

DIGITAL TOOLS FOR HUMANISTS
SUMMER SCHOOL 2021, University of Pisa

Open Data, Linked Data, Linked Open Data,
the Semantic Web, and Knowledge Sharing and
Discovery
Friday, 4 June 2021



Dr Seamus Ross,
Professor, Faculty of Information, University of Toronto

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

1

1

Some of the slides in this presentation contain copyright material which has been copied and made available to you under section 30.01 of Canada's Copyright Act. Wherever possible the sources of this material has been flagged on the individual slides themselves either as links listed at the bottom of the slide or because the source of the screenshots is clearly identifiable

Recording the slides of this lecture would infringe copyright.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

2

2

INTRODUCTION

- Welcome and Introduction
- Who am I
- Overview of the day
 - Lectures in Morning
 - Interactive Activities
 - Experimentation in Afternoon
- Timetable
 - 09:00 – 10:30 Lecture
 - 10:30 – 11:00 Break
 - 11:00 – 12:30 Lecture
 - 12:30 – 14:00 Lunch
 - 14:00 – 15:30 Team Activities
 - 15:30 – 16:00 Break
 - 16:00 – 17:00 Experimentation
 - 17:00 – 17:30 Discussion

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 3

3

WHO AM I

- Seamus Ross
- Professor Faculty of Information
University of Toronto
- Fields of Engagement:
 - Information
 - Digital Curation/Preservation
 - Digital Archaeology
 - Cultural Heritage
 - Knowledge Representation and Reasoning

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 4

4



<https://euclid-project.eu/>



Elena Simperl, et al., 2013, *Using Linked Data Effectively*, The Euclid Project Consortium,
<https://books.apple.com/gb/book/using-linked-data-effectively/id783647393>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 5

5

SUGGESTED READINGS/RESOURCES: (More about them during the lecture).

- ❑ Elena Simperl, et al., 2013, *Using Linked Data Effectively*, The Euclid Project Consortium, <https://books.apple.com/gb/book/using-linked-data-effectively/id783647393>
- ❑ Eero Hyvonen, 2019, "Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery," *Semantic Web – Interoperability, Usability, Applicability*, (Tracking #2310-3523), <http://semantic-web-journal.net/content/using-semantic-web-digital-humanities-shift-data-publishing-data-analysis-and-serendipitous#>
- ❑ Kaylan Dutia and John Stack, 2021, "Heritage Connector: A machine learning framework for building linked open data from museum collections," *Applied AI Letters*, 3 May 2021, <https://doi.org/10.1002/aill2.23>
- ❑ Dominik Lukas, Claudia Engel and Camilla Mazzucato, 2018, "Towards a Living Archive: Making Multi Layered Research Data and Knowledge Generation Transparent", *Journal of Field Archaeology*, 43:sup1, S19-S30, DOI: 10.1080/00934690.2018.1516110
- ❑ *Cogan Shimizu, Pascal Hitzler, Quinn Hirt, Dean Rehberger, Seila Gonzalez Estrecha, Catherine Foley, Alicia M. Sheill, Walter Hawthorne, Jeff Mixter, Ethan Watrall, Ryan Carty, Duncan Tarr, 2020, "The enslaved ontology: Peoples of the historic slave trade," *Journal of Web Semantics*, V 63, <https://doi.org/10.1016/j.jwebsem.2020.100567>.
- ❑ *Lyne Da Sylva, 2018. "Towards linked data: Some consequences for researchers in the social sciences and humanities". *Proceedings of the Association for Information Science and Technology*, 55(1), 94–103. <https://doi.org/10.1002/pr2.2018.14505501011>
- ❑ *Mauro Dragoni, Sara Tonelli, and Giovanni Moretti. 2017. A Knowledge Management Architecture for Digital Cultural Heritage. *J. Comput. Cult. Herit.* 10, 3, Article 15 (August 2017), 18 pages. DOI:<https://doi.org/10.1145/3012289>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 6

6

WHAT WE WILL COVER

- Open Data Concept and Origins
- Open Data
- Linked Data
- Linked (Open) Data
- RDF
- SPARQL
- Semantic Web
- Meaning Making and Knowledge Discovery
- Open Science
- Digital Humanities

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 7

7

Download : [Download high-res image \(118KB\)](#) Download : [Download full-size image](#)

Fig. 1. Components of Digital Humanities.

From: Lora Aroyo, Franciska de Jong, Eero Hyvönen, Sara Tonelli, 2020, "Web Semantics for Digital Humanities," *Journal of Web Semantics*, Volumes 61–62, Figure 1.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 8

8

LINKED
(OPEN)
DATA

- Concept of “Open”
- Transparency
- Reproducibility
- Interconnectedness of scholarship or business need
- Semantic Potential and Value
- Knowledge discovery
- Quality, Data, Links, Ontologies/Vocabularies
- Assessment and Validation

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 9

9

STARTING
WITH
WHY

- Open Data Concept and Origins
- Open Data
- Linked Data
- Linked (Open) Data
- RDF
- SPARQL
- Semantic Web
- Meaning Making and Knowledge Discovery
- Open Science

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 10

10

The diagram illustrates the architecture of a Linked Data Server. It features a central 'Link Manager' (represented by a blue circle with yellow nodes) connected to a 'Web' (globe icon), a 'Client' (computer icon), a 'Data Source' (server rack icon), a 'Data Transformer' (server rack icon), and a 'Data Manager' (server rack icon). Arrows indicate the flow of data and interactions between these components. The entire system is enclosed in a light blue oval.

Fig. 1. Linked Data Server

Nicola Aloia, Cesare Concordia, and Carlo Meghini, 2013, "The Europeana Linked Open Data Pilot Server" in Maristella Agosti, Floriana Esposito, Stefano Ferilli, and Nicola Ferro. Digital Libraries and Archives: 8th Italian Research Conference, IRCDL 2012, Bari, Italy, February 9-10, 2012, Revised Selected Papers. Springer Berlin Heidelberg, 2013. P 253 -

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 11

11

FAIR DATA PRINCIPLES

FAIR: Findability, Accessibility, Interoperability, and Reusability.

Should FAIR data principles from e-Science apply beyond e-Science?

Not a concept inherently tied to Linked Data, but closely linked with Open Data.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 12

12

Mark D Wilkinson, et al., 2016, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, 3
<https://dx.doi.org/10.1038/sdata.2016.18>

Box 2 | The FAIR Guiding Principles

To be Findable:
 F1. (meta)data are assigned a globally unique and persistent identifier
 F2. data are described with rich metadata (defined by R1 below)
 F3. metadata clearly and explicitly include the identifier of the data it describes
 F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
 A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 A1.1 the protocol is open, free, and universally implementable
 A1.2 the protocol allows for an authentication and authorization procedure, where necessary
 A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
 I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
 I2. (meta)data use vocabularies that follow FAIR principles
 I3. (meta)data include qualified references to other (meta)data

To be Reusable:
 R1. meta(data) are richly described with a plurality of accurate and relevant attributes
 R1.1. (meta)data are released with a clear and accessible data usage license
 R1.2. (meta)data are associated with detailed provenance
 R1.3. (meta)data meet domain-relevant community standards

13

CARE DATA PRINCIPLES

CARE: Collective Benefit, Authority to Control, Responsibility, Ethics

Do CARE data principles any new ways of thinking about Gov't Open Data?

