of distinct Semantic Web terms.

When indexing Semantic Web documents and terms, Swoogle builds the metadata about the Semantic Web by analyzing the semantic content and structure of Semantic Web documents. In general, Swoogle's metadata consists of the following categories:

- the annotation metadata about an Semantic Web document, e.g. URL and length
- the annotation metadata about an Semantic Web term, e.g., URI and local-name
- meta-usage of terms in documents
- instantiation of *rdfs:domain* and *rdfs:range* in instance data
- triples that have contribution to term definition

## 2 Definitions

### 2.1 Basic Concepts

#### **Definition 1 (Semantic Web document (SWD))**

Semantic Web Document is a class of Web documents serializing one or several RDF graphs. The URL of a Semantic Web document has three senses: (i) the address of the document on the Web, (ii) the unique identifier of the document in RDF graph world, and (iii) the unique identifier of the RDF graph serialized in the document.

#### RDF grammar

Currently, three RDF grammars have been recommended by W3C to syntactically serialize Semantic Web documents:

- RDF/XML, see <http://www.w3.org/TR/rdf-syntax-grammar/>
- N-Triples, see <http://www.w3.org/TR/rdf-testcases/#ntriples>
- N3, see <http://www.w3.org/DesignIssues/Notation3>

#### Embedded RDF

A Semantic Web document can be classified as pure or embedded based on whether its entire content is encoded using RDF grammar. For example, some HTML documents may embed a small RDF graph (serialized by RDF/XML) that stores Creative Commons License data.

#### Filetype extensions

The widely known filetype extensions of Semantic Web document are 'rdf', 'owl', 'rdfs', 'nt', and 'n3'. We also noticed that 'rss', 'foaf' and 'xml' are frequently used. However, more than 60% SWDs indexed by Swoogle do not have filetype extensions.

#### **Definition 2 (Semantic Web Term (SWT))**

Semantic Web Term refers to a special class of RDF resources, each of which has valid URI reference and has meta-usage (e.g., being defined, referenced or populated as a class or a property) in at least one Semantic Web document.

#### **Definition 3 (Semantic Web Ontology (SWO))**

Semantic Web Ontology refers to a special class of Semantic Web documents that define at least one Semantic Web term.

#### **Definition 4 (Semantic Web Namespace)**

Semantic Web Namespace refers to a special class of named RDF resources, each of which has been used as the namespace of some RDF resources. Currently, we only consider the namespaces used by Semantic Web terms.

## 2.2 Meta-Usage: Class and Property

The semantics of a Semantic Web term depends on its usage in residential RDF graph. In general, the usage of a Semantic Web term is either a class or property.

**Class-Usage:** A class refers to a named *rdfs:Resource* which has been used as the instance of *rdfs:Class* in SWDs. We care about three types of class-usages:

### Being defined as a class

A resource *X* is defined as a class if there exists a triple like  $(X, rdfs:type, C)$  where *C* is *rdfs:subClassOf rdfs:Class*. For example, *foaf:Person* is defined as a class according to triple *t3* in Figure 1.

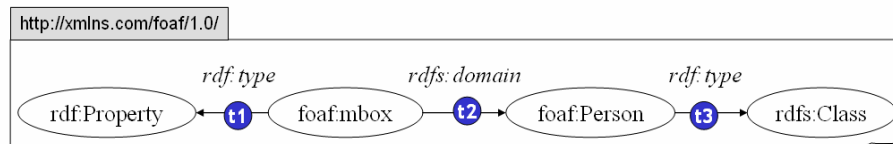


Figure 1 A example of meta-usage

### Being populated as a class

A resource *X* is populated (or instantiated) as a class if there exists a triple  $(R, rdfs:type, X)$  where *R* can be any resource. For example, *rdfs:Class* is populated as a class according to triple *t3* in Figure 1.

### Being referenced as a class

A resource *X* is referenced as a class in a triple if *X* is of type *rdfs:Class* according to the vocabulary of Semantic Web languages without the involvement of *rdfs:type*. For example, *foaf:Person* is referenced as a class by triple *t2* in Figure 1 because *rdfs:domain* has a known range *rdfs:Class*.

**Property-Usage:** A property refers to a named *rdfs:Resource* which has been used as the instance of *rdf:Property* in SWDs. We care about three types of property-usages:

### Being defined as a property

A resource *X* is defined as a property if there exists a triple  $(X, rdfs:type, P)$  where *P* is *rdfs:subClassOf rdf:Property*. For example, *foaf:mbox* is defined as a property by triple *t1* in Figure 1.

### Being populated as a property

A resource *X* is populated (or instantiated) as a property if there exists a triple  $(S, X, O)$  where *S* and *O* can be any resource (or literal). For example, *rdfs:type* is populated as a property by triple *t3* in Figure 1.

### Being referenced as a property

A resource *X* is referenced as a property in a triple if *X* is of type *rdf:Property* according to the vocabulary of Semantic Web languages without the involvement *rdfs:type*. For example, *foaf:mbox* is referenced as a property by triple *t2* in Figure 1.

### 2.3 Instantiation of domain/range Usage

As shown in previous section, the *predicate* in a triple indicates a property-usage (populated as a property) of a Semantic Web term. We may additionally observe the instantiations of *rdfs:domain* and *rdfs:range*.

In Figure 2, we observe the instantiation of a domain definition (*foaf:name rdfs:domain foaf:Person*) from triple t2 and t1: once *foaf:name* has been observed being populated as a property in triple t2, a user may pursue the background RDF graph for the *rdf:type* of the subject of t2 and thus get the match (i.e., t1).

Similarly, we may learn the instantiation of a range definition (*foaf:name rdfs:range rdf:Literal*) since "Tim Finin" is a literal *rdf:object* of *foaf:name*.

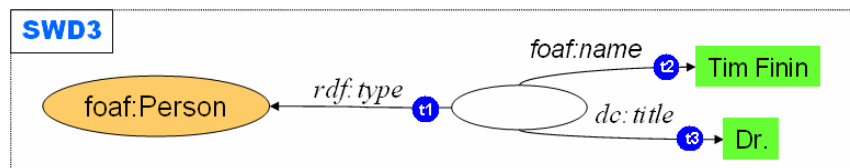


Figure 2 An example for instantiation of domain/range usage

## 3 Database Design and E-R Diagram

Swoogle stores its metadata in a relational database (we use MySQL 4.0.16). Table 1 briefly describes the tables in Swoogle's metadata database (swoogle3meta).

Table 1 A brief description of each table in Swoogle metadata database

| table name               | brief description  | identifiers (exclude surrogate key)        |
|--------------------------|--|--|
| digest_swd               | Metadata of SWD  | idurl and url                              |
| digest_swt               | Metadata of SWT  | uri  |
| digest_ns                | Metadata of SWN  | uri  |
| rel_swd_swt              | Relation between SWD and SWT   | (idswd, idswt)                             |
| rel_swd_ns_prefix        | Relation between SWD and SWN   | (idswd, idns)                              |
| rel_swd_swt_domain_range | All possible domain-range combinations associated with a SWT within a SWD and their frequencies. | (idswd, idswt, idswt_domain, idswt range)  |
| triple_swd_swd           | Triples describing a SWD. The object of these triples is an instance of <i>rdfs:Resource</i> .   | (idswd_source, subject, predicate, object) |
| triple_swd_annotation    | Triples describing a SWD, The object of these triples has the type <i>rdfs:Literal</i> .         |  |
| triple_swt_swt           | Triples defining a SWT. The object of these triples is an instance of <i>rdfs:Resource</i> .     |  |
| triple_swt_annotation    | Triples defining a SWT. The object of these triples has the type <i>rdfs:Literal</i> .           |  |
| digest_resource          | Resources used in SWT definition   | uid  |
| digest_host              | Metadata of host   | url and (protocol, name, port)             |

Figure 3 depicts the E-R diagram of Swoogle's metadata database. The tables with green background are more important than the ones with yellow. The detailed table description is available in Appendix C.

