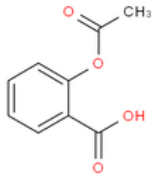
Describing Scientific Datasets: The HCLS Community Profile

Alasdair J G Gray
A.J.G.Gray@hw.ac.uk
alasdairjggray.co.uk
@gray_alasdair

Michel Dumontier
Stanford University

M. Scott Marshall
MAASTRO Clinic



Aspirin

Structure

Draw Molecule

$C_9H_8O_4$

The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5)

AlogP	1.19
# H-Bond Acceptors	4
# H-Bond Donors	1
Mol Weight	180.157

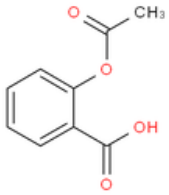
ChemSpider ID	OPS403534
SMILES	CC(=O)OC1=CC=CC=C1C(=O)O
Standard InChI	InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)
Standard InChIKey	BSYNRYMUTXBXSQ-UHFFFAOYSA-N
Protein Binding	High (99.5%) to albumin. Decreases as plasma salicylate concentration increases, with reduced plasma albumin concentration or renal dysfunction, and during pregnancy.
Toxicity	Oral, mouse: LD₅₀ = 250 mg/kg; Oral, rabbit: LD₅₀ = 1010 mg/kg; Oral, rat: LD₅₀ = 200 mg/kg. Effects of overdose include: tinnitus, abdominal pain, hypokalemia, hypoglycemia, pyrexia, hyperventilation, dysrhythmia, hypotension, hallucination, renal failure, confusion, seizure, coma, and death.

Open PHACTS

Open Pharmacological Space

Open PHACTS Explorer Alpha | Aspirin

Home / Aspirin / Compound Information



Aspirin

Pharmacology (2677

Structure

Draw Molecule

The prototypical ana
antipyretic propertie
biosynthesis of pros
arterial and venous t

AlogP

1.19

H-Bond Acceptors

H-Bond Donors

Mol Weight

180.157

ChemSpid

SMILES CC(=O)OC1=CC=CC=C1C(=O)O

Standard InChI InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)

Standard InChIKey BSYNRYMUTXBXSQ-UHFFFAOYSA-N

Protein Binding High (99.5%) to albumin. Decreases as plasma salicylate concentration increases, with reduced plasma albumin concentration or renal dysfunction, and during pregnancy.

Toxicity Oral, mouse: LD₅₀ = 250 mg/kg; Oral, rabbit: LD₅₀ = 1010 mg/kg; Oral, rat: LD₅₀ = 200 mg/kg. Effects of overdose include: tinnitus, abdominal pain, hypokalemia, hypoglycemia, pyrexia, hyperventilation, dysrhythmia