Research Data Alliance International Indigenous Data Sovereignty Interest Group. (September 2019). "CARE Principles for Indigenous Data Governance." The Global Indigenous Data Alliance
https://nnigovernance.arizona.edu/sites/default/files/resources/CARE%20Principles_One%20Partners%20FINAL_Oct_17_2019.pdf

© Seamus Ross, FI at UoT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 14

14

CARE GUIDING PRINCIPLES


Collective Benefit:
 C1: For inclusive development and innovation
 C2: For improved governance and citizen engagement
 C3: For equitable outcomes

Authority to Control:
 A1: Recognizing rights and interests
 A2: Data for governance
 A3: Governance of data

Responsibility:
 R1: For positive relationships
 R2: For expanding capability and capacity
 R3: For Indigenous languages and worldviews

Ethics:
 E1: For minimizing harm and maximizing benefit
 E2: For justice
 E3: For future use

Research Data Alliance International Indigenous Data Sovereignty Interest Group. (September 2019). "CARE Principles for Indigenous Data Governance." The Global Indigenous Data Alliance https://indigenous.arts.utoronto.ca/sites/default/files/resources/CARE%20Principles_One%20Pages%20FINAL_Oct_17_2019.pdf



15

COMMENTARY

- Linked
- Open
- Data

© Seamus Ross, FI at UoT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

16

WINNIE-THER-POOH 'COMING DOWNSTAIRS' AND THE LINKED (OPEN) DATA CONUNDRUM.



“Here is Edward Bear, coming down the stairs now, bump, bump, bump, on the back of his head, behind Christopher Robin. It is, as far as he knows, the only way of coming downstairs, but sometimes he feels that there really is another way, if only he could stop bumping for a moment and think of it. And then he feels that perhaps there isn't. Anyhow, here he is at the bottom, and ready to be introduced to you. Winnie-the-Pooh.”

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

17

17

LET'S START
BRIEFLY
WITH

- OPEN DATA

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

18

18

OPENNESS WHAT IS IT?

WHAT IS OPEN KNOWLEDGE? IS IT 'OPEN KNOWLEDGE' WITHOUT ANY LEGAL, POLITICAL, OR ECONOMIC BARRIERS?

The Open Definition

The **Open Definition** sets out principles that define "openness" in relation to **data and content**.

It makes **precise** the meaning of "open" in the terms "**open data**" and "**open content**" and thereby ensures **quality** and encourages **compatibility** between different pools of open material.

It can be summed up in the statement that:

*"Open means **anyone** can **freely access, use, modify, and share** for **any purpose** (subject, at most, to requirements that preserve provenance and openness)."*

Put most succinctly:

*"Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**."*

Sources: <https://okfn.org/opendata/> & <http://opendefinition.org/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 19

19

OPEN KNOWLEDGE FOUNDATION ARGUES

"Open Data becomes Open Knowledge when it is useful, usable, and used."

- Availability and Access
- Reuse and Redistribution
- Universal Participation

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 20

20

OPEN DATA OF ALL FLAVOUR – VANILLA AND LINKED CAN CREATE

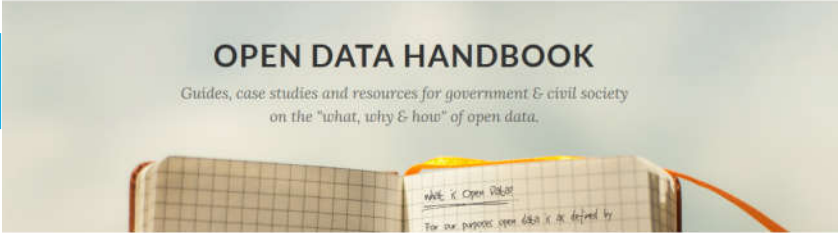
- Economic Value
- Social Good
- Cultural Power
- Research Possibilities

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 21

21

OPEN DATA HANDBOOK

Guides, case studies and resources for government & civil society on the "what, why & how" of open data.



- Open Data Guide**
This guide discusses the legal, social and technical aspects of open data. It can be used by anyone but is especially designed for those seeking to open up data. It discusses why to go open, what open is, and the how to 'open' data.
[Start Reading](#)
- Value Stories**
Use cases, stories and case studies highlighting the social and economic value, the impact and the varied applications of open data from cities and countries across the globe.
[Value Stories](#)
- Resource Library**
A curated collection of open data resources, including articles, longer publications, how-to guides, presentations and videos, produced by the global open data community.
[Open Data Resources](#)

Sources: <http://opendatahandbook.org/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 22

22

15 PRINCIPLES OF UNDERLYING OPEN GOV'T DATA ARISING FROM SEBASTOPOL 2007 AND SUBSEQUENT MTG

- Complete
- Primary
- Timely
- Accessible
- Machine Processible
- Non-discriminatory (e.g., anonymous access)
- Non-proprietary
- License Free
- Online & Free
- Permanent
- Trusted
- A presumption of openness
- Documented
- Safe to open
- Designed with public input

<https://opengovdata.org/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

23

23

What kinds of open data?

There are many kinds of open data that have potential uses and applications:

- **Culture:** Data about cultural works and artefacts – for example titles and authors – and generally collected and held by galleries, libraries, archives and museums.
- **Science:** Data that is produced as part of scientific research from astronomy to zoology.
- **Finance:** Data such as government accounts (expenditure and revenue) and information on financial markets (stocks, shares, bonds etc).
- **Statistics:** Data produced by statistical offices such as the census and key socioeconomic indicators.
- **Weather:** The many types of information used to understand and predict the weather and climate.
- **Environment:** Information related to the natural environment such presence and level of pollutants, the quality and rivers and seas.



- Sources: <https://okfn.org/opendata/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

24

24

JANSSEN, ET. AL, 2012 – A REFLECTIVE PRESPECTIVE WORTH BRING TO LINKED (OPEN) DATA

Open Data Myths

- Publicizing data creates benefits
- All content/data should be released
- Publishing Data is sufficient
- All Citizens can use Open Data
- Open Data make for Open Gov't

Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk, 2012, "Benefits, Adoption Barriers and Myths of Open Data and Open Government," *Information Systems Management*, 29(4), pp., 258-268,

25

CONTRASTING VIEWS FROM 1990 AND EARLY 2000'S

From Peter Weiss, 2002, *Borders in Cyberspace: Conflicting Public Sector Information Policies and their Economic Impacts (Summary Report)*, <http://www.nws.noaa.gov/sp/Borders-report.pdf-p.3> (currently a dead link)

See his presentation at LISBON ERPANET SEMINAR 2003, <https://www.erpanet.org/events/2003/lisbon/presentations/Weiss%20presentation.pdf>

<https://www.nap.edu/report/11030/chapter/18>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 26

26

Majority of Americans say they are more apt to trust research when the data is openly available

% of U.S. adults who say when they hear each of the following, they trust scientific research findings ...

Condition	Less	More	Makes no difference
Data is openly available to the public	8%	57%	34%
Reviewed by an independent committee	10%	52%	37%
Funded by the federal government	26%	23%	48%
Funded by an industry group	58%	10%	32%

% of U.S. adults who say when they hear each of the following, they trust a science practitioner's recommendation ...

