How to get and create RDF Data?
(Microformats, GRDDL, RDFa, POWDER)
Simple approach

- Write RDF/XML, RDFa, or Turtle “manually”
- In some cases that is necessary, but it really does not scale…
Obviously, a huge source of information

By adding some “meta” information, the same source can be reused for, eg, data integration, better mashups, etc

- typical example: your personal information, like address, should be readable for humans and processable by machines

Two solutions have emerged:

- use microformats and convert the content into RDF
- add RDF statements directly into XHTML via RDFa
Microformats

- Not a Semantic Web specification, originally
  - there is a separate microformat community
- Approach: re-use (X)HTML attributes and elements to add “meta” information
  - typically @abbr, @class, @title, …
  - different agreements for different applications
Microformat example: hCalendar

- Goal: “markup” calendaring information on your HTML page
  - use a community agreement using, eg, :
    - @class for event name
    - abbr element for dates
    - @title for date values
    - etc.
- Automatic procedures (ie, calendaring applications) may then get to the right data
Microformat example: hCalendar

Dan Connolly
Research Scientist, MIT/CSAIL

(connolly@w3.org)
32 Vassar Street
Room 32-G506
Cambridge, MA, 02139 USA
+1-617-395-0241, DanC

hCard

standards: HTML WG, TAG, GRDDL WG, RDF Calendar, QA, DAWG/SPARQL, Semantic Web IG, OWL, HTML 2, ESW

research: breadcrumbs journal/weblog, cwm, N3, tabulator, PAW, TAMI, microformats open source
life: family, volleyball, guitar

Dan Connolly is a research scientist at the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) in the Decentralized Information Group (DIG) and a member of the technical staff of the World Wide Web Consortium (W3C). His research interest is

Mar 7-11, 2008: to Austin, TX for SXSW Interactive
Apr 20-22: in Beijing, China for W3C AC meeting, linked data workshop, trip stuff
May 19-May 22: to Bristol for TAG itf, trip stuff
Sep 23 - 25: in Kansas City TAG meeting
Oct 20 - 25: to NCE for W3C TPAC
Dec 08 - 11: to Cambridge, MA TAG meeting
Feb 2 to 7: to Denver hoping to go for Web Directions North

Earlier travel/talks/events include Tools of
Behind the scenes...
To use it on the Semantic Web, microformat data should be converted to RDF.

A simple transformation (e.g., in XSLT) can be defined, yielding:

```xml
<http://www.w3.org/People/Connolly/#sxsw2008>
  a hcal:Vevent;
  hcal:organizer <http://www.w3.org/People/Connolly/#me>;
  hcal:summary "SXSW Interactive";
  hcal:dtstart "2008-03-07"^^xsd:date;
  hcal:dtend "2008-03-12"^^xsd:date;
  hcal:location "Austin, TX" .
```
So far so good, but…

- The XSLT transformation is hCalendar specific
  - each microformat dialect needs its own
- How does a general processor find the right transformation?
- Enter GRDDL
GRDDL defines
- a few extra attribute values to locate the right transformation
- a precise processing model on how the transformation should be applied to generate RDF

Note: we describe GRDDL in terms of XHTML (and microformats) but GRDDL can be used for any XML data
GRDDL: find the right transformation
The GRDDL process: simple case

XHTML/XML

@transformation

XSLT Script (dc-extract.xsl)

W3C RDF
The GRDDL process: merging case
The GRDDL process: indirect case

@namespacetransformation
@profiletransformation

Namespace/profile XHTML/XML

XSLT Script (dc-extract.xsl)

XHTML/XML

W3C RDF
Microformats & GRDDL: pros

- Pros:
  - simple to define/add new vocabularies
    - there is a strong microformat community for this
  - works with all current browsers, markup validators, etc
  - fairly user friendly, easy to understand and use
Microformats & GRDDL: cons

Cons:
- does not scale well for complex vocabularies
  - remember: needs a transformation per vocabulary
- difficult to mix vocabularies within one page
  - what if the usage of an attribute clashes among different vocabularies?
- some of the attributes are meant for other usage
  - eg, the abbr element, the @title attribute, …
An alternative solution: XHTML+RDFa

- RDFa also uses (X)HTML attributes to add “meta” information

- However
  - it also uses proprietary attributes to avoid clashes with the intended usage of the (X)HTML ones
  - it includes a namespace+URI mechanism for disambiguation
  - it is one set of attributes for any vocabularies
Ivan Herman

Who am I?