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:void="http://rdfs.org/ns/void#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:msg0="http://www.openphacts.org/api#"
  xmlns:chembl25="http://rdf.ebi.ac.uk/resource/chembl/molecule/CHEM
  xmlns:drugbank="http://www4.wiwiss.fu-berlin.de/drugbank/resource/
  xmlns:linked-data="http://purl.org/linked-data/api/vocab#">
  <rdf:Description rdf:about="http://www.conceptwiki.org/concept/dd7
    <skos:exactMatch rdf:resource="http://www4.wiwiss.fu-berlin.de/d
    <skos:exactMatch rdf:resource="http://www.conceptwiki.org/concep
    <skos:exactMatch rdf:resource="http://ops.rsc.org/OPS403534"/>
    <skos:exactMatch rdf:resource="http://rdf.ebi.ac.uk/resource/cher
    <void:inDataset rdf:resource="http://www.conceptwiki.org"/>
  </rdf:Description>
    <void:inDataset rdf:resource="http://ops.rsc.org"/>
    <msg0:inchi>InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-
    <msg0:inchikey>BSYNRYMUTXBXSQ-UHFFFAOYSA-N</msg0:inchikey>
    <msg0:logp rdf:datatype="http://www.w3.org/2001/XMLSchema#double"
    <msg0:hba rdf:datatype="http://www.w3.org/2001/XMLSchema#double">
    <msg0:hbd rdf:datatype="http://www.w3.org/2001/XMLSchema#double">
```

Core Platform

Historic Use Case

~January 2012

Open PHACTS v1.3

ChEMBL 16

http://tiny.cc/ops-datasets

ChEMBL
v13

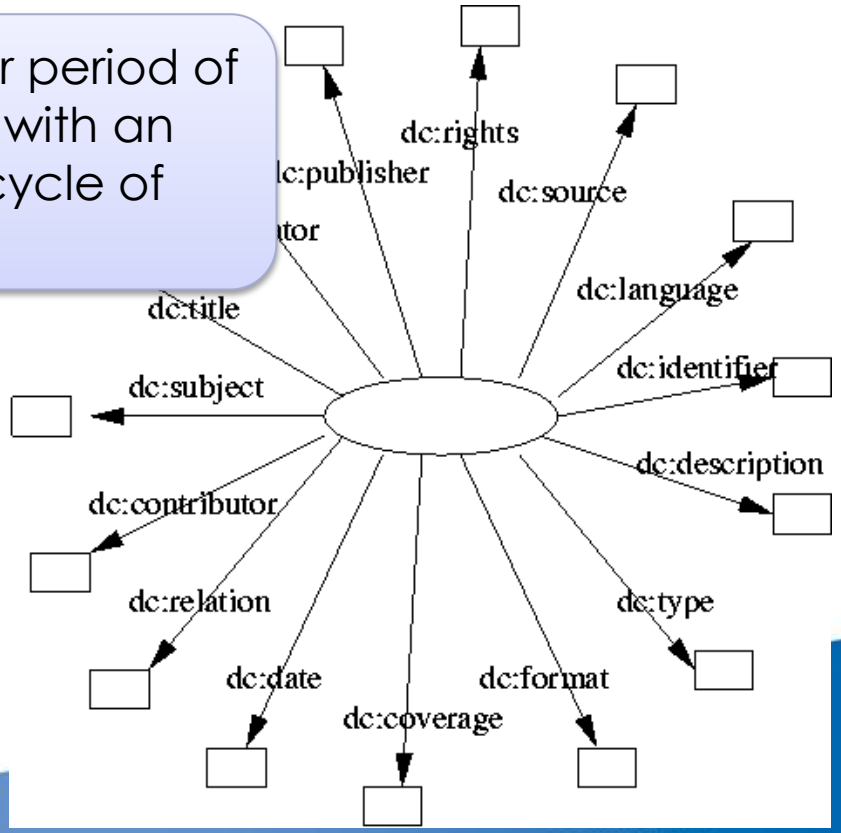
Challenges

- Datasets available
  - In many versions over time
  - In different formats
  - From many mirrors/registries
- Datasets build on each other
- Files do not carry metadata
- Registries
  - Can be out-of-date
  - Can contain conflicting information

Dublin Core Metadata Initiative

- ✓ Widely used
- ✓ Broadly applicable
  - Documents
  - Datasets

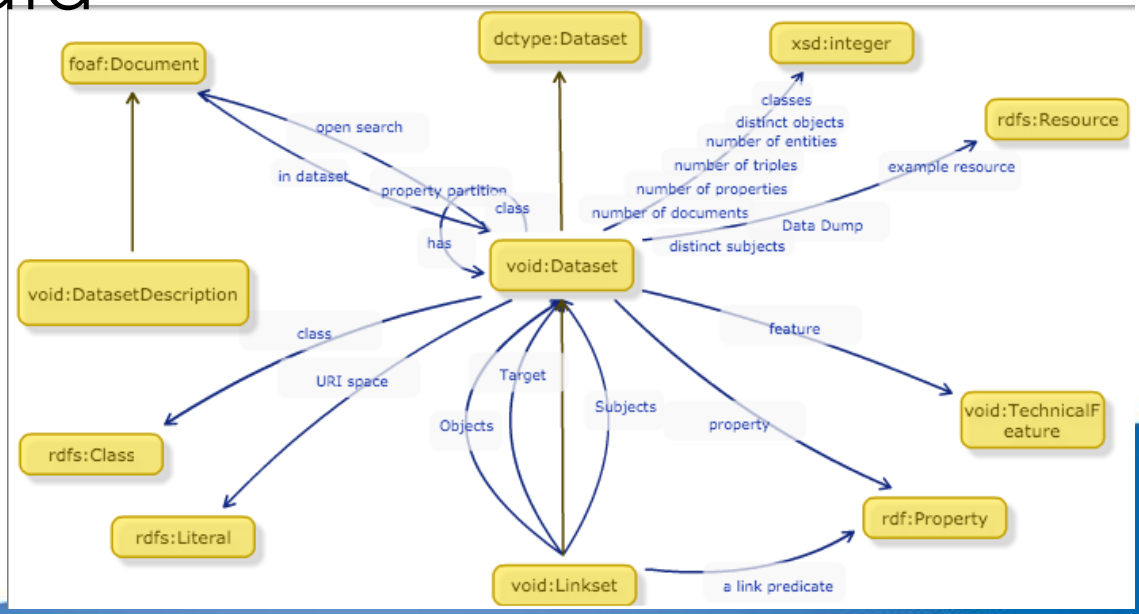
"Date: A point or period of time associated with an event in the lifecycle of the resource."



- ✗ Generic terms
- ✗ Not comprehensive
- ✗ No required properties

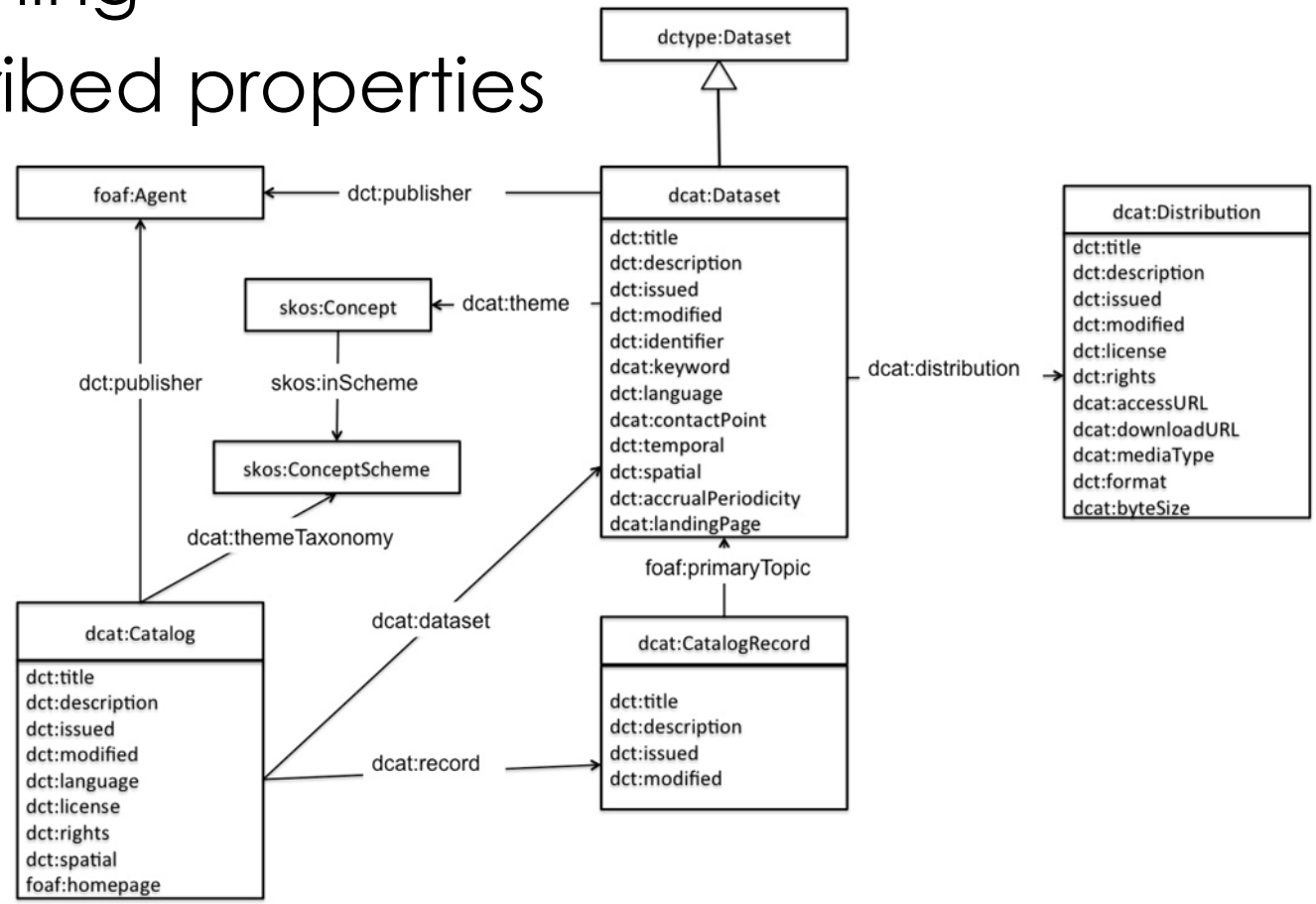
VoID: Vocabulary of Interlinked Datasets

- ✓ Metadata carried with data
  - Directly embedded: `void:inDataset`
- ✗ No versioning
- ✗ No checklist of requisite fields
- ✗ Only for RDF data