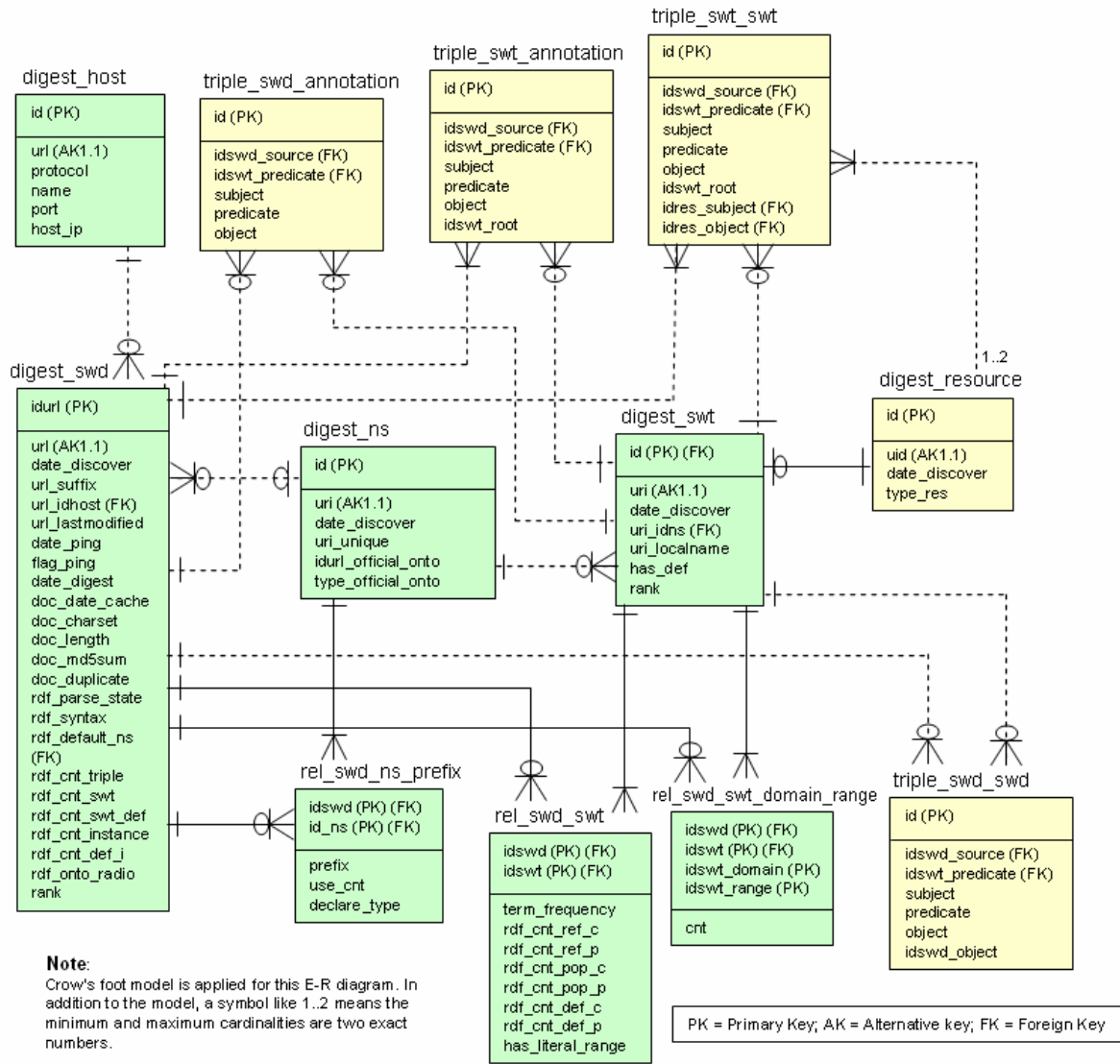


Figure 3 E-R diagram for Swoogle metadata database

## 4 An Example of Swoogle's Metadata

In what follows, we show how Swoogle's metadata is designed and extracted using an example SWD (<http://rdflib.net/2002/InformationStore>). Its original content is in Appendix A, and its N-Triples version is in Appendix B.

### 4.1 Metadata of Semantic Web Document

Swoogle stores three groups of metadata of an SWD in table `digest_swid`:

- Table 2 covers properties collected from crawling and http response
- Table 3 covers properties obtained from conventional document processing
- Table 4 covers properties obtained from RDF parsing

**Table 2 Metadata collected from crawling and HTTP response for the example SWD**

| Field                      | Value   | Note   |
|----------------------------|---|--|
| <code>idurl</code>         | 410868  | Primary key, unique id assigned to this SWD  |
| <code>url</code>           | <a href="http://rdflib.net/2002/InformationStore">http://rdflib.net/2002/InformationStore</a> | The URL of this SWD  |
| <code>date_discover</code> | 2005-02-17  | The date when Swoogle first found this SWD   |
| <code>url_suffix</code>    | ---   | This SWD does not have filetype extension  |
| <code>url_idhost</code>    | 30143   | Unique ID of the "host " part of this URL, also a foreign key to <code>digest_host</code> table. |
| <code>date_ping</code>     | 2006-01-29  | The date when Swoogle last accessed this SWD   |
| <code>flag_ping</code>     | 7   | 7 is the code indicating the SWD is still alive.   |

Notes:

1. `url_suffix`:

The filetype extension of the SWD. It is extracted from URL using several heuristics, e.g. "rdf" can be extracted from the URL "http://foo.com/example.rdf". However, many URLs have no extension, and we use '---' for this case.

2. `flag_ping`:

The status of the SWD according to Swoogle's last access (aka. ping). Possible values are:

- ❖ failed: 1, 2, 3, 4, 5, 6, 20 (see appendix for details)
- ❖ alive:
  - 7 (alive) unchanged;
  - 8 (changed from NSWD to SWD);
  - 9 (changed from SWD to NSWD)

**Table 3 Metadata Collected from conventional document processing for the example SWD**

| Field                       | Value                           | Note                                    |
|-----------------------------|---------------------------------|---|
| <code>doc_date_cache</code> | 2006-01-17                      | The date when Swoogle cached the SWD    |
| <code>doc_charset</code>    | UTF-8                           | The charset used to encode this SWD     |
| <code>doc_length</code>     | 2961                            | The length (in bytes) of the SWD        |
| <code>doc_md5sum</code>     | 6e88dfa4170804e73fd24bbc4d19ff9 | The MD5 hash of the content of the SWD. |

Notes:

1. `doc_charset`:

The charset used for encoding the content of the SWD. This field is critical for converting the byte stream into a proper string representation [Bos2006]. Swoogle uses several heuristics (e.g. the http response header, the first several bytes in byte stream, and the

encoding declaration in XML prolog) to obtain the right encoding. Possible values are "UTF-8", "ISO8859-1", "UTF-16", "SHIFT\_JIS", and so on.

**Table 4 Metadata Collected from RDF parse process for the example SWD**

| Field            | Value | Note   |
|------------------|-------|--|
| rdf_parse_state  | 25    | The code for RDF parsing result of this SWD is '25', which means this SWD is an error-free RDF document. |
| rdf_syntax       | RX    | The syntactic grammar used by this SWD is "RDF/XML".   |
| rdf_default_ns   | 0     | '0' means no default namespace.  |
| rdf_cnt_triple   | 34    | There are 34 triples in the SWD  |
| rdf_cnt_swt      | 20    | There are 20 unique SWTs in the SWD  |
| rdf_cnt_swt_def  | 10    | There are 10 unique SWTs being defined/referenced (not populated) as classes or properties in the SWD    |
| rdf_cnt_instance | 1     | The number of unique instances in the SWD (excluding class and properties)                               |
| rdf_onto_ratio   | 1     | 1 means all triples contribute to class/property definition in the SWD.                                  |

Notes:

1. rdf\_parse\_state:

**Table 5 Possible values of column rdf\_parse\_state**

|              | no RDF graph | RDF graph with parse error | RDF graph with parse warning | RDF graph w/o error |
|--------------|--------------|----------------------------|------------------------------|---------------------|
| not visit    | 0            |                            |                              |                     |
| visited nswd | 2            |                            |                              |                     |
| embedded swd |              | 12                         | 13                           | 15                  |
| swd          |              | 22                         | 23                           | 25                  |

2. rdf\_syntax:

The syntactic grammar used by an SWD. Possible values are: (i)'RE' (RDF/XML, embedded), (ii)'RX' (RDF/XML), (iii)'N3' (Notation 3), and (iv)'NT' (N-Triples).

3. rdf\_default\_ns:

Default namespace is extracted according to W3C standard [Bray et al. 1999]. It is a foreign key of table digest\_ns. When no default namespace is declared, it is set to 0.

4. rdf\_cnt\_swt:

It can be obtained from table rel\_swd\_swt using the following SQL query:

```
SELECT count(*) FROM rel_swd_swt WHERE idswd=410868
```

5. rdf\_cnt\_swt\_def:

It can be obtained from table rel\_swd\_swt using the following SQL query:

```
SELECT count(*) FROM rel_swd_swt WHERE idswd=410868
AND rdf_cnt_ref_c+rdf_cnt_ref_p + rdf_cnt_def_c + rdf_cnt_def_p > 0
```

6. rdf\_onto\_ratio:

The percentage of triples that contributed to ontological definition: (i) triples related to meta-usage (ii) triples related to definition of instances of *owl:Ontology*.

## 4.2 Metadata of Semantic Web Term

Table 6 lists 20 SWTs found in the example SWD:

**Table 6 Corresponding entries in table digest\_swt for the example SWD**

| ID     | uri   | date_discover | uri_idns | uri_localname | has_def |
|--------|---|---------------|----------|---------------|---------|
| 1      | http://www.w3.org/1999/02/22-rdf-syntax-ns#type     | 2006-02-10    | 2        | type          | 1       |
| 259    | http://www.w3.org/2000/01/rdf-schema#label          | 2006-02-10    | 14       | label         | 1       |
| 324    | http://www.w3.org/2002/07/owl#Ontology              | 2006-02-10    | 76       | Ontology      | 1       |
| 506    | http://www.w3.org/2000/01/rdf-schema#comment        | 2006-02-10    | 14       | comment       | 1       |
| 730    | http://www.w3.org/2000/01/rdf-schema#Class          | 2006-02-10    | 14       | Class         | 1       |
| 839    | http://www.w3.org/2000/01/rdf-schema#isDefinedBy    | 2006-02-10    | 14       | isDefinedBy   | 1       |
| 909    | http://www.w3.org/2000/01/rdf-schema#subClassOf     | 2006-02-10    | 14       | subClassOf    | 1       |
| 1334   | http://www.w3.org/2000/01/rdf-schema#domain         | 2006-02-10    | 14       | domain        | 1       |
| 1342   | http://www.w3.org/2000/01/rdf-schema#range          | 2006-02-10    | 14       | range         | 1       |
| 1419   | http://www.w3.org/1999/02/22-rdf-syntax-ns#Property | 2006-02-10    | 2        | Property      | 1       |
| 4999   | http://www.w3.org/2000/01/rdf-schema#Resource       | 2006-02-11    | 14       | Resource      | 1       |
| 14     | http://www.w3.org/2000/01/rdf-schema#Literal        | 2006-02-11    | 14       | Literal       | 1       |
| 129481 | http://rdflib.net/2002/InformationStore#UpdateEvent | 2006-02-13    | 2280     | UpdateEvent   | 1       |
| 129482 | http://rdflib.net/2002/InformationStore#updateEvent | 2006-02-13    | 2280     | updateEvent   | 1       |
| 129483 | http://www.w3.org/2000/01/rdf-schema#Context        | 2006-02-13    | 14       | Context       | 0       |
| 129484 | http://www.w3.org/2000/01/rdf-schema#UpdateEvent    | 2006-02-13    | 14       | UpdateEvent   | 0       |
| 129485 | http://rdflib.net/2002/InformationStore#Context     | 2006-02-13    | 2280     | Context       | 1       |
| 129486 | http://rdflib.net/2002/InformationStore#error       | 2006-02-13    | 2280     | error         | 1       |
| 129487 | http://rdflib.net/2002/InformationStore#source      | 2006-02-13    | 2280     | source        | 1       |
| 129488 | http://rdflib.net/2002/InformationStore#http_status | 2006-02-13    | 2280     | http_status   | 1       |