Condition	Less	More	Makes no difference
Open to getting a second opinion	7%	68%	23%
Based on review from an independent committee	17%	43%	38%
Received financial incentives from the government	37%	14%	48%
Received financial incentives from an industry group	62%	10%	27%

Note: Respondents who did not give an answer are not shown.
Source: Survey conducted Jan. 7-21, 2019.
"Trust and Mistrust in Americans' Views of Scientific Experts"

PEW RESEARCH CENTER

From "3. Americans say open access to data and independent review inspire more trust in research findings," of the Pew Research Center, August 2019, "Trust and Mistrust in Americans' Views of Scientific Experts" (Cary Funk, Meg Heffernon, Brian Kennedy, and Courtney Johnson, <https://www.pewresearch.org/science/2019/08/02/americans-say-open-access-to-data-and-independent-review-inspire-more-trust-in-research-findings/> (see pages 24-27).

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 27

27

REPRESENTATION

- How is Open Data Represented
 - Hierarchically
 - Relationally
 - As a Graph

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 28

28

<http://www.providenceri.gov/police/crime-statistics/>

29

<http://www.providenceri.gov/wp-content/uploads/2019/02/20190224.pdf>

30

CaseNumber	Location	Reported Date	Month	Year	Offense Desc	Statute Code	Statute Desc	Counts	Reporting Officer
2018-00091670	325 WASHINGTON ST	8/31/2018 17:16	8	2018	Medical Aid	Not Used	No violations	0	
2018-00092806	64 EATON ST	9/2/2018 23:41	9	2018	Liquor Law Violations	3/8/2010	POSSESSION OF BEVERAGE- UNDERAGE PERSONS	1	
2018-00093479	MILLER AVE & BROAD ST	9/4/2018 16:21	9	2018	Municipal Code Violation	Sec. 18-21	Sale, use, possession of alcoholic beverages in parks, etc.	1	
2018-00093900	25 SORRENTO ST	9/5/2018 18:54	9	2018	Drug Offenses	21-28-4.01-A1	MANUFAC/POSS/DELIVER SCH 1/II-DRUG DEPEND	1	
2018-00094373	140 PEMBROKE AVE	9/7/2018 0:26	9	2018	Municipal Code Violation	Sec. 16-93	Noise Control - Radios, television sets, and similar devices.	1	
2018-00096462	160 BENEDICT ST	9/12/2018 15:02	9	2018	Larceny, Other	11-41-1	LARCENY/U \$1500 - ALL OTH LARCENY	1	
2018-00097067	284 VEAZIE ST	9/14/2018 2:00	9	2018	Burglary	11/8/2001	BURGLARY	1	
2018-00097093	280 BROAD ST	9/14/2018 10:23	9	2018	Municipal Code Violation	Sec. 18-21	Sale, use, possession of alcoholic beverages in parks, etc.	1	
2018-00097070	340 BROAD ST	9/14/2018 9:18	9	2018	Municipal Code Violation	Sec. 18-21	Sale, use, possession of alcoholic beverages in parks, etc.	1	
2018-00097070	340 BROAD ST	9/14/2018 9:18	9	2018	RI Statute Violation	12/9/2016	WARRANT OF ARREST ON AFFIDAVIT - ALL OTH OFFENSE	1	
2018-00097282	99 KENNEDY PLZ	9/14/2018 18:05	9	2018	Assault, Simple	11/5/2003	SIMPLE ASSAULT OR BATTERY	1	
					Liquor Law		POSSESSION OF BEVERAGE-		

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 31

31

PROPUBLICA

TOPICS ▾ SERIES ▾ NEWS APPS GET INVOLVED IMPACT ABOUT ▾

TRUMP, INC.
How a Nigerian Presidential Candidate Hired a Trump Lobbyist and Elected in Trump's Lobby "Trump, Inc."
 We spent a night at President Trump's hotel in Washington, D.C., and we met some interesting people.
 by Katherine Sullivan, Feb. 27, 8 a.m.

Featured Series
DOLLARS FOR DOCTORS
 How Industry Money Rescues Physicians

The Mission
 To expose abuses of power and betrayals of the public trust by government, business, and other institutions, using the moral force of investigative journalism to spur reform through the sustained spotlighting of wrongdoing.

THE NAVY'S DISASTER IN THE PACIFIC
Navy Leaders Taken to Task by Lawmakers, Including One Who Was Grilling a Former Boss
 by T. Christian Miller and Robert Faruqi, Feb. 26, 6:58 p.m. EST

HEART FAILURE
Numerous Mistakes Led to Fatal Blood Transfusion at St. Luke's in Houston, Report Finds
 by Mike Isaacbaugh, Houston Chronicle, and Chae Park, Feb. 26, 6:02 p.m. EST

ProPublica is an independent, nonprofit newsroom that produces investigative journalism with moral force. We dig deep into important issues, shining a light on abuses of power and betrayals of public trust — and we stick with those issues as long as it takes to hold power to account.

With a team of more than 75 dedicated journalists, ProPublica covers a range of topics including government and politics, business, criminal justice, the environment, education, health care, immigration, and technology. We focus on stories with the potential to spur real-world impact. Among other positive changes, our reporting has contributed to the passage of new laws; reversals of harmful policies and practices; and accountability for leaders at local, state and national levels.

<https://www.propublica.org/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 32

32

PROPUBLICA Analytics & Data Newsletters About

DOLLARS FOR DOCTORS

We Found Over 700 Doctors Who Were Paid More Than a Million Dollars by Drug and Medical Device Companies

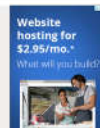
ProPublica has been tracking drug company spending on doctors since 2010. We just updated our database and found that companies are still paying private doctors huge sums for promotional talks and consulting.

By Elizabeth Gibbons, Tracy Iversen and Katherine Searles | June 16, 2018, 10:46 AM

ProPublica is a nonprofit newsroom that investigates abuses of power. Our reporters are based in Phoenix, Chicago, New York and other cities.

Back in 2014, ProPublica revealed what seemed a shocking development in the pharmaceutical industry's drive to win the prescriptions of the nation's doctors: In just four years, one doctor had received \$1 million in drug promotional talks and consulting. For drug companies, 20 doctors had each taken more than \$200,000.

Site opens later — Right-click here to open this page in a new window.