DCAT: Data Catalog

- ✓ Separates Dataset and Distribution
- ✗ No versioning
- ✗ No prescribed properties



Dataset Descriptions: HCLS Community Profile

Editors working draft.

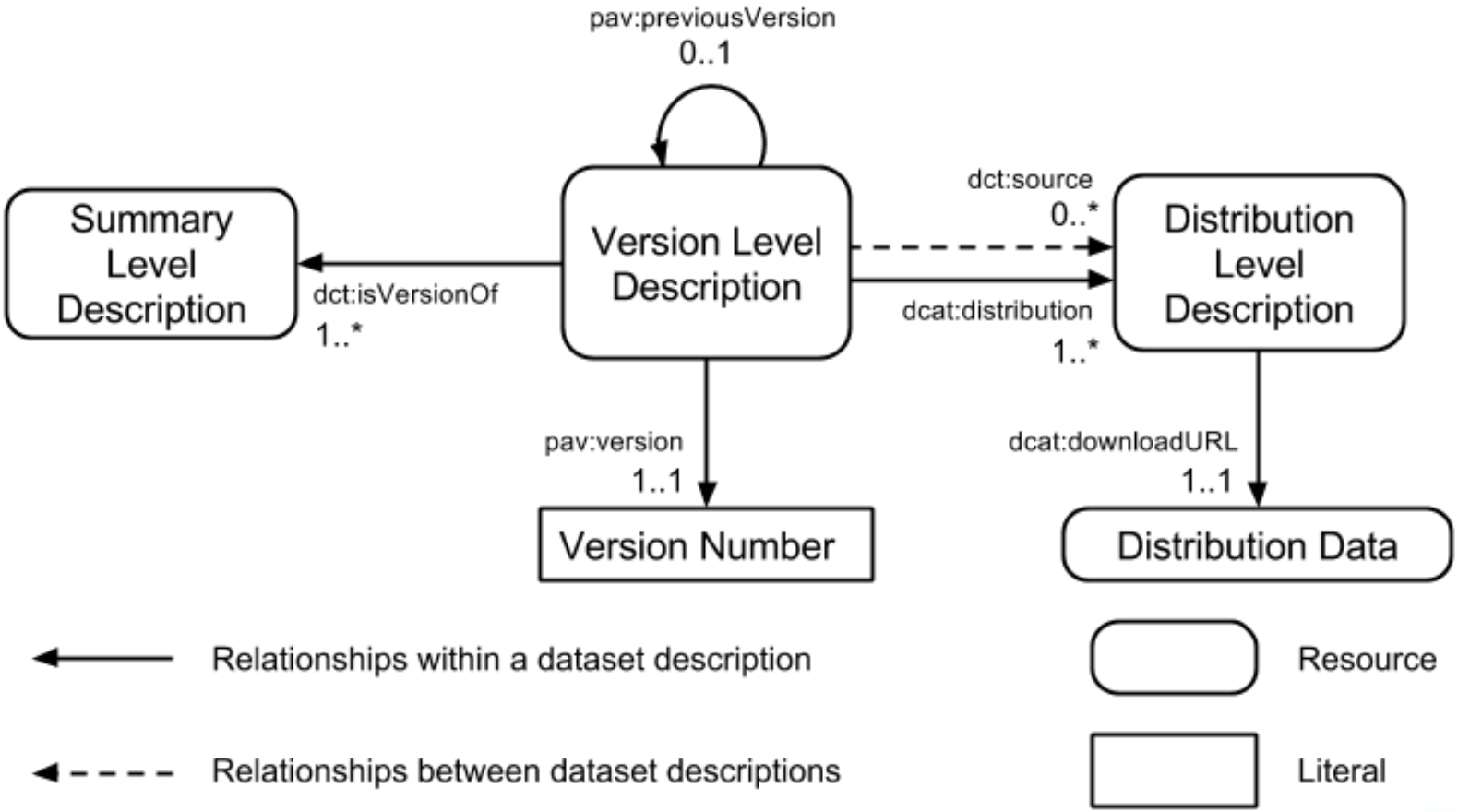
Editors:

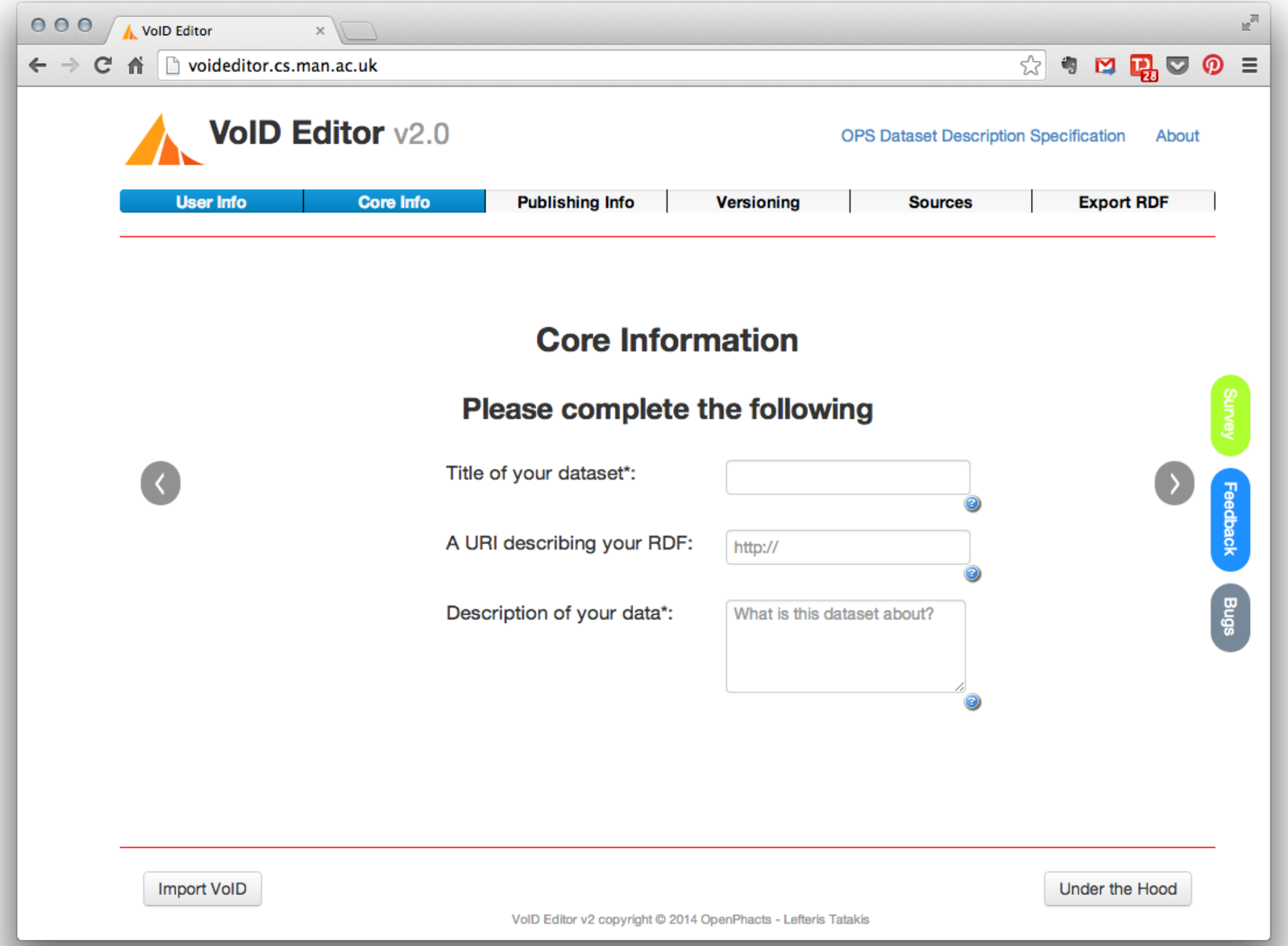
Alasdair J.G. Gray, Heriott-Watt University, UK <A.J.G.Gray@hw.ac.uk>
Michel Dumontier, Stanford University, USA <michel.dumontier@stanford.edu>
M. Scott Marshall, MAASTRO Clinic, The Netherlands <m.scott.marshall@maastro.nl>
Joachim Baran, Stanford University, USA <joachim.baran@stanford.edu>

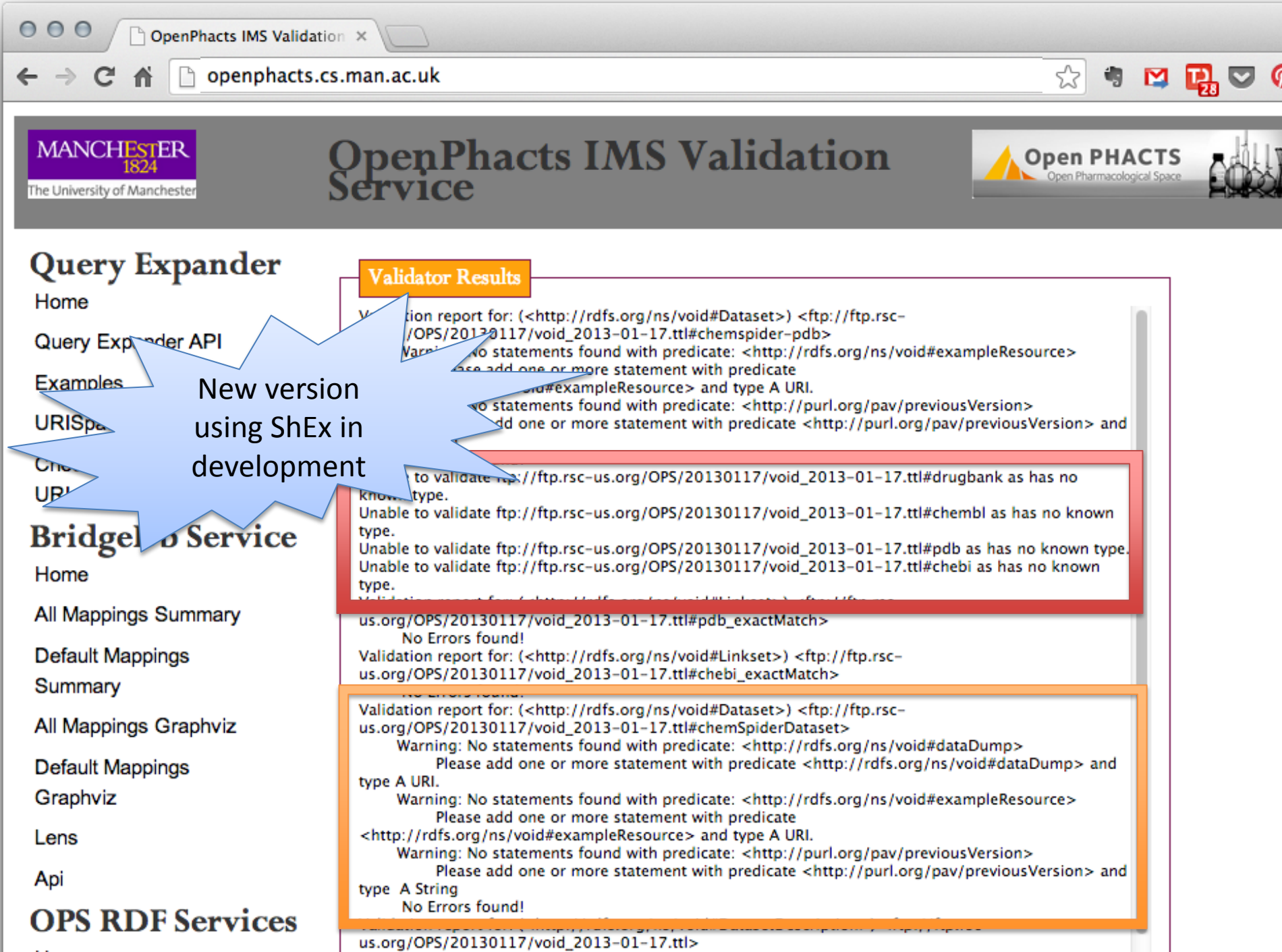
Contributors:

Peter Ansell, CSIRO, Australia <peter.ansell@csiro.au>
Gary D. Bader, The Donnelly Centre, University of Toronto, Canada <gary.bader@utoronto.ca>
Asuka Bando, NBDC, Japan <bando@biosciencedbc.jp>
Jerven Bolleman, SIB Swiss Institute of Bioinformatics, Switzerland <jerven.bolleman@isb-sib.ch>
Alison Callahan, Carleton University, Canada <alison.callahan@carleton.ca>
José Cruz-Toledo, Carleton University, Canada <josecruztoledo@cmail.carleton.ca>