Notes:

1. How does Swoogle identify SWTs?

A SWT is either a class or property. The main task of Swoogle is to study the constitution and meta-usage of classes and properties (i.e. SWT) in the semantic web. Besides classes and properties, a much larger number of individual instances also exist in the semantic web. In practice, Swoogle identifies SWTs by investigating each triple according to the meta-usage definition (see section 2.2).

Consider the SWT highlighted in Table 6. We should be able to find the meta-usage of the URI ‘http://rdflib.net/2002/InformationStore#UpdateEvent’ in at least one triple in the example SWD, e.g., the triple in Table 7.

**Table 7 A triple showing a meta-usage of ‘http://rdflib.net/2002/InformationStore#UpdateEvent’**

| No. | subject   | predicate                                       | object                                     |
|-----|---|---|--|
| 9   | http://rdflib.net/2002/InformationStore#UpdateEvent | http://www.w3.org/1999/02/22-rdf-syntax-ns#type | http://www.w3.org/2000/01/rdf-schema#Class |

However, the URI ‘http://rdflib.net/2002/InformationStore’ does not have meta-usage in the triple in table 8. Since no other triples in the example SWD contain meta-usage of this URI, it is not counted as a SWT.

**Table 8 A triple not showing a meta-usage of ‘http://rdflib.net/2002/InformationStore’**

| No. | subject                                 | predicate                                       | object                                 |
|-----|---|---|--|
| 1   | http://rdflib.net/2002/InformationStore | http://www.w3.org/1999/02/22-rdf-syntax-ns#type | http://www.w3.org/2002/07/owl#Ontology |

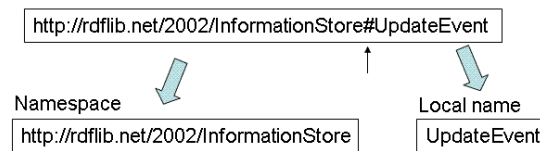


All SWTs in the example SWD can also be obtained from the following SQL query:

```
SELECT * FROM digest_swt
WHERE id IN (SELECT idswt FROM rel_swd_swt WHERE idswd = 410868)
```

## 2. Splitting the URI of a SWT

We use various heuristics to split a SWT's URI into a pair of namespace and local name. Among these heuristics, the simplest one is to identify the location of '#' in a SWT, and the string to the left of '#' is the namespace, the string to the right of '#' is the local name.



**Figure 4 Split the URI of a SWT**

If the extracted namespace is new to Swoogle, it will be inserted into table `digest_ns`, which maintains all discovered namespaces.

- column `uri_idns`: stores the foreign key id of the namespace portion.
- column `uri_localname`: stores the literal content of the local name portion

## 3. Column `date_discover`

One SWT may occur in multiple SWDs. The discover date is the first time Swoogle found the SWT and added new entry to table `digest_swt`. When Swoogle met the same SWT later in other SWDs, it would not modify the corresponding entry in table `digest_swt`.

## 4. Column `has_def`

The `has_def` column is set to 1 if there exists a SWD in our database, in which the target SWT has been defined or referenced. The value of `has_def` has a global range. It is not only confined to one SWD.

## 5. Case-sensitive

SWTs are case-sensitive. Table 9 shows two SWTs that are different.

**Table 9 Two different SWTs**

| id     | URI   | date_discover | uri_idns | uri_localname | has_def |
|--------|---|---------------|----------|---------------|---------|
| 129481 | http://rdflib.net/2002/InformationStore#UpdateEvent | 2006-02-13    | 2280     | UpdateEvent   | 1       |
| 129482 | http://rdflib.net/2002/InformationStore#updateEvent | 2006-02-13    | 2280     | updateEvent   | 1       |

### 4.3 Term Frequency and Meta Usage

Table 6 lists the term frequency and meta-usage frequency of the 20 SWTs in Table 10.

**Table 10** Corresponding entries in table rel\_swd\_swt for the example SWD

| idswd  | idswt  | term_frequency | rdf_cnt_ref_c | rdf_cnt_ref_p | rdf_cnt_pop_c | rdf_cnt_pop_p | rdf_cnt_def_c | rdf_cnt_def_p | has_literal_range |
|--------|--------|----------------|---------------|---------------|---------------|---------------|---------------|---------------|-------------------|
| 410868 | 1      | 7              | 0             | 0             | 0             | 2             | 7             | 0             | 0                 |
| 410868 | 14     | 2              | 2             | 0             | 0             | 0             | 0             | 0             | 0                 |
| 410868 | 259    | 7              | 0             | 0             | 0             | 0             | 7             | 0             | 1                 |
| 410868 | 324    | 1              | 0             | 0             | 1             | 0             | 0             | 0             | 0                 |
| 410868 | 506    | 3              | 0             | 0             | 0             | 3             | 0             | 0             | 1                 |
| 410868 | 730    | 2              | 0             | 0             | 2             | 0             | 0             | 0             | 0                 |
| 410868 | 839    | 7              | 0             | 0             | 0             | 7             | 0             | 0             | 0                 |
| 410868 | 909    | 2              | 0             | 0             | 0             | 2             | 0             | 0             | 0                 |
| 410868 | 1334   | 4              | 0             | 0             | 0             | 5             | 4             | 0             | 0                 |
| 410868 | 1342   | 4              | 0             | 0             | 0             | 4             | 0             | 0             | 0                 |
| 410868 | 1419   | 4              | 0             | 0             | 3             | 4             | 0             | 0             | 0                 |
| 410868 | 4999   | 4              | 4             | 0             | 0             | 0             | 0             | 0             | 0                 |
| 410868 | 129481 | 2              | 1             | 0             | 0             | 0             | 1             | 0             | 0                 |
| 410868 | 129482 | 3              | 0             | 4             | 2             | 0             | 0             | 1             | 0                 |
| 410868 | 129483 | 1              | 6             | 1             | 0             | 0             | 0             | 0             | 0                 |
| 410868 | 129484 | 3              | 3             | 0             | 0             | 0             | 0             | 0             | 0                 |
| 410868 | 129485 | 2              | 1             | 0             | 0             | 0             | 1             | 0             | 0                 |
| 410868 | 129486 | 3              | 0             | 2             | 0             | 0             | 0             | 1             | 0                 |
| 410868 | 129487 | 3              | 0             | 2             | 0             | 0             | 0             | 1             | 0                 |
| 410868 | 129488 | 3              | 0             | 2             | 0             | 0             | 0             | 1             | 0                 |

Notes:

1. Keys

The primary key of table rel\_swd\_swt is (idswd, idswt). The idswt is a foreign key referencing digest\_swt.id. The idswd is a foreign key referencing digest\_swt.idurl.

2. Column has\_literal\_range

'1' means the range of the SWT represented by a row is 'rdfs:Literal'. '0' means the range is a subclass of 'rdfs:resource'.

3. Counting meta usage

We give three example triples to illustrate how to count meta-usages. Every meta-usage of a SWT increases the corresponding field by one.

**Example 1: definition and instantiation**

Table 11 shows a triple that contributes three meta-usages (definition and instantiation).

**Table 11** Example triple: definition and instantiation

| No. | subject   | predicate                                       | object  |
|-----|---|---|---|
| 18  | http://rdflib.net/2002/InformationStore#updateEvent | http://www.w3.org/1999/02/22-rdf-syntax-ns#type | http://www.w3.org/1999/02/22-rdf-syntax-ns#Property |

- Being defined as a property  
 'http://rdflib.net/2002/InformationStore#updateEvent' is counted as being defined as a property because the predicate is 'rdf:type' and the object is 'rdf:Property' (see

Section 2.2 for defining of meta-usage). This meta-usage increases the cell labeled by ① in the table 10 by one.

- Being populated/instantiated as a property  
‘http://www.w3.org/1999/02/22-rdf-syntax-ns#type’ is counted as being populated as a property because it is the predicate in the triple. This meta-usage increases the cell labeled by ② in table 10 by one.
- Being populated/instantiated as a class  
‘http://www.w3.org/1999/02/22-rdf-syntax-ns#Property’ is counted as being populated as a class because it is the object in a triple having the form (x, ‘rdfs:Type’, y). This meta-usage increases the cell labeled by ③ in table 10 by one.