2014	2015	2016	2017	2018
Brylcreem \$25.8M	Xarelto \$20.7M	Xarelto \$20.2M	Xarelto \$23.1M	Xarelto \$17.0M
Invokana \$21.3M	Humira \$24.9M	Eliglustat \$18.8M	Envarsio \$18.7M	Farigpa \$12.8M
Xarelto \$20.2M	Invokana \$21.2M	Invokana \$18.2M	Jardiance \$17.0M	Humira \$10.2M
Eliglustat \$19.2M	Vedolizumab \$19.2M	Humira \$19.2M	Invokana \$19.0M	Jardiance \$12.2M
Brintellix \$17.7M	Eliglustat \$19.2M	Treosulf \$14.7M	Eliglustat \$10.4M	Keytruda \$11.7M
Brintellix \$15.4M	Kytrona \$19.8M	Tenipos \$13.8M	Farigpa \$14.8M	Eliglustat \$11.6M
Vincosa \$15.2M	Androgel \$15.2M	Farigpa \$13.5M	Humira \$14.4M	Rapatha \$10.9M
Letrova \$13.9M	Synthroid \$14.7M	Envarsio \$13.2M	Ashgale \$13.9M	Ashgale \$10.7M
Humira \$13.0M	Laprom \$14.0M	Rapatha \$10.0M	Rapatha \$11.4M	Envarsio \$10.2M
Ashgale \$10.0M	Vincosa \$11.9M	Onsior \$12.0M	Keytruda \$11.3M	Onsior \$10.2M
Synthroid \$8.2M	Ashgale \$11.3M	Vibralta \$11.0M	Onsior \$10.5M	Dapivonil \$10.0M
Copaxone \$8.74M	Tenipos \$11.2M	Ashgale \$10.0M	Treditor \$10.67M	Vincosa \$9.69M
Humira \$8.57M	Brintellix \$10.9M	Lincosa \$10.2M	Treosulf \$10.65M	Invokana \$9.09M
Aldyly Maltensa \$8.45M	Onsior \$10.1M	Trisenlla \$9.52M	Vincosa \$9.20M	Treditor \$8.95M
Gilroya \$7.7M	Jardiance \$9.29M	Humira \$9.10M	Letradia \$9.13M	Conveya \$7.95M
Rubra \$7.51M	Brintellix \$9.19M	Letradia \$9.09M	Hypocidil \$8.52M	Onsior \$7.70M
Protonix \$7.44M	Letradia \$8.83M	Vincosa \$7.88M	Trisenlla \$8.13M	Letradia \$7.57M
Breen \$7.23M	Glyxambi \$8.69M	Keytruda \$7.51M	Lincosa \$7.58M	Alimta \$7.54M
Levetir \$6.84M	Rapatha \$8.20M	Vincosa \$8.06M	Vibralta \$7.93M	Brintellix \$7.52M
Humira \$6.04M	Soliris \$8.07M	Letradia \$8.04M	Letradia \$7.65M	Treditor \$7.57M

Credit: Moiz Syed/ProPublica. Source: ProPublica analysis of Open Payments data from the Centers for Medicare and Medicaid Services.

<https://www.propublica.org/article/we-found-over-700-doctors-who-were-paid-more-than-a-million-dollars-by-drug-and-medical-device-companies>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 33

33

BREXIT AND OPEN DATA AT THE BOUNDARY OF LINKED OPEN DATA

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 34

34

https://petition.parliament.uk/petitions/241584 80% ... Search

We use cookies to make this service simpler. Find out more about cookies

Petitions
UK Government and Parliament

Petition
Revoke Article 50 and remain in the EU.

The government repeatedly claims exitin. We need to put a stop to this claim by pr support now, for remaining in the EU. A F so vote now.

Sign this petition

5,812,736 signatures

Show on a map 100,000

Parliament will debate this petition
Parliament will debate this petition on 1 April 2019

35

Created by
Margaret Anne Georgiadou

Deadline
20 August 2019

All petitions run for 6 months

Get petition data (json format)

About petition data

The data shows the number of people who have signed the petition by country as well as in the constituency of each Member of Parliament. This data is available for all petitions on the site. It is not a list of people who have signed the petition. The only name that is shared on the site is that of the petitioner.

The petition data is available in JSON format. The data is available for all petitions on the site. It is not a list of people who have signed the petition. The only name that is shared on the site is that of the petitioner.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 36

36

Revoke Article 50 and remain in the EU.

[SIGN HERE](#)

The government repeatedly claims exiting the EU is 'the will of the people'. We need to put a stop to this claim by proving the strength of public support now, for remaining in the EU. A People's Vote may not happen - so vote now.

Signatures: 5,812,914
AVERAGE: 29 PER MINUTE

Last update: a few seconds ago

<input type="checkbox"/> England Signatures 4,598,109	<input type="checkbox"/> More signatures than MP's majority 228 *	UK Signatures 5,573,164 (85.88%)
<input type="checkbox"/> Northern Ireland Signatures 123,200	<input type="checkbox"/> More signatures than MP's GE votes 6 *	Overseas Signatures 239,752 (4.12%)
<input type="checkbox"/> Scotland Signatures 535,427		Percentage of UK electorate signed 12.41%
<input type="checkbox"/> Wales Signatures 215,470		

CONSTITUENCY	MP	SIGNATURES +	% ELECTORATE	% POPULATION
1 Bristol West <small>VOTED REMAIN 78.3%</small>	Thangam Debbonaire <small>LABOUR MAJORITY 37,239</small>	35,840 <small>8 PER HR</small>	38.5% <small>OF 93,302</small>	25.8% <small>OF 138,009</small>
2 Horsey and Wood Green <small>VOTED REMAIN 73%</small>	Catherine West <small>LABOUR MAJORITY 33,739</small>	31,041 <small>4 PER HR</small>	38.8% <small>OF 79,846</small>	24.1% <small>OF 138,666</small>
3 Brighton, Pavilion <small>VOTED REMAIN 74.1% (EST.)</small>	Caroline Lucas <small>GREEN PARTY MAJORITY 14,888</small>	28,086 <small>8 PER HR</small>	37.2% <small>OF 75,488</small>	25.0% <small>OF 112,196</small>
4 Hampstead and Kilburn <small>VOTED REMAIN 73.3% (EST.)</small>	Tulip Siddiq <small>LABOUR MAJORITY 30,211</small>	28,086 <small>8 PER HR</small>	Exceeded majority Exceeded GE2017 votes	

Signature numbers in bold are those that have exceeded the majority.

- Signatures data source: UKGov Petition Page
- Electorate taken from: 2017 General Election dataset
- EU Referendum results from House of Commons Library. The original results of the EU referendum were published across the UK by counting area and not constituency. In cases where the result had to be estimated for a constituency, the figure is marked with (est.). The HoC Library page explains more in detail how the numbers were worked out.

[@LiveFromBrexit](#)

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 37

37

Data

Introduction

Collection of accurate and reliable data is a prerequisite for informed risk assessment and risk management. Both scientists carrying out risk assessments and decision makers in Europe need up-to-date and comparable information across Member States on hazards found in the food chain and on food consumption.

When a new hazard is found in the food chain - for instance the recent cases of melamine found in various foods or dioxin contamination of pork - scientists must quickly assess who is exposed, through which foods and at what levels. This is in order to provide a rapid and reliable risk assessment and to help risk managers take appropriate action to protect consumers.

By collecting data at the EU level we can find out for example how often foods are contaminated with bacteria or chemicals and at what levels. This information, combined with reliable information on food consumption in the Member States, makes it possible for risk assessors to assess consumer exposure to a certain hazard both at the EU- and country-level. The assessments also show scenarios to make recommendations for the prevention, reduction, and monitoring of these hazards in the food chain.

Access to harmonised data supports risk managers in making informed decisions to protect and promote consumer health, for instance in assessing how dietary choices of salt compare with targets set for healthy diets. Such data can also be utilised in evaluating the effectiveness of EU actions and programmes aimed at reducing the occurrence of biological and chemical risks in food and in animal populations.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 38

38

https://data.europa.eu/data/datasets?locale=en

Overview of the components of the EDP data request

```

{
  "@context": "http://www.w3.org/2018/02/linkeddata",
  "@type": "Dataset",
  "name": "https://www.w3.org/2018/02/linkeddata",
  "description": "https://www.w3.org/2018/02/linkeddata",
  "url": "https://www.w3.org/2018/02/linkeddata",
  "image": "https://www.w3.org/2018/02/linkeddata"
}

```

Mapping of DCAT-AP to CDS and vice versa

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 39

39

**OPEN DATA,
BUT
NO SEMANTICS,
NO INHERENT INTEROPERABILITY,
NO MACHINE UNDERSTANDABILITY**

□ Link (Open) Data perceived solution.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 40

40

AN OBSERVATION FROM THE CREATOR OF THE WEB

Tim Berners-Lee and colleagues observed in the late 1990s that web pages were machine-processible but that they were not machine-understandable.*

Concepts encapsulated in Tim Berners-Lee's "Universal Resource Identifiers -- Axioms of Web Architecture" from 1996 at <https://www.w3.org/DesignIssues/Axioms.html>

**** The next wave**
The Semantic Web we aspire to makes substantial reuse of existing ontologies and data. It's a linked information space in which data is being enriched and added. It lets users engage in the sort of serendipitous reuse and discovery of related information that's been a hallmark of viral Web uptake.