I graduated as mathematician from the Eötvös Loránd University of Budapest, Hungary, in 1979. After a brief scholarship at the Université Paris VI I joined the Hungarian research institute in computer science (SZTAKI) where I worked for 6 years (and turned into a computer scientist...). I left Hungary in 1986 and, after a few years in industry in Munich, Germany, I joined the Centre for Mathematics and Computer Sciences (CWI) in Amsterdam where I have a tenure position since 1988. I received a PhD degree in Computer Science in 1990 at the University of Leiden, in the Netherlands. I joined the World Wide Web Consortium (W3C) Team as Head of W3C Offices in January 2001 while maintaining my position at CWI. I served as Head of Offices until June 2006, when I was asked to take the Semantic Web Activity Lead position, which is now my principal work at W3C.

Before joining W3C I worked in quite different areas (distributed and dataflow programming, language design, system programming), but I spend most of my research years in computer graphics and information visualization. I also participated in various graphics related ISO standardization activities and software developments. My “professional” home page contains a list of my publications, my public presentations, and details of the various projects I participated in the past. There is also a dblp entry for my publications generated automatically (although I am not sure it is complete...). (B.t.w., based on my publications, my Erdős number is ≤4...)

In my previous life (i.e., before joining W3C...) I was member of the Executive Committee of the Eurographics Association for 15 years, and I was vice-chair of the Association between 2000 and 2002. I was the co-chair of the 9th World Wide Web Conference, in Amsterdam, May 2000; since then, I have also been member of IW3C2 (International World Wide Web Conference Committee) in many capacities (editor, chair, etc).
Same example behind the scenes...

```html
<!--

<span rel="owl:sameAs" resource="http://community.linkeddata.org/dataspce/person/ivan#this">
  <a href="http://community.linkeddata.org/">Linked Open Data community</a></span>
</div>

</div>

<div id="content">
  <h1><span property="foaf:title" class="forRDFOnly">Dr</span><span property="foaf:name">Ivan Herman</span></h1>
  <h2>Who am I?</h2>
  <p>I graduated as mathematician at the <a rel="foaf:schoolHomepage" href="http://www.elte.hu/">Eötvös Loránd University of Budapest</a>, Hungary, in 1979. After a brief scholarship at the Université Paris VI I joined the Hungarian research institute in computer science (<a href="http://www.sztaki.hu">SZTAKI</a>) where I worked for 6 years (and turned into a computer scientist...). I left Hungary in 1986 and, after a few years in industry in Munich, Germany, I joined the <a rel="foaf:workplaceHomepage" href="http://www.cwi.nl/">Centre for Mathematics and Computer Sciences</a> in Amsterdam where I have a tenure position since 1988. I received a PhD degree in Computer Science in 1990 at the <a href="http://www.leidenuniv.nl/">University of Leiden</a> in the Netherlands. I joined the <a rel="foaf:workplaceHomepage" href="http://www.w3.org/">World Wide Web Consortium (W3C)</a> Team as Head of <a rel="foaf:pastProject" href="http://www.w3.org/Consortium/Offices">W3C Offices</a> in January 2001 while maintaining my position at <a abbr title="Centrum Wiskunde en Informatica" xml:lang="nl">CWI</a>. I served as Head of Offices until June 2006, when I was asked to take the <a rel="foaf:workInfoHomepage">W3C Offices</a>.
```
Same example behind the scenes...
In a slightly more readable format...

I graduated as mathematician at the Eötvös Loránd University of Budapest.
In a slightly more readable format...

I graduated as mathematician at the [Eötvös Loránd University of Budapest](http://www.elte.hu/).
... yielding

@prefix foaf: <http://xmlns.com/foaf/0.1/>.
@prefix dc:   <http://purl.org/dc/terms/>.

<http://www.ivan-herman.net/me>
  foaf:schoolHomepage <http://www.elte.hu/>.

<http://www.elte.hu/>
  dc:title "Eötvös Loránd University of Budapest".
Microformats or RDFa?

- There has been many unnecessary controversies
- For simple applications microformats are enough
  - GRDDL bridges them to the rest of the Semantic Web
- For more complex documents RDFa is great
- It often boils down to matters of taste…
Data on the Web are mostly stored in databases

“Bridges” are being defined:

- a layer between RDF and the relational data
  - RDB tables are “mapped” to RDF graphs, possibly on the fly
  - different mapping approaches are being used
- a number RDB systems offer this facility already (e.g., Oracle, OpenLink, …)

R2RML is W3C’s evolving standard in this area
How to “assign” RDF data to resources?