### **Example 2: reference and instantiation**

Table 12 shows a triple that contributes three meta-usages (reference and instantiation).

**Table 12 Example triple: reference and instantiation**

| No. | subject   | predicate                                   | object                                       |
|-----|---|---|--|
| 20  | http://rdflib.net/2002/InformationStore#updateEvent | http://www.w3.org/2000/01/rdf-schema#domain | http://www.w3.org/2000/01/rdf-schema#Context |

- Being referenced as a property:  
‘http://rdflib.net/2002/InformationStore#updateEvent’ is counted as being referenced as a property because this triple has the form (x, *rdfs:domain*, y), and *rdfs:domain* is a property whose domain is *rdf:Property*. Consequently, we can infer that ‘x’ is a property. This meta-usage increases the cell labeled by ④ in table 10 by one.
- Being populated/instantiated as property:  
‘http://www.w3.org/2000/01/rdf-schema#domain’ is counted as being populated as a property because it is the predicate in the triple. This meta-usage increases the field labeled by ⑤ in table 10 by one.
- Being referenced as a class:  
‘http://www.w3.org/2000/01/rdf-schema#Context’ is counted as being referenced as a class. This triple has the form (x, ‘rdfs:domain’, y), where ‘rdfs:domain’ is a property whose range is *rdfs:Class*. Consequently, we can infer that ‘y’ is a class. This meta-usage increases the field labeled by ⑥ in table 10 by one.

### **Example 3: no meta-usage**

Table 13 shows a triple that contributes only one meta-usage (i.e., the predicate "rdfs:label" is populated/instantiated as a property). The subject has no meta-usage in this triple because from the domain/range definitions of ‘rdfs:label’, we can only know the subject is a resource but cannot further infer whether it is a class or a property. According to our meta usage definition, the subject is neither ‘being defined as SWT’ nor ‘being referenced as SWT’. Moreover, the literal "update event" does not contribute meta-usage either since it is a literal.

**Table 13 An example triple for explaining meta usage**

| No. | subject   | predicate                                  | object         |
|-----|---|--|----------------|
| 19  | http://rdflib.net/2002/InformationStore#updateEvent | http://www.w3.org/2000/01/rdf-schema#label | "update event" |

## 4.4 Counting domain/range usage in instance data

We count the domain/range usage only for SWTs defined in most popular namespaces, such as RDF, RDFS and OWL.

**Table 14** Corresponding entries in table `rel_swd_swt_domain_range` for the example SWD

| idswd  | idswt | idswt_domain | idswt_range | cnt |
|--------|-------|--------------|-------------|-----|
| 410868 | 259   | 324          | 14          | 1   |
| 410868 | 259   | 730          | 14          | 2   |
| 410868 | 259   | 1419         | 14          | 4   |
| 410868 | 506   | 324          | 14          | 1   |
| 410868 | 506   | 1419         | 14          | 2   |
| 410868 | 839   | 324          | 324         | 1   |
| 410868 | 839   | 730          | 324         | 2   |
| 410868 | 839   | 1419         | 324         | 4   |
| 410868 | 909   | 730          | 0           | 2   |
| 410868 | 1334  | 1419         | 0           | 4   |
| 410868 | 1342  | 1419         | 0           | 4   |

Notes:

1. Swoogle skips triples whose predicate are "rdf:type". We do not count triples in the form of (?x rdf:type ?y) for domain/range usage. For example, Swoogle skips the triple shown in table 15.

**Table 15** Example triple: Swoogle does not count domain/range usage

| No. | subject   | predicate                                       | object                                     |
|-----|---|---|--|
| 5   | http://rdflib.net/2002/InformationStore#Context | http://www.w3.org/1999/02/22-rdf-syntax-ns#type | http://www.w3.org/2000/01/rdf-schema#Class |

2. How to extract domain/range usage from instance data.

Given triple 8 in Table 16, we explore domain-usage and range-usage of the property "rdfs:isDefinedBy" by finding the rdf:type of the triple's subject and object respectively:

- triple 5 indicates the rdf:type of the subject is "rdfs:Class"
- triple 1 indicates the rdf:type of the object is "owl:Ontology"

Since the 'idswt' of 'rdfs:isDefinedBy', 'rdfs:Class', and 'owl:Ontology' are '839', '730', and '324' respectively, we can increase the "cnt" field of the highlighted entry in table 14 to store the case of instantiation by one.

**Table 16** Examples triples: how Swoogle count domain/range usage

| No | subject   | predicate  | object                                     |
|----|---|--|--|
| 8  | http://rdflib.net/2002/InformationStore#Context | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore    |
| 5  | http://rdflib.net/2002/InformationStore#Context | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/2000/01/rdf-schema#Class |
| 1  | http://rdflib.net/2002/InformationStore         | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/2002/07/owl#Ontology     |

## 4.5 Namespaces in SWD

Table 17 lists all namespaces involved in the example SWD.

**Table 17 Corresponding entries in table digest\_ns for the example SWD**

| id   | uri   | date_discover | uri_unique                                 | idurl_official_onto | type_official_onto |
|------|---|---------------|--|---------------------|--------------------|
| 2    | http://www.w3.org/1999/02/22-rdf-syntax-ns# | 2006-02-10    | http://www.w3.org/1999/02/22-rdf-syntax-ns | 3,581               | 2                  |
| 14   | http://www.w3.org/2000/01/rdf-schema#       | 2006-02-10    | http://www.w3.org/2000/01/rdf-schema       | 3,583               | 2                  |
| 76   | http://www.w3.org/2002/07/owl#              | 2006-02-10    | http://www.w3.org/2002/07/owl              | 3,580               | 2                  |
| 2280 | http://rdflib.net/2002/InformationStore#    | 2006-02-13    | http://rdflib.net/2002/InformationStore    | 410,868             | 2                  |

Notes:

### 1. Where namespaces are extracted

We have discussed how to split a SWT into a pair of namespace and local name. If the extracted namespace does not exist in the table digest\_ns, a new entry for this namespace will be added.

### 2. uri\_unique

The uri\_unique is just a revised URI by removing the *fragment* (see RFC URI definition [Berners-Lee et al. 1998]) starting from '#', and does not affect web addressing.

### 3. idurl\_official\_onto

This column points to an official ontology defining the namespace. Official ontology is derived by Swoogle from SWDs collected in the semantic web.

## 4.6 Relations between SWD and SWN

Table 18 lists three namespaces used or declared in the example SWD.

**Table 18 Corresponding entries in table rel\_swd\_ns\_prefix for the example SWD**

| idswd  | idns | prefix | use_cnt | declare_type |
|--------|------|--------|---------|--------------|
| 410868 | 2    | rdf    | 11      | 1            |
| 410868 | 14   | rdfs   | 39      | 1            |
| 410868 | 76   | owl    | 1       | 1            |
| 410868 | 2280 |        | 16      | 0            |

Notes:

### 1. Keys

The primary key of table rel\_swd\_ns\_prefix is a composite key of 'idswd' and 'idns'. 'idswd' is a foreign key referencing the primary key of table digest\_swd. 'idns' is a foreign key referencing the primary key of table digest\_ns.

## 2. Count namespace-usage

Swoogle counts the occurrence of a namespace only for SWTs that have meta-usage. Given a namespace X, the value of ‘use\_cnt’ may be less than the actual number of the occurrences of RDF resources using X in the entire SWD. For example, the number of occurrences of the namespace ‘http://rdflib.net/2002/InformationStore’ in the N-triples version of the example SWD (see appendix B) is larger than the corresponding field highlighted in Table 18.

Given an SWT “http://rdflib.net/2002/InformationStore#Context”, table 19 lists all triples which have the SWT as the subject. Since "http://rdflib.net/2002/InformationStore" is used as the namespace of the SWT, each occurrence of the SWT may contribute one usage of the namespace; however, we selectively count the usage of the namespace as the following:

- In triple 5, the SWT is defined as a class; therefore, the namespace is counted.
- In triple 6, the SWT has no meta-usage; therefore, the namespace is not counted.
- In triple 7, the SWT is referenced as a class; therefore, the namespace is counted.
- In triple 8, the SWT has no meta-usage; therefore, the namespace is not counted.