* T. Berners-Lee, J. Hendler, and O. Lassila, 2001, "The Semantic Web," *Scientific American*, (May), pp. 34–43.


** N. Shadbolt, W. Hall, and T. Berners-Lee, "The semantic Web revisited," *Intelligent Systems, IEEE*, vol. 21, no. 1, pp. 96–101, Jan. 2006.

41



From @ 8m 50s into Tim Berners-Lee, 2009 (Feb), "The Next Web," TED2009, https://www.ted.com/talks/tim_berniers_lee_the_next_web#t-525235

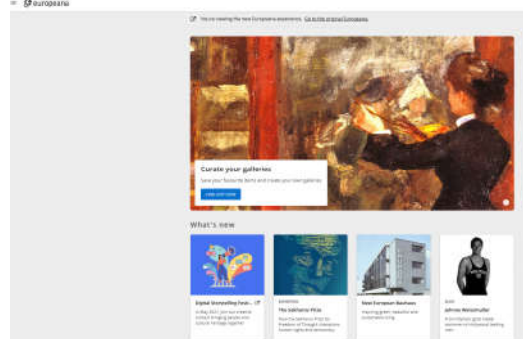
42



- ❑ Data Instantiated in a way that makes it interoperable
- ❑ Machine readable/executable
- ❑ Laid on the Resource Description Format (RDF) -- a data representation language/standard
- ❑ Provides a foundational component for Semantic Web
- ❑ Uses shared vocabularies and ontologies
- ❑ Remember Not all Linked Data are Open Data
- ❑ Not all Open Data are Linked Data
- ❑ --REFLECT BACK to the Sebastopol Principles of Open Data
- ❑ *"Linked Data refers to a set of best practices for publishing and interlinking structured data for access by both humans and machines via the use of the RDF (Resource Description Framework) family of standards for data interchange [RDF-CONCEPTS] and SPARQL for query. RDF and Linked Data are not synonyms. Linked Data however could not exist without the consistent underlying data model that we call RDF [RDF-CONCEPTS]. Understanding the basics of RDF is helpful in leveraging Linked Data."
- ❑ *<https://www.w3.org/TR/ld-bp/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)
43

43



<https://www.europeana.eu/en>

Reflecting on the European Data Model
Silvio Peroni¹, Francesca Tomas², and Fabio Vitali³
¹ Department of Computer Science, University of Bologna Italy
² Department of Computer Science, University of Bologna Italy
³ Department of Chemical Engineering and Applied Mathematics, University of Toronto, Canada

Abstract. We describe some issues arising while using Europeana, and analyze some features of the European Data Model (EDM), system from the perspective of the project. Some aspects of the historical work, already mostly done by the project, are presented. The project's main goal is to provide a single point of access to the metadata and the content of the objects in the European Digital Library (EDL) and the EDM. The metadata are the objects in the EDM, and the content is the objects in the EDL. The project's main goal is to provide a single point of access to the metadata and the content of the objects in the EDM. The metadata are the objects in the EDM, and the content is the objects in the EDL.

Keywords: EDM, Linked Data, RDF, DC, ODC, ODM, FRBR

1 Introduction
Europeana¹ is the European Digital Library, a distributed access point to Europe's multilingual cultural heritage in digital form. The main aim of the project is to collect metadata from a large number of providers, mainly cultural institutions, across Europe, and to make them available in a single point of access. The project's main goal is to provide a single point of access to the metadata and the content of the objects in the EDM. The metadata are the objects in the EDM, and the content is the objects in the EDL.

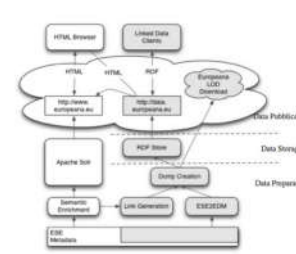


Fig. 3. LOD server technical architecture

Nicola Aloia, Cesare Concolato, and Carlo Meghini, 2013, "The European Linked Open Data Pilot Server" in Agosti, Mariastella, Floriana Esposito, Stefano Ferilli, and Nicola Ferro, Digital Libraries and Archives: 8th Italian Research Conference, IRCDL 2012, Bari, Italy, February 9-10, 2012, Revised Selected Papers. Springer Berlin Heidelberg, 2013. P 253 -

The European Linked Open Data Pilot Server
Nicola Aloia, Cesare Concolato, and Carlo Meghini
Institute of Informatics and Telematic Systems, University of Bari
nicola.aloia@iit.uniba.it, cesare.concolato@iit.uniba.it, carlo.meghini@iit.uniba.it


Abstract. The Linked Data is a set of principles and technologies providing a publishing paradigm for sharing and linking data. The Linked Data Pilot Server (LPS) is a web-based system that allows users to publish and link their data to the European Linked Open Data Pilot Server. The LPS is a web-based system that allows users to publish and link their data to the European Linked Open Data Pilot Server. The LPS is a web-based system that allows users to publish and link their data to the European Linked Open Data Pilot Server.

Keywords: Linked Data, Linked Open Data, Europeana

1 Introduction
The Linked Data is a set of principles and technologies providing a publishing paradigm for sharing and linking data. The Linked Data Pilot Server (LPS) is a web-based system that allows users to publish and link their data to the European Linked Open Data Pilot Server. The LPS is a web-based system that allows users to publish and link their data to the European Linked Open Data Pilot Server. The LPS is a web-based system that allows users to publish and link their data to the European Linked Open Data Pilot Server.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)
44

44



Linked Data

Data

<subject> <predicate> <object>

using W3C standards (e.g. RDF)

Open

Freely accessible using the Sebastopol Principles.

Linked

Links between "elements" from different data sets

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 45

45

Tim Berners-Lee
 Date: 2006-07-27, last change: \$Date: 2009/06/18 18:24:33 \$
 Status: personal view only. Editing status: imperfect but published.
 Up to Design Issues

Linked Data

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow.

