PII Resource Descriptions and the DCMI Abstract Model (DCAM)

Supplement to Advanced Methods

Foreword

Personally Identifiable Information, redacted from a document with RUST has had its value deleted but nonetheless retains its class – as in *classification.* If the only reason to fool with normally hidden meta data is to improve differentiation between documents, then it follows that *nullified but once visible*, non-propagating, data is still useful.

Redact Unless Static Text (RUST) is a rule for meta data embedded in Plain Text based on a tokenized the meta data "Value". This "Value" is part of the Property-Value Pair which forms the basis of the DCMI Abstract Model.

Each Property has a Range and a Domain.

- All properties in the PII name space are of DCMIType:Text, meaning that the "stuff redacted" was plain text and not images, etc.. This is the beginnings of a PII ontology. As methods of bounding meta data in other genre are developed, this can be extended to other DCMIType's. In DCMI parlance, the "Range" is simple text or equivalent (that which a human can read as text). The idea of bounding the range of PII by genre is crucial to the evidentiary uses of personal identity – in a perfect world, no one looks as bad as their driver's license picture.

- All properties in the PII name space are exposed (described) in the RDF of a document resource (instance document, eg a GRDDL transform) as members of the class BibliographicResource. In DCMI parlance, the "Domain" is <dcterms:BibliographicResource>

The described meta data element would be identical to <dcterms:bibliographicCitation> but has the advantage of extensibility. While Optical Character Recognition (OCR) was foreseen, and so a facsimile included in genre "Text", biometric data, for example an Iris Scan or a gene sequence, is PII but not a human readable citation or reference by any abstraction. Likewise, a movie of a Field Sobriety Test is PII, but not human "readable". For this reason we hesitate limiting ourselves to a <Literal> "Value" as called for by <dcterms:bibliographicCitation>. Nonetheless, the "Domain" is shared because the "Value" token occurs in printed documents.

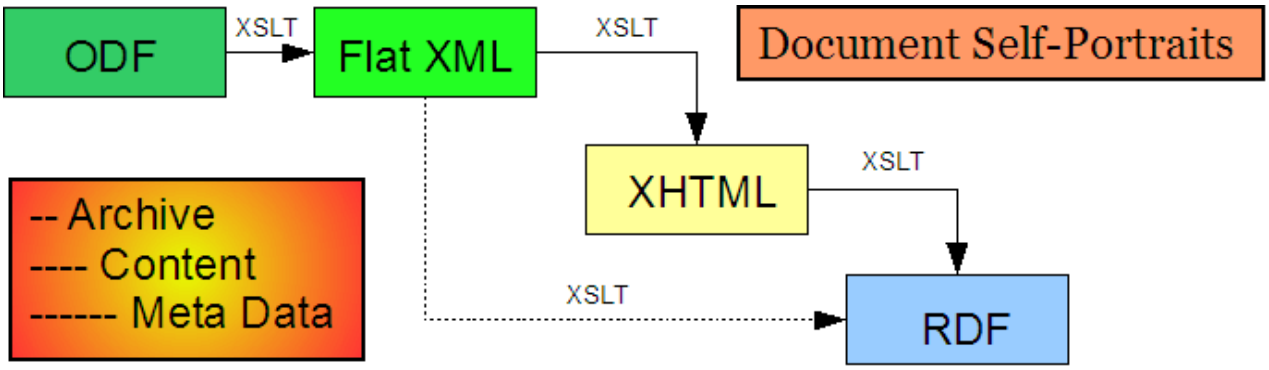
The token may be rendered as needed or displayed as an icon – with one critical security feature: There is only one icon source, a PURL, which must be used to insure that a picture link points to the PII name space resource. With PURLs, it is possible to restrict redirection, and to validate that no redirection has taken place by simple examination of the link source code. While this option may be added to rel="help" links in the future, no benefit to understanding the meta data is seen for rel="term", those links which become an RDF definition set with a GRDDL transform.

Why does it have to be so complicated?

"Well I've never been to Spain, but I kind of like the music ..."

– *3 Dog Night*

This is a micro-format tag link. It can be visible meta data, depending upon the link, but 1) it does not RUST, and 2) does not have sufficient capacity to describe the resource it references *at a distance*. Imagine trying to sort out the references in the quote above with its layers of misdirection! Fuzzy references exist all over the place and have their uses, but when it comes to re-connectible references (meta data) we have to draw the line somewhere. The most straightforward way to do this is to ignore links that do not state explicitly that their target is a resource framework, and exclude them from the RDF Definition Set.