Pascale Gaudet, SIB Swiss Institute of Bioinformatics, Switzerland <pascale.gaudet@isb-sib.ch>
Erich Gombocz, IO Informatics, USA <egombocz@io-informatics.com>
Alejandra Gonzalez-Beltran, University of Oxford, UK <alejandra.gonzalez.beltran@gmail.com>
Paul Groth, VU University Amsterdam, The Netherlands < p.t.groth@vu.nl>
Melissa Haendel, Oregon Health and Science University, USA <haendel@ohsu.edu>
Maori Ito, NIBIO, Japan <maori@nibio.go.jp>
Simon Jupp, EMBL-EBI, UK <jupp@ebi.ac.uk>
Toshiaki Katayama, Database Center for Life Sciences, Japan <ktym@dbcls.jp>
Kalpana Krishnaswami, Metaome, USA <kalpana@metaome.com>
Simon Lin, Marshfield Clinic Research Foundation, USA <lin.simon@mcrf.mfldclin.edu>
Michael Miller, Institute for Systems Biology, USA <mmiller@systemsbiology.org>
Chris Mungall, Lawrence Berkeley National Laboratory, USA <cjm@berkeleybop.org>
Nicolas Le Novère, Babraham Institute, UK <n.lenovere@gmail.com>
Camille Laibe, EMBL-EBI, UK <laibe@ebi.ac.uk>
Nick Juty, EMBL-EBI, UK <juty@ebi.ac.uk>
James Malone, EMBL-EBI, UK <malone@ebi.ac.uk>