- This is important when the RDF data is used as “metadata”

- Some examples:
  - copyright information for your photographs
  - is a Web page usable on a mobile phone and how?
  - bibliographical data for a publication
  - annotation of the data resulting from a scientific experiment
  - etc
If I know the URI of the resource (photograph, publication, etc), how do I find the relevant RDF data?
The data might be embedded

- Some data formats allow the direct inclusion of (RDF) metadata:
  - SVG (Scalable Vector Graphics)
    - direct inclusion of RDF/XML
    - via RDFa attributes
  - XHTML with RDFa or microformats+GRDDL
  - JPG files using the comment area and, eg, Adobe’s XMP technology
Simple linkage

- Some formats have special link statements. Eg, in (X)HTML:

```html
<html>
  <head>
    <link rel="meta" href="meta.rdf"/>
  ...
</html>
```

- Similar facilities might exist in other formats (eg, SMIL)
POWDER

POWDER provides for a more elaborate scenarios:

- define a set of resources by constraints on the URIs; eg
  - URIs must begin with http://www.example.com/bla/
  - the port number in the URI-s should be XYZW
- define “description resources” that bind those resources to additional information
- get such description resources, eg, via a link statements, via HTTP, via SPARQL, …

Use cases: licensing information, mobileOK (and other) trustmarks, finding accessible Web sites, content labeling (eg, child protection), …
A POWDER scenario: copyright for photos

1. GET .... index

2. Return .... descr.xml

3. GET .... descr.xml

4. GET .... http://ex3.org/img/imgXXX.jpg

5. Deduce triplets

<http://www.ex3.org/img/imgXXX.jpg> cc:license <http://cp...>
The “description resource” is an XML file:

```xml
<powder xmlns="http://www.w3.org/2007/05/powder#"
         xmlns:cc="http://creativecommons.org/ns#">
  <attribution>
    <issuedby src="http://www.ivan-herman.net/me"/>
  </attribution>
  <dr>
    <iriset>
      <includehosts>www.ex2.org</includehosts>
      <includepathstartswith>/img/</includepathstartswith>
    </iriset>
    <descriptorset>
      <cc:license rdf:resource="http://..."/>
    </descriptorset>
  </dr>
</powder>
```
Powder processors may then return
  - direct RDF triples, eg:


- or can transform this XML file into an RDF (OWL) for more
generic processors
  - a canonical processing of the XML file is defined by the
    POWDER specification
Online POWDER service can be set up:

- a Web service with
  - submit a URI and a resource description file
  - return the RDF statements for that URI
- such service should be set up, eg, at W3C

A GRDDL transform is also defined
But there is a hidden “hiccup”

- RDF makes a strong separation between
  - URI as an ID for a resource
  - URI as a datatype (xsd:anyURI)
  - there is no “bridge” between the two

- POWDER includes a small extension to the formal semantics of RDF for two properties:
  - wdrs:matchregex and wdrs:notmatchregex such that
    - (R wdrs:matchregex Regex) holds iff the URI of R matches Regex
If you want the OWL version...

```owl
<> wdrs:issuedBy <http://www.ivan-herman.net/me> .

_:iriset_1 a owl:Class; owl:intersectionOf (  
  [ a owl:Restriction;  
    owl:onProperty wsdr:matchregex ;  
    owl:hasValue "..ugly regex for ex2.org"^^xsd:string ]  
  [ a owl:Restriction;  
    owl:onProperty wsdr:matchregex ;  
    owl:hasValue "..ugly regex for /img"^^xsd:string ]  
).

_:desc_1 a owl:Restriction;  
  owl:onProperty cc:license;  
  owl:hasValue <http://...>.  

_:iriset_1 rdfs:subClassOf _:desc_1 .
```
Consequences of the “hiccup”

- In practice this means that only “POWDER aware” agents can fully handle the description files
  - note that the extension is fairly easy to add, so it is not a big implementation issue…

- Existence of the services to provide the triplets automatically relieve the pain…
Other POWDER features

- There are a number of additional features:
  - built in authentication: description resources must be attributed and this is open for authentication
  - assignments may carry validity dates
  - packaging several resource descriptions in one, possibly control their processing order
  - using tags to identify resources instead of URI patterns
Linking Open Data
Linking Open Data Project

- Goal: “expose” open datasets in RDF
- Set RDF links among the data items from different datasets
- Set up, if possible, query endpoints
Example data source: DBpedia