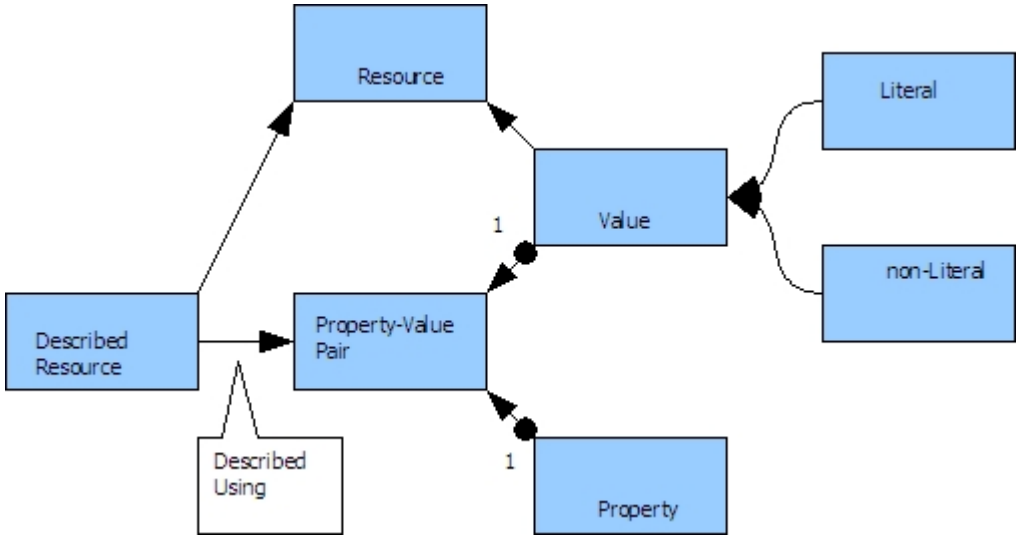


If what we are after is a Document self-portrait then it is unnecessary to duplicate this type of link in the RDF or "meta data" section. There is an ulterior motive at work here: almost since their inception, links have been used for misdirection.

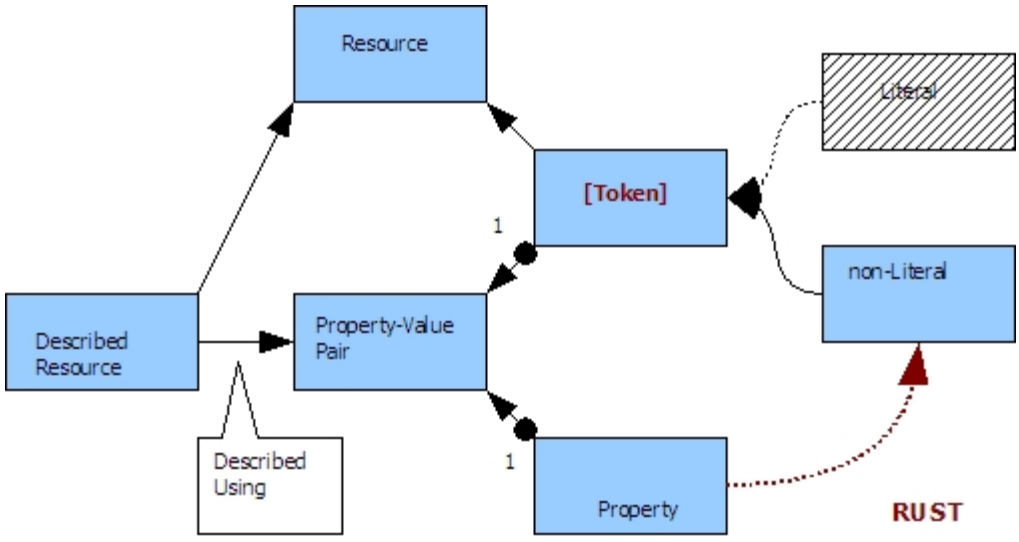
The DCMI Abstract Models

The DCMI has published several articles about their Abstract Model.

The DCMI Resource Model



Redaction with RUST applies this model, although in a purely text context.



Nonetheless, the model facets remain in place. Generalized redaction (deletion), where the token is blank space or the empty set is an abstraction we can do without – these Tokens appear as packed, everywhere. A Classified Document, containing secrets has the same model, with token replacement by degree. The process of Declassification can be seen as a further replacement of tokens by their original content. For simplicity, the PII name space tokens are all DCMIType::Text, but a generalized method of Declassification would include the other genre property values. This genre property refers to the information tokenized and not

to the rendering of the token itself. PII being a very serious business; WWW Silliness; Icons, Animated Icons, Sound Effects etc. have no place and lower trust in the integrity of resource location and identification.