**Table 19. Example triples for explaining namespace count**

| No. | subject   | predicate  | object  |
|-----|---|--|---|
| 5   | http://rdflib.net/2002/InformationStore#Context | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/2000/01/rdf-schema#Class    |
| 6   | http://rdflib.net/2002/InformationStore#Context | http://www.w3.org/2000/01/rdf-schema#label       | "Context"                                     |
| 7   | http://rdflib.net/2002/InformationStore#Context | http://www.w3.org/2000/01/rdf-schema#subClassOf  | http://www.w3.org/2000/01/rdf-schema#Resource |
| 8   | http://rdflib.net/2002/InformationStore#Context | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore       |

In fact, we can write a SQL query to derive the "use\_cnt" value indirectly from table rel\_swd\_swt by summing up the frequency of terms under the given namespace.

```
SELECT SUM(term_frequency)
FROM rel_swd_swt, digest_swt, digest_ns
WHERE rel_swd_swt.idswt = digest_swt.id
AND digest_swt.uri_idns = digest_ns.id
AND digest_ns.uri = 'http://rdflib.net/2002/InformationStore#'
AND rel_swd_swt.idswd = '410868';
```

## 3. declare type

The column is about how the namespace is declared in the SWD. Possible values are: 0: not declared but being used; 1: declared; 2: declared as default namespace.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#">
```

**Figure 5. The prefix declaration in the example SWD**

As shown in the Figure 5, ‘rdf,’ ‘rdfs,’ and ‘owl’ are declared as prefixes that substitute corresponding namespaces. No default namespace has been declared. The namespace ‘http://rdflib.net/2002/InformationStore’ is not declared.

## 4.7 Triples

Swoogle specially records some triples in SWDs because they are used to define/describe SWDs or SWTs. Swoogle classifies those triples into four groups and stores them in the following four tables:

- triple\_swd\_swd
- triple\_swd\_annotation
- triple\_swt\_swt
- triple\_swt\_annotation

**Common structure:** All the four tables share a common structure and have their own extensions. The general structure of a triple table is listed in Table 20:

**Table 20 Common columns used by triple tables**

| column name      | data type               | column description  |
|------------------|-------------------------|---|
| id:              | int(11)                 | Surrogate key for stored RDF triples.                     |
| idswd_source:    | int(11)                 | The unique id of the SWD containing this triple.          |
| idswt_predicate: | int(11)                 | The unique id of the predicate of this triple.            |
| subject:         | varchar(250)            | The URI of the subject.                                   |
| predicate        | varchar(250)            | The URI of the predicate. (redundant to avoid table join) |
| * object         | varchar(250)<br>OR text | The URI of the object.<br>The text content of the object. |

\* object: an object of a triple can be either a resource or a literal. We store the triples having literal-object and resource-object in different tables for (i) storage efficiency (ii) link (when the object resource is an SWD, we can record its SWD id in another column).

### 4.7.1 Table triple\_swd\_swd, triple\_swd\_annotation

Table 21 and Table 22 together record triples that contribute descriptions for the example SWD. In practice, we use triple\_swd\_annotation to store triples with literal object, and use triple\_swd\_swd for triples with resource object. In table 21, if the object URI stands for an existing SWD, the idswd\_object stores the id for the SWD. Otherwise, '0' is used.

**Table 21. Corresponding entries in table triple\_swd\_swd**

| id     | idswd_source | idswt_predicate | subject                                 | predicate  | object                                  | idswd_object |
|--------|--------------|-----------------|---|--|---|--------------|
| 537737 | 410868       | 839             | http://rdflib.net/2002/InformationStore | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore | 410868       |

**Table 22 Corresponding entries in table triple\_swd\_annotation**

| id     | idswd_source | idswt_predicate | subject                                 | predicate                                    | Object  |
|--------|--------------|-----------------|---|--|---|
| 738405 | 410868       | 259             | http://rdflib.net/2002/InformationStore | http://www.w3.org/2000/01/rdf-schema#label   | InformationStore                                      |
| 738404 | 410868       | 506             | http://rdflib.net/2002/InformationStore | http://www.w3.org/2000/01/rdf-schema#comment | A document defining a vocabulary used by an Inform... |

## 4.7.2 triple\_swt\_swt, triple\_swt\_annotation

In Table 23 and Table 24, Swoogle records triples that contribute definitions of SWTs in the example SWD.

**Table 23. Corresponding entries in table triple\_swt\_swt**

| id      | idswd_source | idswt_predicate | subject   | predicate  | object   | idswt_root | idres_subject | idres_object |
|---------|--------------|-----------------|---|--|--|------------|---------------|--------------|
| 8090939 | 410868       | 839             | http://rdflib.net/2002/InformationStore#Context     | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore          | 129485     | 129485        | 129489       |
| 8090938 | 410868       | 1342            | http://rdflib.net/2002/InformationStore#error       | http://www.w3.org/2000/01/rdf-schema#range       | http://www.w3.org/2000/01/rdf-schema#Literal     | 129486     | 129486        | 14           |
| 8090937 | 410868       | 1334            | http://rdflib.net/2002/InformationStore#http_status | http://www.w3.org/2000/01/rdf-schema#domain      | http://www.w3.org/2000/01/rdf-schema#UpdateEvent | 129488     | 129488        | 129484       |
| ...     | ...          | ...             | ...   | ...  | ...  | ...        | ...           | ...          |

**Table 24. Corresponding entries in table triple\_swt\_annotation**

| id      | idswd_source | idswt_predicate | subject   | predicate                                  | object       | idswt_root |
|---------|--------------|-----------------|---|--|--------------|------------|
| 2808604 | 410868       | 259             | http://rdflib.net/2002/InformationStore#updateEvent | http://www.w3.org/2000/01/rdf-schema#label | update event | 129482     |
| 2808603 | 410868       | 259             | http://rdflib.net/2002/InformationStore#source      | http://www.w3.org/2000/01/rdf-schema#label | source       | 129487     |
| ...     | ...          | ...             | ...   | ...  | ...          | ...        |

Note:

### 1. idres\_subject and idres\_object

These two columns contain the unique resource id for the subject and object of the target triple respectively. The reason why we use resource id instead of SWT id is that subject and object of a triple can be anonymous node or resources which are not identified as SWT.

### 2. Anonymous resource in definition triples

Anonymous resources (aka. blank-node [Klyne et al. 2004]) can be involved in defining a class. Swoogle also stores triples containing anonymous resources which contribute to the definition of a class. We can trace anonymous resources in triples to find essential terminal nodes, which is either a URI or literal, to make clear how the definition of a class is constructed. We will use an example to illustrate this feature.

The example RDF graph in Figure 6 demonstrates how class ‘owl:Thing’ is defined in the SWD ‘http://www.w3.org/2002/07/owl’.



```

<Class rdf:ID="Thing">
  <rdfs:label>Thing</rdfs:label>
  <unionOf rdf:parseType="Collection">
    <Class rdf:about="#Nothing"/>
    <Class>
      <complementOf rdf:resource="#Nothing"/>
    </Class>
  </unionOf>
</Class>

```

**Figure 6. The definition of ‘owl:Thing’ in ‘http://www.w3.org/2002/07/owl’**

The corresponding triples which compose the definition of the class ‘owl:Thing’ can be easily collected with a SQL query to the Swoogle database.