- DBpedia is a community effort to
  - extract structured (“infobox”) information from Wikipedia
  - provide a query endpoint to the dataset
  - interlink the DBpedia dataset with other datasets on the Web
@prefix dbpedia <http://dbpedia.org/resource/>.
@prefix dbterm  <http://dbpedia.org/property/>.

dbpedia:Amsterdam
  dbterm:officialName "Amsterdam" ;
  dbterm:longd "4" ;
  dbterm:longm "53" ;
  dbterm:longs "32" ;
  dbterm:website <http://www.amsterdam.nl> ;
  dbterm:populationUrban "1364422" ;
  dbterm:areaTotalKm "219" ;
...

dbpedia:ABN_AMRO
  dbterm:location dbpedia:Amsterdam ;
...
Automatic links among open datasets

Processors can switch automatically from one to the other...
The LOD “cloud”, March 2008
The LOD “cloud”, September 2008

As of September 2008
The LOD “cloud”, March 2009
The LOD “cloud”, June 2009

As of July 2009
The LOD “cloud”, September 2010
Application specific portions of the cloud

- Eg, “bio” related datasets
  - done, partially, by the “Linking Open Drug Data” task force of the HCLS IG at W3C
Three technologies have emerged

- To re-use thesauri, glossaries, etc: **SKOS**
- To define more complex vocabularies with a strong logical underpinning: **OWL**
- Generic framework to define rules on terms and data: **RIF**
SKOS

- Represent and share classifications, glossaries, thesauri, etc
  - for example:
    - Dewey Decimal Classification, Art and Architecture Thesaurus, ACM classification of keywords and terms…
    - classification/formalization of Web 2.0 type tags
- Define classes and properties to add those structures to an RDF universe
  - allow for a quick port of this traditional data, combine it with other data
Example: the term “Fiction”, as defined by the Library of Congress

**Fiction**

- **URI**: <http://id.loc.gov/authorities/sh85048050#concept>
- **Type**: Topical Term
- **Alternate Labels**: Fiction--Philosophy; Metafiction; Novellas (Short novels); Novels; Stories
- **Broader Terms**:
  - Literature
  - Prose literature
- **Narrower Terms**:
  - Adventure stories
  - Allegories
  - Alternative histories (Fiction)
  - Bildungsromans
  - Biographical fiction
Example: the term “Fiction”, as defined by the Library of Congress
Thesauri have identical structures...

- The structure of the LOC page is fairly typical
  - label, alternate label, narrower, broader, …
  - there is even an ISO standard for these
- SKOS provides a basic structure to create an RDF representation of these
LOC’s “Fiction” in SKOS/RDF
Usage of the LOC graph

The Glass Palace

http://.../isbn/...

Historical Fiction

Fiction

skos:broader

skos:prefLabel

dc:title

dc:subject

skos:prefLabel

skos:prefLabel

rdf:type

skos:Concept

Historical Fiction

Fiction

The Glass Palace

http://.../isbn/...
Same serialized

<http://isbn/000651409X>
   dc:title "The Glass Palace"@en;
   dc:subject <http://id.loc.gov/authorities/sh85061165#concept>;
   ...

<http://id.loc.gov/authorities/sh85061165#concept>
   a skos:Concept;
   skos:prefLabel "Historical Fiction"@en;
   skos:broader <http://id.loc.gov/authorities/sh85048050#concept>;
   ...

<http://id.loc.gov/authorities/sh85048050#concept>
   a skos:Concept;
   skos:prefLabel "Fiction"@en;
   skos:narrower <http://id.loc.gov/authorities/sh85061165#concept>;
   ...
SKOS terms overview

- Classes and Properties:
  - Basic description (Concept, ConceptScheme,...)
  - Labeling (prefLabel, altLabel,...)
  - Documentation (definition, historyNote,...)
  - Semantic relations (broader, narrower, related,...)
  - Collections (Collection, OrderedCollection,...)
  - Concept mappings (broadMatch, narrowMatch,...)
Importance of SKOS

- SKOS provides a simple bridge between the “print world” and the (Semantic) Web
- Thesauri, glossaries, etc, from the library community can be made available
  - LOC is a good example
- SKOS can also be used to organize, eg, tags, annotate other vocabularies, …
Importance of SKOS

- Anybody in the World can refer to common concepts
  - they mean the same for everybody
- Applications may exploit the relationships among concepts
  - eg, SPARQL queries may be issued on the library data+LOC