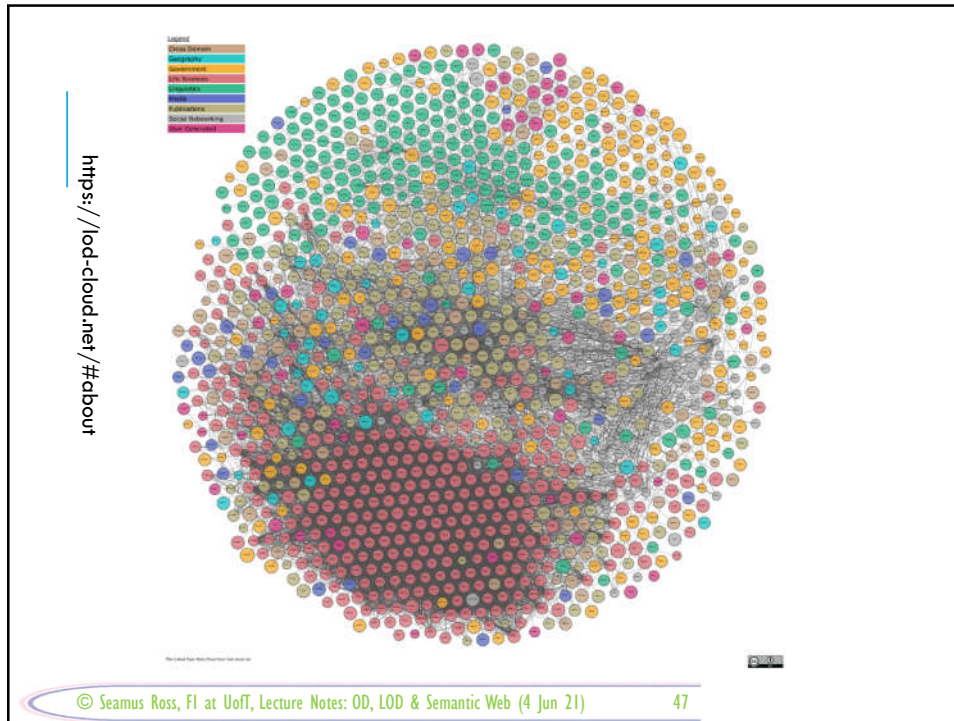
1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs, so that they can discover more things.



<https://www.w3.org/DesignIssues/LinkedData>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 46

46



47

Contributing to the Diagram

First, make sure that you publish data according to the [Linked Data principles](#). We interpret this as:

- There must be *resolvable* `http://` (or `https://`) URIs.
- They must resolve, with or without content negotiation, to *RDF* data in one of the popular RDF formats (RDFa, RDF/XML, Turtle, N-Triples).
- The dataset must contain at least 1000 triples. (Hence, your FOAF file most likely does not qualify.)
- The dataset must be connected via *RDF* links to a dataset that is already in the diagram. This means, either your dataset must use URIs from the other dataset, or vice versa. We arbitrarily require at least 50 links.
- Access of the entire dataset must be possible via *RDF* crawling, via an *RDF* dump, or via a SPARQL endpoint.

You may add a dataset by submitting it at this [form](#) [here](#). The process for adding datasets is still under development, please contact [John P. McCrae](#) for any issues.

<https://lod-cloud.net/#about>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 48

48

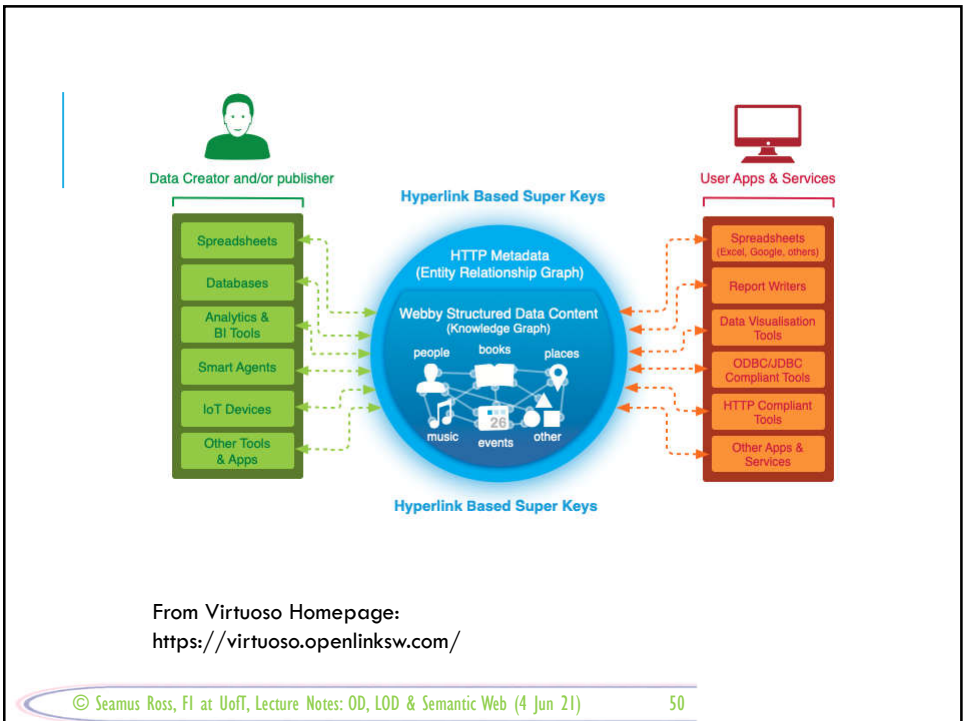
As Dominik Lukas, Claudia Engel and Camilla Mazzucato, 2018, "Towards a Living Archive: Making Multi Layered Research Data and Knowledge Generation Transparent", *Journal of Field Archaeology*, 43:sup1, S19-S30, DOI: [10.1080/00934690.2018.1516110](https://doi.org/10.1080/00934690.2018.1516110) explain in a discussion of the Çatalhöyük Research Project and its (living) archive. Quotes from pages S22 and S23.

While traditional long-term archives rely on the possible standardization of data types, 'Linked Open Data' approaches, like the OpenContext (Kansa 2012: 510) platform, in contrast, respond to the problem of identifying the informational structure of the datasets by assigning commonly understood attributes and allowing to link between similar information. In the 'Linked Open Data' approach a specific type of information—like a dataset describing a faunal bone—can be identified in the data cloud and can be accessed with a specific web-address. It is possible to link similar information residing in different archives by using the same publication approach. However, while this improves data interaction, the role of the data in a particular research context remains difficult to capture.

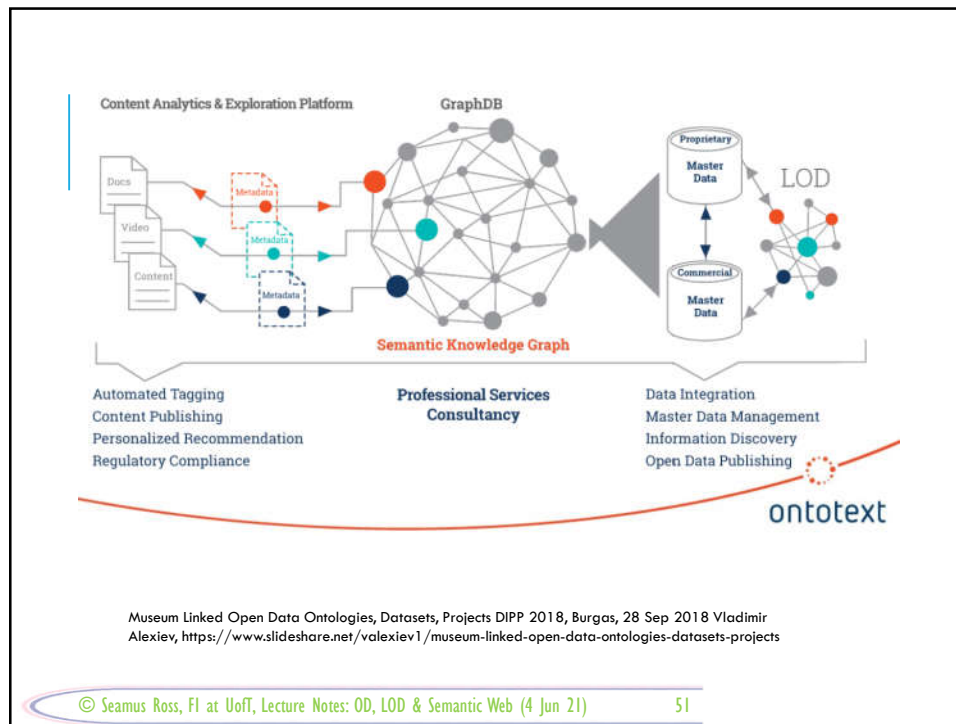
In this respect, the mentioned 'Linked Open Data' has to be understood as a merely pragmatic approach to the coding of knowledge, which has also gained popularity because of the difficulty to implement ontological conceptualizations. In contrast to complex ontologies like the CIDOC-CRM, other ontologies have proven to be more easily applicable (such as Open Annotation) (Isaksen et al. 2014: 198). OpenContext follows a similar pragmatic approach: ontologies are implemented depending on the specific context of the information (e.g. Encyclopedia of Life and Uberon for faunal remains) but are not integrated with a central reference model like the CIDOC-CRM (Faniel et al. 2013: 302). A first implementation for the Çatalhöyük data had been