```

SELECT *
FROM triple_swt_swt
WHERE idswt_root = 672 and idswd_source = 3580

```

where 672 is the unique id of the SWT ‘owl:Thing’ and 3580 is the unique id of the SWD ‘http://www.w3.org/2002/07/owl’. The result is in Table 25.

**Table 25. A list of triples contributing to the definition of ‘owl:Thing’ in Swoogle database**

| id      | idswd_source | idswt_predicate | subject                             | predicate  | object   | idswt_root | idres_subject | idres_object |
|---------|--------------|-----------------|-------------------------------------|--|--|------------|---------------|--------------|
| 6832239 | 3580         | 7522            | swd:3580#93df2c:109a3655232:-672b   | http://www.w3.org/2002/07/owl#complementOf       | http://www.w3.org/2002/07/owl#Nothing          | 672        | 1956347       | 7666         |
| 6832218 | 3580         | 1               | swd:3580#93df2c:109a3655232:-672b   | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/2002/07/owl#Class            | 672        | 1956347       | 950          |
| 6832211 | 3580         | 4746            | http://www.w3.org/2002/07/owl#Thing | http://www.w3.org/2002/07/owl#unionOf            | swd:3580#93df2c:109a3655232:-672c              | 672        | 672           | 1956345      |
| 6832210 | 3580         | 1               | http://www.w3.org/2002/07/owl#Thing | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/2002/07/owl#Class            | 672        | 672           | 950          |
| 6832182 | 3580         | 954             | swd:3580#93df2c:109a3655232:-672a   | http://www.w3.org/1999/02/22-rdf-syntax-ns#rest  | http://www.w3.org/1999/02/22-rdf-syntax-ns#nil | 672        | 1956346       | 960          |
| 6832174 | 3580         | 954             | swd:3580#93df2c:109a3655232:-672c   | http://www.w3.org/1999/02/22-rdf-syntax-ns#rest  | swd:3580#93df2c:109a3655232:-672a              | 672        | 1956345       | 1956346      |
| 6832163 | 3580         | 951             | swd:3580#93df2c:109a3655232:-672c   | http://www.w3.org/1999/02/22-rdf-syntax-ns#first | http://www.w3.org/2002/07/owl#Nothing          | 672        | 1956345       | 7666         |
| 6832255 | 3580         | 951             | swd:3580#93df2c:109a3655232:-672a   | http://www.w3.org/1999/02/22-rdf-syntax-ns#first | swd:3580#93df2c:109a3655232:-672b              | 672        | 1956346       | 1956347      |

The field idswt\_root stores the unique id of the SWT being defined. Its value does not have to be the same as the value of "idres\_subject" due to the existence of blank node. In fact, when we encounter blank node in term definition, we always pursue the definition of the blank node to ensure the completeness of the definition.

## 4.8 Resources related to SWT definition (digest\_resource)

**Table 26** Corresponding entries in table `digest_resource` for the example SWD

| id     | uid   | date discover | type res |
|--------|---|---------------|----------|
| 1      | <a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#type">http://www.w3.org/1999/02/22-rdf-syntax-ns#type</a>             | 2006-02-10    | -2       |
| 14     | <a href="http://www.w3.org/2000/01/rdf-schema#Literal">http://www.w3.org/2000/01/rdf-schema#Literal</a>                   | 2006-02-10    | -3       |
| 259    | <a href="http://www.w3.org/2000/01/rdf-schema#label">http://www.w3.org/2000/01/rdf-schema#label</a>                       | 2006-02-10    | -2       |
| 324    | <a href="http://www.w3.org/2002/07/owl#Ontology">http://www.w3.org/2002/07/owl#Ontology</a>                               | 2006-02-10    | -2       |
| 506    | <a href="http://www.w3.org/2000/01/rdf-schema#comment">http://www.w3.org/2000/01/rdf-schema#comment</a>                   | 2006-02-10    | -2       |
| 730    | <a href="http://www.w3.org/2000/01/rdf-schema#Class">http://www.w3.org/2000/01/rdf-schema#Class</a>                       | 2006-02-10    | -2       |
| 839    | <a href="http://www.w3.org/2000/01/rdf-schema#isDefinedBy">http://www.w3.org/2000/01/rdf-schema#isDefinedBy</a>           | 2006-02-10    | -2       |
| 909    | <a href="http://www.w3.org/2000/01/rdf-schema#subClassOf">http://www.w3.org/2000/01/rdf-schema#subClassOf</a>             | 2006-02-10    | -2       |
| 1334   | <a href="http://www.w3.org/2000/01/rdf-schema#domain">http://www.w3.org/2000/01/rdf-schema#domain</a>                     | 2006-02-10    | -2       |
| 1342   | <a href="http://www.w3.org/2000/01/rdf-schema#range">http://www.w3.org/2000/01/rdf-schema#range</a>                       | 2006-02-10    | -2       |
| 1419   | <a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#Propert...">http://www.w3.org/1999/02/22-rdf-syntax-ns#Propert...</a> | 2006-02-10    | -2       |
| 4999   | <a href="http://www.w3.org/2000/01/rdf-schema#Resource">http://www.w3.org/2000/01/rdf-schema#Resource</a>                 | 2006-02-11    | -2       |
| 129481 | <a href="http://rdflib.net/2002/InformationStore#UpdateEven...">http://rdflib.net/2002/InformationStore#UpdateEven...</a> | 2006-02-13    | -2       |
| 129482 | <a href="http://rdflib.net/2002/InformationStore#updateEven...">http://rdflib.net/2002/InformationStore#updateEven...</a> | 2006-02-13    | -2       |
| 129483 | <a href="http://www.w3.org/2000/01/rdf-schema#Context">http://www.w3.org/2000/01/rdf-schema#Context</a>                   | 2006-02-13    | -2       |
| 129484 | <a href="http://www.w3.org/2000/01/rdf-schema#UpdateEvent">http://www.w3.org/2000/01/rdf-schema#UpdateEvent</a>           | 2006-02-13    | -2       |
| 129485 | <a href="http://rdflib.net/2002/InformationStore#Context">http://rdflib.net/2002/InformationStore#Context</a>             | 2006-02-13    | -2       |
| 129486 | <a href="http://rdflib.net/2002/InformationStore#error">http://rdflib.net/2002/InformationStore#error</a>                 | 2006-02-13    | -2       |
| 129487 | <a href="http://rdflib.net/2002/InformationStore#source">http://rdflib.net/2002/InformationStore#source</a>               | 2006-02-13    | -2       |
| 129488 | <a href="http://rdflib.net/2002/InformationStore#http_statu...">http://rdflib.net/2002/InformationStore#http_statu...</a> | 2006-02-13    | -2       |
| 129489 | <a href="http://rdflib.net/2002/InformationStore">http://rdflib.net/2002/InformationStore</a>                             | 2006-02-13    | -2       |

### 1. What resources are stored in the Swoogle

Swoogle only stores a subset of RDF resources of a SWD in table `digest_resource`. The resources are chosen because they are either SWTs or used in SWT definitions. Consider the triple in Table 27, whose subject is an identified SWT. This triple is a description or qualification on the SWT “<http://rdflib.net/2002/InformationStore#Context>”. In Swoogle, we record this triple and the object resource “<http://rdflib.net/2002/InformationStore>”, even though it is not a SWT.

**Table 27.** A triple showing what resource Swoogle stores

| No. | subject   | predicate   | object  |
|-----|---|---|---|
| 8   | <a href="http://rdflib.net/2002/InformationStore#Context">http://rdflib.net/2002/InformationStore#Context</a> | <a href="http://www.w3.org/2000/01/rdf-schema#isDefinedBy">http://www.w3.org/2000/01/rdf-schema#isDefinedBy</a> | <a href="http://rdflib.net/2002/InformationStore">http://rdflib.net/2002/InformationStore</a> |

### 2. What resources are not stored

If both the subject and object has no meta-usage in a triple, and the subject of a triple is not an SWT identified by any triple in a SWD, we disregard this triple and any resources appearing in this triple.

According to the triple in Table 28, neither “<http://rdflib.net/2002/InformationCenter>” nor “<http://rdflib.net/2002/InformationStore>” have meta-usage. Hence, the triple will not be added to table `triple_swt_swt`, and the former will not be added to table `digest_resource`.

**Table 28.** A triple showing what resource Swoogle don’t store

| subject   | predicate   | object  |
|---|---|---|
| <a href="http://rdflib.net/2002/InformationStore">http://rdflib.net/2002/InformationStore</a> | <a href="http://www.w3.org/2000/01/rdf-schema#isDefinedBy">http://www.w3.org/2000/01/rdf-schema#isDefinedBy</a> | <a href="http://rdflib.net/2002/InformationCenter">http://rdflib.net/2002/InformationCenter</a> |

### 3. Handling an anonymous resource

Consider the entry in Table 29, which records an anonymous resource. This anonymous resource must also appear in a triple stored in table `triple_swt_swt`. The column ‘`type_res`’ stores the id of the SWD containing this anonymous resource. The string shown in ‘`uid`’ is a dynamically assigned id by a RDF parser.

**Table 29. An anonymous resource entry**

| id      | uid                              | date discover | type res |
|---------|----------------------------------|---------------|----------|
| 1930096 | swd:186#93df2c:109a35cc253:-7fee | 2006-02-25    | 186      |

## 5 Conclusion

In this paper, we describe Swoogle's database design for representing and storing the metadata of the Semantic Web. We further use an example Semantic Web document to explain how the metadata is extracted and stored.

## References

[Ding et al. 2005] Li Ding, Rong Pan, Tim Finin, Anupam Joshi, Yun Peng and Pranam Kolari, *Finding and Ranking Knowledge on the Semantic Web*, in proceedings of ISWC 2005, November, 2005.

[Ding et al. 2004] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal C Doshi and Joel Sachs, *Swoogle: A Search and Metadata Engine for the Semantic Web*, in proceedings of the 13th ACM Conference on Information and Knowledge Management, 2004.

[Bos2006] Bert Bos, *Character encodings*, <http://www.w3.org/International/O-charset.html>, (version 2006-01-09 18:43 GMT).

[Berners-Lee et al. 1998] T. Berners-Lee, R. Fielding, U.C. Irvine, L. Masinter, *Uniform Resource Identifiers (URI): Generic Syntax*, <http://www.faqs.org/rfcs/rfc2396.html> (version August 1998).

[Klyne et al. 2004] Graham Klyne, Jeremy J. Carroll, *Resource Description Framework (RDF): Concepts and Abstract Syntax*, <http://www.w3.org/TR/rdf-concepts/#section-blank-nodes> (version 10-February-2004).

[Bray et al. 1999] Tim Bray, Dave Hollander, Andrew Layman, *Namespaces in XML*, <http://www.w3.org/TR/1999/REC-xml-names-19990114> (version 14-January-1999).

## Appendix A: the example ontology in RDF/XML grammar

```
<?xml version="1.0" encoding="UTF-8" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#">

  <owl:Ontology rdf:about="http://rdflib.net/2002/InformationStore">
    <rdfs:label>InformationStore</rdfs:label>
    <rdfs:comment>A document defining a vocabulary used by an InformationStore, a TripleStore with support for
      multiple contexts.</rdfs:comment>
    <rdfs:isDefinedBy rdf:resource="http://rdflib.net/2002/InformationStore"/>
  </owl:Ontology>

  <rdfs:Class rdf:about="http://rdflib.net/2002/InformationStore#Context">
    <rdfs:label>Context</rdfs:label>
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
    <rdfs:isDefinedBy rdf:resource="http://rdflib.net/2002/InformationStore"/>
  </rdfs:Class>

  <rdfs:Class rdf:about="http://rdflib.net/2002/InformationStore#UpdateEvent">
    <rdfs:label>An Event for updating a Context</rdfs:label>
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
    <rdfs:isDefinedBy rdf:resource="http://rdflib.net/2002/InformationStore"/>
  </rdfs:Class>

  <rdf:Property rdf:about="http://rdflib.net/2002/InformationStore#source">
    <rdfs:label>source</rdfs:label>
    <rdfs:domain rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
    <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Resource"/>
    <rdfs:isDefinedBy rdf:resource="http://rdflib.net/2002/InformationStore"/>
  </rdf:Property>

  <rdf:Property rdf:about="http://rdflib.net/2002/InformationStore#updateEvent">
    <rdfs:label>update event</rdfs:label>
    <rdfs:domain rdf:resource="http://www.w3.org/2000/01/rdf-schema#Context"/>
    <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#UpdateEvent"/>
    <rdfs:isDefinedBy rdf:resource="http://rdflib.net/2002/InformationStore"/>
  </rdf:Property>