Laurens Rietveld, VU University Amsterdam, The Netherlands <laurens.rietveld@vu.nl>
Sarala M. Wimalaratne, EMBL-EBI, UK <sarala@ebi.ac.uk>

HCLS Dataset Descriptions



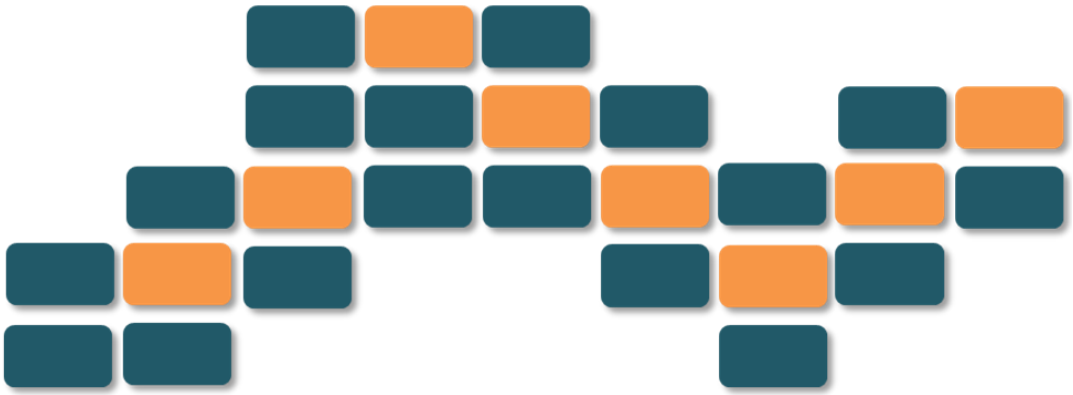

VoID Editor
voideditor.cs.man.ac.uk
VoID Editor v2.0
OPS Dataset Description Specification
About
User Info
Core Info
Publishing Info
Versioning
Sources
Export RDF
Core Information
Please complete the following
Title of your dataset*:
A URI describing your RDF:
http://
Description of your data*:
What is this dataset about?
Survey
Feedback
Bugs
Import VoID
Under the Hood
VoID Editor v2 copyright © 2014 OpenPhacts - Lefteris Tatakis


OpenPhacts IMS Validation
openphacts.cs.man.ac.uk
MANCHESTER 1824
The University of Manchester
OpenPhacts IMS Validation Service
Open PHACTS
Open Pharmacological Space
Query Expander
Home
Query Expander API
Examples
URISpace
URI
Bridgedb Service
Home
All Mappings Summary
Default Mappings Summary
All Mappings Graphviz
Default Mappings Graphviz
Lens
Api
OPS RDF Services
New version using ShEx in development
Validator Results
tion report for: (<http://rdfs.org/ns/void#Dataset>) <ftp://ftp.rsc-