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 49

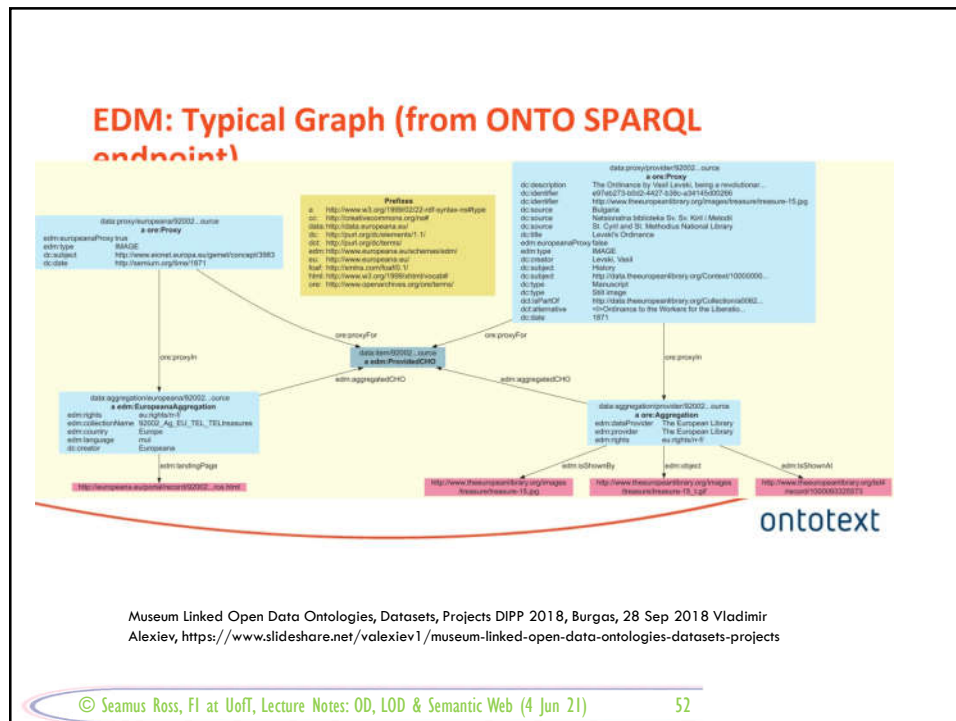
49



50



51



52

OA Example: Annotating SVG Part of Image (ResearchSpace)

Museum Linked Open Data Ontologies, Datasets, Projects DIPP 2018, Burgas, 28 Sep 2018 Vladimir Alexiev, <https://www.slideshare.net/valexiev1/museum-linked-open-data-ontologies-datasets-projects>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 53

53

Best Practices for Publishing Linked Data

W3C Working Group Note 09 January 2014

This version: <http://www.w3.org/TR/2014/NOTE-ld-bp-20140109/>
Latest published version: <http://www.w3.org/TR/ld-bp/>

Editors:
 Bernadette Hyland, 1 Round Stone, Inc.
 Shalun Hernandez, RUBSCOM
 Boris Vilazon, Intellecta, ISOCO, Intelligent Software Components S.A.

Copyright © 2014 W3C® (MIT, ERCIM, Keio, Beihai). All Rights Reserved. W3C (tm) and W3C (org) are trademarks of the World Wide Web Consortium. All other marks are the property of their respective owners.

Abstract

This document sets out a series of best practices designed to facilitate development and delivery of open government data as [Linked Open Data](#). [Linked Open Data](#) makes the World Wide Web into a global set of [Linked Data](#) from multiple sources of data and combine it without the need for a single common schema but all data shares. Prior to international data exchange standards for data on the Web, it was time to government data is published on the Web, best practices are evolving too. The goal of this document is to compile the most relevant data management practices for the publication and use of high quality data.

Status of This Document

This section describes the status of this document at the time of its publication. Other documents may supersede this document. A list of current W3C publications and the latest revision of this technical report is available at <http://www.w3.org/2002/02/tracker/>.

This document was published by the [Government Linked Data Working Group](#) as a First Public Working Group Note. If you wish to make comments regarding this document, please send them to public-ld-bp@w3.org; ending, the group might not officially respond to comments, but individual members may. As usual, comments are [publicly archived](#), available to both readers and any group updating this document in the future. Publication as a Working Group Note does not imply endorsement by the W3C Membership. This is a draft document and may be updated, replaced or obsoleted by other documents at any time. It is inappropriate to cite this document as a standard.

This document was produced by a group operating under the [5 February 2004 W3C Patent Policy](#). W3C maintains a [public list of any patent disclosures](#) made in connection with the deliverables of the group; that which the individual believes contains [Essential Claims](#) must disclose the information in accordance with [section 6 of the W3C Patent Policy](#).

Table of Contents

1. Prepare Stakeholders
2. Select a Dataset
3. Model the Data
4. Specify an Appropriate License
5. The Role of "Good URIs" for Linked Data
6. Standard Vocabularies
7. Convert Data to Linked Data
8. Provide Machine Access to Data
9. Announce to the Public
10. Social Contract of a Linked Data Publisher
- A. Acknowledgments
- B. References
- B.1 Informative references

<https://www.w3.org/TR/ld-bp/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 54

54

Summary of Best Practices

The following best practices are discussed in this document and listed here for convenience.