The DCMI Description Set Model and The DCMI Vocabulary Model are not affected by RUST behavior in a text environment. We should mention though that RUST is a Vocabulary Encoding Scheme, and whether intentionally or not, the vexing problem of *cultural ciphers* found in Anchor●Buoy●Boat Diagrams is well represented in the Description Set Model, if we only identify "Buoy" with "Property". This exercise is left to the reader.

Resource Definition Framework (RDF)

When PII is "exposed" with RUST in an XHTML export only one anchor/link type (publisher='RDF') is passed on to the GRDDL transform. One should remember that the *Personally Identifiable Information* has already been redacted, that is tokenized and deleted. The token represents meta data residue, not the original literal text.

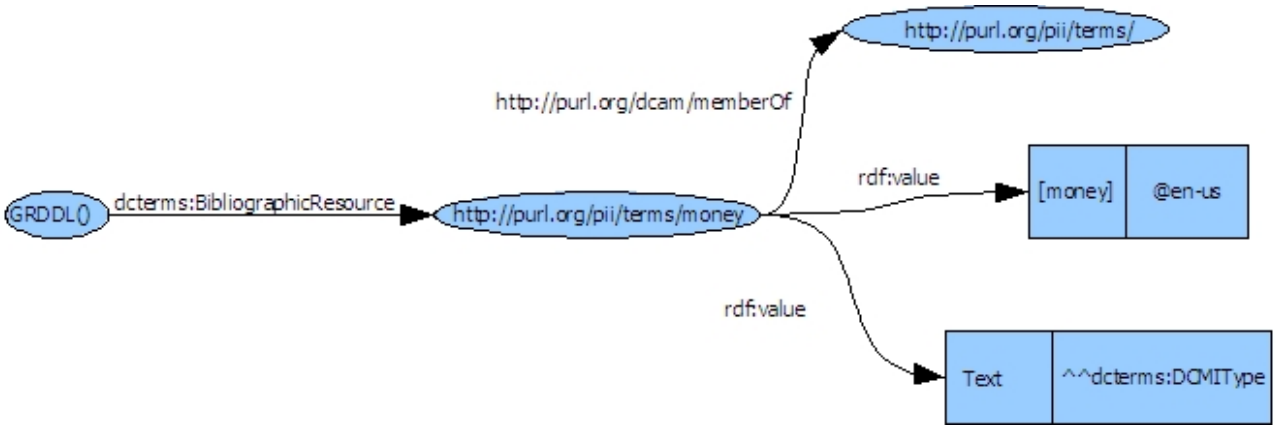
In the XHTML page it's a link : [money]