/OPS/20130117/void_2013-01-17.ttl#chemspider-pdb>
No statements found with predicate: <http://rdfs.org/ns/void#exampleResource>
ase add one or more statement with predicate
id#exampleResource> and type A URI.
o statements found with predicate: <http://purl.org/pav/previousVersion>
dd one or more statement with predicate <http://purl.org/pav/previousVersion> and
to validate ftp://ftp.rsc-us.org/OPS/20130117/void_2013-01-17.ttl#drugbank as has no
type.
Unable to validate ftp://ftp.rsc-us.org/OPS/20130117/void_2013-01-17.ttl#chembl as has no known
type.
Unable to validate ftp://ftp.rsc-us.org/OPS/20130117/void_2013-01-17.ttl#pdb as has no known type.
Unable to validate ftp://ftp.rsc-us.org/OPS/20130117/void_2013-01-17.ttl#chebi as has no known
type.
us.org/OPS/20130117/void_2013-01-17.ttl#pdb_exactMatch>
No Errors found!
Validation report for: (<http://rdfs.org/ns/void#Linkset>) <ftp://ftp.rsc-
us.org/OPS/20130117/void_2013-01-17.ttl#chebi_exactMatch>
Validation report for: (<http://rdfs.org/ns/void#Dataset>) <ftp://ftp.rsc-
us.org/OPS/20130117/void_2013-01-17.ttl#chemSpiderDataset>
Warning: No statements found with predicate: <http://rdfs.org/ns/void#dataDump>
Please add one or more statement with predicate <http://rdfs.org/ns/void#dataDump> and
type A URI.
Warning: No statements found with predicate: <http://rdfs.org/ns/void#exampleResource>
Please add one or more statement with predicate
<http://rdfs.org/ns/void#exampleResource> and type A URI.
Warning: No statements found with predicate: <http://purl.org/pav/previousVersion>
Please add one or more statement with predicate <http://purl.org/pav/previousVersion> and
type A String
No Errors found!
us.org/OPS/20130117/void_2013-01-17.ttl>

Future Vision

- Provide rich and accurate provenance trail of data
  - Write once, use many times
    - Automatic pipeline from description file to registries
- FAIR Data

Find
Access
Interoperate
Re-use
Data



Thank you

Editors' Draft:

http://tiny.cc/hcls-datadesc-ed

W3C Interest Group Note:

http://tiny.cc/hcls-datadesc

Acknowledgements to W3C HCLS Group

- www.alasdairjggray.co.uk
- A.J.G.Gray@hw.ac.uk
- @gray_alasdair