  <rdf:Property rdf:about="http://rdflib.net/2002/InformationStore#error">
    <rdfs:label>Error</rdfs:label>
    <rdfs:comment>This property is used to capture information about errors that occurred while updating the
      context.</rdfs:comment>
    <rdfs:domain rdf:resource="http://www.w3.org/2000/01/rdf-schema#UpdateEvent"/>
    <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
    <rdfs:isDefinedBy rdf:resource="http://rdflib.net/2002/InformationStore"/>
  </rdf:Property>

  <rdf:Property rdf:about="http://rdflib.net/2002/InformationStore#http_status">
    <rdfs:label>http status</rdfs:label>
    <rdfs:comment>This property is used for the http status that was returned when updating the
      context.</rdfs:comment>
    <rdfs:domain rdf:resource="http://www.w3.org/2000/01/rdf-schema#UpdateEvent"/>
    <rdfs:range rdf:resource="http://www.w3.org/2000/01/rdf-schema#Literal"/>
    <rdfs:isDefinedBy rdf:resource="http://rdflib.net/2002/InformationStore"/>
  </rdf:Property>

</rdf:RDF>
```

## Appendix B: the example ontology in N-Triples grammar

| No. | Subject   | Predicate  | Object  |
|-----|---|--|---|
| 1   | http://rdflib.net/2002/InformationStore             | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/2002/07/owl#Ontology  |
| 2   | http://rdflib.net/2002/InformationStore             | http://www.w3.org/2000/01/rdf-schema#label       | "InformationStore"  |
| 3   | http://rdflib.net/2002/InformationStore             | http://www.w3.org/2000/01/rdf-schema#comment     | "A document defining a vocabulary used by an InformationStore, a TripleStore with support for multiple contexts." |
| 4   | http://rdflib.net/2002/InformationStore             | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore   |
| 5   | http://rdflib.net/2002/InformationStore#Context     | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/2000/01/rdf-schema#Class  |
| 6   | http://rdflib.net/2002/InformationStore#Context     | http://www.w3.org/2000/01/rdf-schema#label       | "Context"   |
| 7   | http://rdflib.net/2002/InformationStore#Context     | http://www.w3.org/2000/01/rdf-schema#subClassOf  | http://www.w3.org/2000/01/rdf-schema#Resource   |
| 8   | http://rdflib.net/2002/InformationStore#Context     | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore   |
| 9   | http://rdflib.net/2002/InformationStore#UpdateEvent | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/2000/01/rdf-schema#Class  |
| 10  | http://rdflib.net/2002/InformationStore#UpdateEvent | http://www.w3.org/2000/01/rdf-schema#label       | "An Event for updating a Context"   |
| 11  | http://rdflib.net/2002/InformationStore#UpdateEvent | http://www.w3.org/2000/01/rdf-schema#subClassOf  | http://www.w3.org/2000/01/rdf-schema#Resource   |
| 12  | http://rdflib.net/2002/InformationStore#UpdateEvent | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore   |
| 13  | http://rdflib.net/2002/InformationStore#source      | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/1999/02/22-rdf-syntax-ns#Property   |
| 14  | http://rdflib.net/2002/InformationStore#source      | http://www.w3.org/2000/01/rdf-schema#label       | "source"  |
| 15  | http://rdflib.net/2002/InformationStore#source      | http://www.w3.org/2000/01/rdf-schema#domain      | http://www.w3.org/2000/01/rdf-schema#Resource   |
| 16  | http://rdflib.net/2002/InformationStore#source      | http://www.w3.org/2000/01/rdf-schema#range       | http://www.w3.org/2000/01/rdf-schema#Resource   |
| 17  | http://rdflib.net/2002/InformationStore#source      | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore   |
| 18  | http://rdflib.net/2002/InformationStore#updateEvent | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/1999/02/22-rdf-syntax-ns#Property   |
| 19  | http://rdflib.net/2002/InformationStore#updateEvent | http://www.w3.org/2000/01/rdf-schema#label       | "update event"  |
| 20  | http://rdflib.net/2002/InformationStore#updateEvent | http://www.w3.org/2000/01/rdf-schema#domain      | http://www.w3.org/2000/01/rdf-schema#Context  |
| 21  | http://rdflib.net/2002/InformationStore#updateEvent | http://www.w3.org/2000/01/rdf-schema#range       | http://www.w3.org/2000/01/rdf-schema#UpdateEvent  |
| 22  | http://rdflib.net/2002/InformationStore#updateEvent | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore   |
| 23  | http://rdflib.net/2002/InformationStore#error       | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/1999/02/22-rdf-syntax-ns#Property   |
| 24  | http://rdflib.net/2002/InformationStore#error       | http://www.w3.org/2000/01/rdf-schema#label       | "Error"   |
| 25  | http://rdflib.net/2002/InformationStore#error       | http://www.w3.org/2000/01/rdf-schema#comment     | "This property is used to capture information about errors that occurred while updating the context."             |
| 26  | http://rdflib.net/2002/InformationStore#error       | http://www.w3.org/2000/01/rdf-schema#domain      | http://www.w3.org/2000/01/rdf-schema#UpdateEvent  |
| 27  | http://rdflib.net/2002/InformationStore#error       | http://www.w3.org/2000/01/rdf-schema#range       | http://www.w3.org/2000/01/rdf-schema#Literal  |
| 28  | http://rdflib.net/2002/InformationStore#error       | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore   |
| 29  | http://rdflib.net/2002/InformationStore#http_status | http://www.w3.org/1999/02/22-rdf-syntax-ns#type  | http://www.w3.org/1999/02/22-rdf-syntax-ns#Property   |
| 30  | http://rdflib.net/2002/InformationStore#http_status | http://www.w3.org/2000/01/rdf-schema#label       | "http status"   |
| 31  | http://rdflib.net/2002/InformationStore#http_status | http://www.w3.org/2000/01/rdf-schema#comment     | "This property is used for the http status that was returned when updating the context."                          |
| 32  | http://rdflib.net/2002/InformationStore#http_status | http://www.w3.org/2000/01/rdf-schema#domain      | http://www.w3.org/2000/01/rdf-schema#UpdateEvent  |
| 33  | http://rdflib.net/2002/InformationStore#http_status | http://www.w3.org/2000/01/rdf-schema#range       | http://www.w3.org/2000/01/rdf-schema#Literal  |
| 34  | http://rdflib.net/2002/InformationStore#http_status | http://www.w3.org/2000/01/rdf-schema#isDefinedBy | http://rdflib.net/2002/InformationStore   |

## Appendix C: database schema description

### 1. Table `digest_swd`: metadata of SWD

Table C.1 Column Description for table `digest_swd`

| column name                    | data type                 | column description   |
|--------------------------------|---------------------------|--|
| <code>idurl</code>             | <code>int(11)</code>      | Surrogate key for SWDs.  |
| <code>url</code>               | <code>varchar(250)</code> | URL for each SWD.  |
| <code>date_discover</code>     | <code>date</code>         | The date on which the SWD was discovered by Swoogle.   |
| * <code>url_suffix</code>      | <code>varchar(10)</code>  | The filetype of the SWD. e.g. "rdf", "owl".  |
| <code>url_idhost</code>        | <code>int(11)</code>      | Unique ID of the host (protocol, name, port) of the URL. Foreign key referencing 'id' column of table <code>digest_host</code> .             |
| <code>url_lastmodified</code>  | <code>int(11)</code>      | The last modified time of the SWD. If it is not specified in http response header, we use the date of the latest cached version.             |
| <code>date_ping</code>         | <code>date</code>         | the date when Swoogle last accessed the URL  |
| * <code>flag_ping</code>       | <code>tinyint(1)</code>   | the status of the SWD according Swoogle's last access  |
| <code>date_digest</code>       | <code>date</code>         | INTERNAL-USE   |
| <code>doc_date_cache</code>    | <code>date</code>         | The date when the SWD was last cached by Swoogle   |
| <code>doc_charset</code>       | <code>varchar(20)</code>  | The charset used to encode the SWD, e.g. "utf-8"   |
| <code>doc_length</code>        | <code>int(11)</code>      | The length (in bytes) of the SWD.  |
| <code>doc_md5sum</code>        | <code>varchar(32)</code>  | The md5 hash of the SWD.   |
| <code>doc_duplicate</code>     | <code>tinyint(1)</code>   | INTERNAL-USE.  |
| * <code>rdf_parse_state</code> | <code>tinyint(1)</code>   | The RDF parse result for the SWD.  |
| <code>rdf_syntax</code>        | <code>varchar(2)</code>   | The RDF syntax used by the SWD. Values: 'RE' (RDF/XML, embedded), 'RX' (RDF/XML), 'N3' (Notation 3 – RDF), 'NT' (N-Triples).                 |
| <code>rdf_default_ns</code>    | <code>int(11)</code>      | Unique ID for the default namespace declared by the SWD. This is a foreign key referencing the 'id' column of table <code>digest_ns</code> . |
| <code>rdf_cnt_triple</code>    | <code>int(11)</code>      | The number of triples in the SWD   |
| <code>rdf_cnt_swt</code>       | <code>int(11)</code>      | The number of unique SWTs in the SWD (see glossary)  |
| <code>rdf_cnt_swt_def</code>   | <code>int(11)</code>      | The number of unique SWTs being defined/referenced (not populated) as classes or properties in the SWD                                       |
| <code>rdf_cnt_instance</code>  | <code>int(11)</code>      | The number of unique class-instances in the SWD.   |
| <code>rdf_cnt_def_i</code>     | <code>int(11)</code>      | INTERNAL-USE.  |
| <code>rdf_onto_ratio</code>    | <code>float</code>        | The percentage of triples used for class/property definition.  |
| <code>rank</code>              | <code>double</code>       | Swoogle's rank of the SWD.   |

\* **url\_suffix**: This feature is extracted from URL using several heuristics. Many URLs do not have extensions because they are dynamically generated. '---' means no extension.

\* **flag\_ping**: The status of the SWD according to Swoogle's last access (aka ping). Possible values are:

- ❖ failed
  - 1 (cannot connect due to robots.txt)
  - 2 (URL unreachable, maybe offline)
  - 3 (response forbidding further access)
  - 4 (response indicating URL redirection)
  - 5 (connection failed for unknown reason)
  - 6 (interrupted during download)
  - 20 (download failed because the file size is too large)
- ❖ alive
  - 7 (ALIVE) unchanged

- 8 (MODIFIED from NSWD to SWD)