It looks like this after the transform:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:pii="http://purl.org/pii/terms/"
 xmlns:dcterms="http://purl.org/dc/terms/"
 xmlns:dcam="http://purl.org/dc/dcam/">
<rdf:Description rdf:about="GRDDL()">
  <dcterms:BibliographicResource>
    <rdf:Description rdf:about="http://purl.org/pii/terms/money">
      <dcam:memberOf rdf:resource="http://purl.org/pii/terms/" />
      <rdf:value xml:lang="en-US">[money]</rdf:value>
      <rdf:value rdf:datatype="http://purl.org/dc/terms/DCMIType">
         Text
      </rdf:value>
    </rdf:Description>
  </dcterms:BibliographicResource>
</rdf:Description>
</rdf:RDF>
```

RDF Graph of GRDDL Transform

The **G**leaning **R**esource **D**escriptions from **D**ialects of **L**anguages (GRDDL) Transform is an XSLT transform run on an XML file which scrapes Resource Descriptions. When a token from the PII name space is encountered, the Resource Description gleaned above results in an RDF Graph.



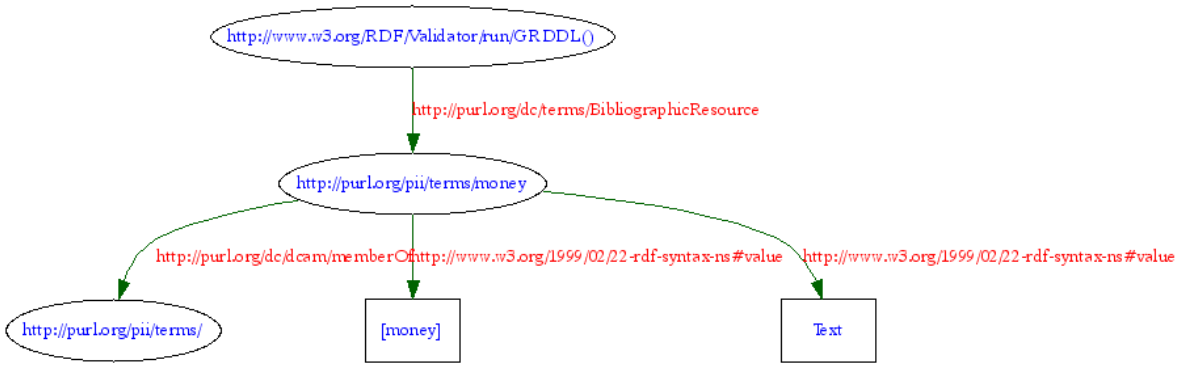
That is the theory, anyway. Similar graphs can be drawn for each of the tokens.

The features we wish to see are:

- The Value URI of the resource is a token, in this case [money].
- The Value URI represents "Text" (as dcterms:DCMIType)
- The Value URI is a member of PII Terms.
- This resource is a member of the class dcterms:BibliographicResource.

Validation

The W3C RDF-Validator agrees.



RDF Triples

Subject	Predicate	Object
GRDDL()	http://purl.org/dc/terms/BibliographicResource	http://purl.org/pii/terms/money
http://purl.org/pii/terms/money	http://purl.org/dc/dcam/memberOf	http://purl.org/pii/terms/
http://purl.org/pii/terms/money	http://www.w3.org/1999/02/22-rdf-syntax-ns#value	"[money]"@en-US
http://purl.org/pii/terms/money	http://www.w3.org/1999/02/22-rdf-syntax-ns#value	"Text"^^http://purl.org/dc/terms/DCMIType

What about other DCMI Vocabulary Encoding Schemes ?

The other Vocabulary Encoding Schemes should write the appropriate link in XHTML (e.g. Medical Subject Headings - MeSH). We already know what we want the other schemes to look like in RDF:

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:pii="http://purl.org/pii/terms/"
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:dcam="http://purl.org/dc/dcam/">
  <rdf:Description rdf:about="GRDDL()">
    <dcterms:BibliographicResource>
       <rdf:Description rdf:about="http://purl.org/pii/terms/MESH">
         <dcam:memberOf rdf:resource="http://purl.org/pii/terms/" />
         <rdf:value xml:lang="en-US">funny bone</rdf:value>
       </rdf:Description>
    </dcterms:BibliographicResource>
  </rdf:Description>
</rdf:RDF>
```

In XHTML the link should look like this: "funny bone"[MeSH-001]

or this, if we borrow the MeSH search notation:

```
<a
     rel="term"
     title="MeSH/Anatomy/Human/Bones/Arm/"
     type="application/rdf+xml"
     charset="UTF-8"
     xml:lang="en"
     href="http://purl.org/dc/terms/MESH">"funny bone"[MeSH-001]</a>
```

How about validation ? Well, you can't validate this a priori, that's what resources are all about, and where Author as in Authority comes into play. The Dublin Core is saying that they know where Medical Subject Headings are kept and who keeps them.

This bothers people who think that meta data is a logical assertion. Hollywood used to use live phone numbers in TV and film, causing the poor person who had the phone number to be bombarded with calls from viewers wishing to comment on plot twists and characters. Feel free to keep thinking meta data is a logical assertion and someday you will be famous – Hollywood will use your phone number. You deserve that.

There are several vocabularies you can use. They include DCMIType, DDC, IMT, LCC, LCSH, MESH, NLM, TGN, and UDC. The acronym is the dcterms name. The easiest way to insert a reference is to make a new Bibliography Entry (from the document). You will need to fill in a Short Name, Author, Title and Publisher.

- Short Name … A unique identifier like [MeSH-001]. Unlike redacted information, these references have a unique payload.
- Title … This is the complete Property URI for the resource.
- Publisher … 'RDF' (fixed)
- Author … This will be the link's Title, normally implemented in browsers as a tool-tip. This can be a generic title, but best practice should be to use an XPATH if one exists which will aid in the understanding of the link in XHTML. One should take care that the 'Author' field is recognizably associated with the 'Title' the Hyperlink reference (href) – the tool-tip should not mislead – but in any case, the 'Author' field value has no effect on the RDF representation.

Using an XPATH in a link, while it enables validation, does not insure absolute accuracy as in a logical assertion. Just as the tool-tip should not hide the destination for the link, so also is it unreliable as proof of the destination.