- 9 (MODIFIED from SWD to NSWD)

\* **rdf\_parse\_state**: result code of RDF parsing. Possible values are given in Table C.2

**Table C.2 Options for column rdf\_parse\_state**

|              | no RDF graph | RDF graph with parse error | RDF graph with parse warning | RDF graph error free |
|--------------|--------------|----------------------------|------------------------------|----------------------|
| not visit    | 0            |                            |                              |                      |
| visited nswd | 2            |                            |                              |                      |
| embedded swd |              | 12                         | 13                           | 15                   |
| Swd          |              | 22                         | 23                           | 25                   |

## 2. Table digest\_swt: metadata of SWT

**Table C.3 Column Description for table digest\_swt**

| column name   | data type    | column description   |
|---------------|--------------|--|
| id            | int(11)      | Surrogate key for semantic web terms. It is also a foreign key referencing 'id' column of table digest_resource.     |
| uri           | varchar(250) | The URI for the semantic web term. The value of this column cannot be blanknode.                                     |
| date_discover | date         | The date on which the term was discovered by Swoogle.  |
| uri_idns      | int(11)      | The unique id of the namespace used by the term. This is foreign key referencing the 'id' column of table digest_ns. |
| uri_localname | varchar(200) | The local name of the term.  |
| has_def       | tinyint(1)   | A flag showing whether this term has been defined or referenced as a class or property.                              |
| rank          | double       | Swoogle's rank for a semantic web term.  |

## 3. Table digest\_ns: metadata of SWN

Currently, we only store SWNs used by SWTs.

**Table C.4 Column Description for table digest\_ns**

| column name         | data type    | column description   |
|---------------------|--------------|--|
| id                  | int(11)      | Surrogate key for namespaces.  |
| uri                 | varchar(250) | Unique URI for each namespace.   |
| date_discover       | date         | The date when the namespace was discovered   |
| uri_unique          | varchar(250) | The revised URI by removing '#' because it is part of the 'fragment', and does not affect web addressing |
| idurl_official_onto | int(11)      | Unique ID of the official ontology SWD of the namespace as derived by Swoogle.                           |
| type_official_onto  | tinyint(1)   | INTERNAL USE   |

#### 4. Table rel\_swd\_swt: relations between SWD and SWT

Besides term frequency, Swoogle additionally stores the frequency of meta-usage of SWT. Swoogle counts frequency as the occurrences of SWT in N-Triples serialization of the SWDs.

**Table C.5 Column Description for table rel\_swd\_swt**

| column name       | data type  | column description  |
|-------------------|------------|---|
| idswd             | int(11)    | The unique id of the SWD. This is a foreign key referencing column 'idurl' of table digest_swd.   |
| idswt             | int(11)    | The unique id of the SWT. Foreign key referencing column 'id' of table digest_swt.  |
| term_frequency    | int(11)    | No. of occurrences of the SWT in the SWD  |
| rdf_cnt_ref_c     | int(11)    | No. of occurrences of the SWT being referenced as class.  |
| rdf_cnt_ref_p     | int(11)    | No. of occurrences of the SWT being referenced as property.   |
| rdf_cnt_pop_c     | int(11)    | No. of occurrences of the SWT being populated as class.   |
| rdf_cnt_pop_p     | int(11)    | No. of occurrences of the SWT being populated as property.  |
| rdf_cnt_def_c     | int(11)    | No. of occurrences of the SWT being defined as class.   |
| rdf_cnt_def_p     | int(11)    | No. of occurrences of the SWT being defined as property.  |
| has_literal_range | tinyint(3) | Only applies to properties, indicating whether there exist at least one triple using the term as predicate and having a literal object. |

#### 5. Table rel\_swd\_swt\_domain\_range: instantiation of domain/range

Swoogle extracts the instantiation/usage of domain/range definition from instance data. Given an SWD, Swoogle extracts all possible (property, range, domain) combinations from each triple except for the triples in form of (?x, rdf:type, ?y).

**Table C.6 Column Description for table rel\_swd\_swt\_domain\_range**

| column name  | Data type | column description   |
|--------------|-----------|--|
| idswd        | int(11)   | The unique id of the SWD. This is a foreign key referencing column 'idurl' of table digest_swd.          |
| idswt        | int(11)   | The unique id of the SWT. This is a foreign key referencing column 'id' of table digest_swt.             |
| idswt_domain | int(11)   | A type of the subject element in a triple with the SWT as predicate.                                     |
| idswt_range  | int(11)   | A type of the object element in a triple with the SWT as predicate.                                      |
| cnt          | int(11)   | The number of occurrences of the domain and range types being associated with this predicate in the SWD. |



## 6. Table rel\_swd\_ns\_prefix: relations between SWD and SWN

Table C.7 Column Description for table rel\_swd\_ns\_prefix

| column name  | data type   | column description  |
|--------------|-------------|---|
| idswd        | int(11)     | The unique id of the SWD. Foreign key referencing column 'idurl' of table digest_swd.   |
| id_ns        | int(11)     | The unique id of the namespace. Foreign key referencing column 'id' of table digest_ns.   |
| prefix       | varchar(40) | The prefix used as QName of the namespace in the SWD.   |
| *use_cnt     | int(11)     | No. of meta-usage of the namespace in the SWD   |
| declare_type | tinyint(1)  | The attribute is about how the namespace is declared in the SWD. Options are:<br>0: not declared<br>1: declared<br>2: declared as default namespace |

\* use\_cnt: computed by sum up the term frequency of all SWTs using the SWN.

## 7. Table digest\_resource: RDF resources used in SWT definition

Table C.8 Column Description for table digest\_resource

| column name   | data type    | column description  |
|---------------|--------------|---|
| id            | int(11)      | Surrogate key for resources.  |
| uid           | varchar(250) | Unique original resource id retrieved from a SWD, usually in the form of hostname followed by localname.  |
| date_discover | Date         | The date on which the resource was discovered by Swoogle.   |
| type_res      | int(11)      | The type of the resource. The options include:<br>>0: id of SWD containing this anonymous Resource<br>-1: bad URI<br>-2: basic correct URI<br>-3: URI in the form of http representation. |

## 8. Triple tables

Swoogle specially records some triples from SWDs because they are used to define the SWD or SWTs in the SWD. All these tables share a common structure and have their own extensions. The general structure of triple tables is listed in table C.9. Their distinct extensions are listed in table C.10 to C.13 respectively.

Table C.9 Common columns used in four triple tables

| column name     | data type               | column description   |
|-----------------|-------------------------|--|
| id              | int(11)                 | Surrogate key for stored RDF triples.  |
| idswd_source    | int(11)                 | The unique id of the SWD containing this triple. Foreign key referencing column 'idurl' of table digest_swd. |
| idswt_predicate | int(11)                 | The unique id of the predicate of this triple. Foreign key referencing column 'id' of table digest_swt.      |
| subject         | varchar(250)            | The URI of the subject.  |
| predicate       | varchar(250)            | The URI of the predicate. (redundant to avoid table join)  |
| * object        | varchar(250)<br>OR text | The URI of the object.<br>The text content of the object.  |

\* object: We store the triples with literal object and with resource object in different tables for (i) storage efficiency (ii) link (in the case that the object resource is itself a SWD, we also record its unique SWD id in another column).

**Table C.10 Additional columns used in table triple\_swd\_swd**

| column name  | data type    | column description   |
|--------------|--------------|--|
| object       | varchar(250) | The URI of the object.   |
| idswd_object | int(11)      | The unique id of the SWD if the object element is a SWD; if not, '0' displays instead. |

**Table C.11 Additional columns used in table triple\_swd\_annotation**

| column name | data type | column description              |
|-------------|-----------|---------------------------------|
| object      | Text      | The text content of the object. |

**Table C.12 Additional columns used in table triple\_swt\_swt**

| column name   | data type    | column description   |
|---------------|--------------|--|
| object        | varchar(250) | The URI of the object.   |
| idswt_root    | int(11)      | For a class defined with construction of anonymous classes, all sustaining triples containing anonymous classes have the term id of the defined class as its idswt_root value. This field enables us to trace along anonymous classes to make clear how every class is defined, even some anonymous classes are used for its creation. |
| idres_subject | int(11)      | The resource id for the subject of the target triple. This is a foreign key referencing column 'id' of table digest_resource.  |
| idres_object  | int(11)      | The resource id for the object of the target triple. This is a foreign key referencing column 'id' of table digest_resource.   |

**Table C.13 Additional Columns used in table triple\_swt\_annotation**

| column name | data type | column description                            |
|-------------|-----------|---|
| object      | text      | Text content of the object.                   |
| idswt_root  | int(11)   | Same as the corresponding filed in Table C.12 |