<p>STEP #1 PREPARE STAKEHOLDERS: Prepare stakeholders by explaining the process of creating and maintaining Linked Open Data.</p>
<p>STEP #2 SELECT A DATASET: Select a dataset that provides benefit to others for reuse.</p>
<p>STEP #3 MODEL THE DATA: Modeling Linked Data involves representing data objects and how they are related in an application-independent way.</p>
<p>STEP #4 SPECIFY AN APPROPRIATE LICENSE: Specify an appropriate open data license. Data reuse is more likely to occur when there is a clear statement about the origin, ownership and terms related to the use of the published data.</p>
<p>STEP #5 GOOD URIs FOR LINKED DATA: The core of Linked Data is a well-considered URI naming strategy and implementation plan, based on HTTP URIs. Consideration for naming objects, multilingual support, data change over time and persistence strategy are the building blocks for useful Linked Data.</p>
<p>STEP #6 USE STANDARD VOCABULARIES: Describe objects with previously defined vocabularies whenever possible. Extend standard vocabularies where necessary, and create vocabularies (only when required) that follow best practices whenever possible.</p>
<p>STEP #7 CONVERT DATA: Convert data to a Linked Data representation. This is typically done by script or other automated processes.</p>
<p>STEP #8 PROVIDE MACHINE ACCESS TO DATA: Provide various ways for search engines and other automated processes to access data using standard Web mechanisms.</p>
<p>STEP #9 ANNOUNCE NEW DATA SETS: Remember to announce new data sets on an authoritative domain. Importantly, remember that as a Linked Open Data publisher, an implicit social contract is in effect.</p>
<p>STEP #10 RECOGNIZE THE SOCIAL CONTRACT: Recognize your responsibility in maintaining data once it is published. Ensure that the dataset(s) remain available where your organization says it will be and is maintained over time.</p>

<https://www.w3.org/TR/ld-bp/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 55

55

ALL DATA ARE
NOT WORTH
MAKING
OPEN OR
LINKING

Data released but not usable

- For example,
 - Without adequate metadata
 - Inadequate paradata
 - In complex formats
 - Without data dictionary support
 - Occasionally it is released in print, or in pdf format
- The business or research case for making these data open or linking them is limited at best.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 56

56

SO WHAT
ABOUT DATA
QUALITY AND
OPEN DATA

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

57

57

ALL OPEN
PUBLIC &
RESEARCH
DATA ARE NOT
OPEN
ALL DATA IS
NOT AMENABLE
TO LINKING

- Conditions of Use may be set on the data:
- Complexity of the dataset in terms of the number of records and variables;
- Is the Data about living people;
- Is the Data anonymized;
- Is it raw and granular or is it aggregated:
- Quantitative or qualitative;
- how often the dataset is updated or replaced;
- How is the data generated: part of a public activity or
- What kind of content dataset;
- the electronic or non-electronic format of the dataset;
- the ways in which the public sector information dataset is distributed;
- the cost of generating/collecting/maintaining/updating the public sector information dataset.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

58

58

QUALITY DIMENSIONS

Table 2. Notable data quality dimensions

Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

From: Yair Wand and Richard. Y. Wang, 1996, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, 39(11), p., 92. From Wang, R.Y., Storey, V.C., and Firth, C.P. A framework for analysis of data quality research. *IEEE Trans. on Knowl. Data Eng.* 7, 4 (1995), pp. 623–640.

59

DESIGN AND PRODUCTION

THE **QUALITY** of data depends on the **DESIGN**
 AND **PRODUCTION PROCESSES** involved in
GENERATING THE DATA. To design for better quality,
 it is necessary first to understand **WHAT QUALITY**
MEANS and **HOW IT IS**
MEASURED.

From: Yair Wand and Richard. Y. Wang, 1996, "Anchoring data quality dimensions in ontological foundations," *Communications of the ACM*, 39(11), p., 89.

60

DATA
CONSUMERS
AND QUALITY

From: Richard Y. Wang and Diane M Strong, 1996. "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, 12(4), pp., 5-33:

Accuracy and Precision: "inaccuracy implies that information system represents a real-world state different from the one that should have been represented."

reliability indicates whether the data can be counted on to convey the right information-- can be viewed as correctness of data"

timelines refers only to the delay between a change of the real-world state and the resulting modification of the information system state."

completeness is the ability of an information system to represent every meaningful state of the represented real world system."

Consistency in the "data values" as representations of real-world data values.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)
61

61

DATA
CONSUMERS
AND QUALITY

From: Richard Y. Wang and Diane M Strong, 1996. "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, 12(4), pp., 5-33:

"Data consumers have a much broader data quality conceptualization than IS Professionals"

"Fitness for Use"

Data Quality = "data that are fit for use by data consumers"

Examined approaches: "intuitive," "theoretical" and "empirical"

Focus on constructing "a comprehensive framework of data quality from data consumers' perspectives"

Adopted a multi-stage research survey approach

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)
62

62

APPROACHES TO EVALUATING DATA QUALITY

Category	Dimension	Definition: the extent to which ...
Intrinsic	Believability	data are accepted or regarded as true, real and credible
	Accuracy	data are correct, reliable and certified free of error
	Objectivity	data are unbiased and impartial
	Reputation	data are trusted or highly regarded in terms of their source and content
Contextual	Value-added	data are beneficial and provide advantages for their use
	Relevancy	data are applicable and useful for the task at hand
	Timeliness	the age of the data is appropriate for the task at hand
	Completeness	data are of sufficient depth, breadth, and scope for the task at hand
	Appropriate amount of data	the quantity or volume of available data is appropriate
Representational	Interpretability	data are in appropriate language and unit and the data definitions are clear
	Ease of understanding	data are clear without ambiguity and easily comprehended
	Representational consistency	data are always presented in the same format and are compatible with the previous data
	Concise representation	data are compactly represented without being overwhelmed
Accessibility	Accessibility	data are available or easily and quickly retrieved
	Access security	access to data can be restricted and hence kept secure

Fig. 2.10. Dimensions proposed in the empirical approach

From: Carlo Batini and Monica Scannapieca. 2006, *Data Quality: concepts, methodologies and techniques*, New York: Springer, p. 38, but derived from Wang and Strong 1996 :

63

BUT WHAT IS THE BORDER OF THE QUESTION OF QUALITY FOR OPEN AND LINKED DATA—THE QUESTION OF PROVENANCE AS SEEN THROUGH: PARADATA & METADATA

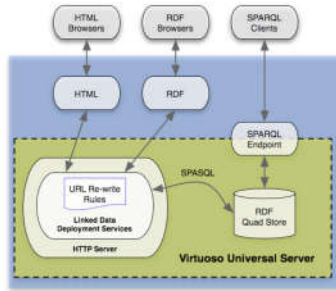
- Provenance of Open Data a poorly explored topic
- Provenance of Linked Data an equally inadequately explored topic
- Quality Parameters do not address issues of Paradata or Metadata.
- Metadata should contain Paradata.
- Paradata provides evidence as to the “processes” related to data collection/construction
- And in the case of linked data evidence as to the processes by which it was linked
- Paradata could reflect human, machine or a combination of both processes.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 64

64

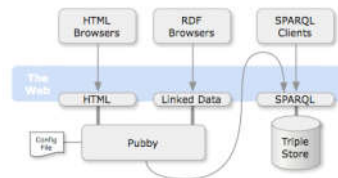
DBpedia is a crowd-sourced community effort to extract structured content from the information created in various Wikimedia projects. This structured information resembles an **open knowledge graph (OKG)** which is available for everyone on the Web. A knowledge graph is a special kind of database which stores knowledge in a machine-readable form and provides a means for information to be collected, organised, shared, searched and utilised. Google uses a similar approach to create those

Illustration of Current DBpedia Data Provision Architecture



<https://www.dbpedia.org/about/>

Illustration of Deprecated Architecture



W3C Working Group Note

W3C

Linked Data Glossary

W3C Working Group Note 27 June 2013

This version:
<http://www.w3.org/TR/2013/NOTE-lid-glossary-20130627/>

Latest published version:
<http://www.w3.org/TR/2013/lid-glossary/>

Previous version:

Editors:
 Deropete Inghel, [3 Round Stones](#)
 Charles Bessany, [EUSC2/M](#)
 Michael Pendleton, [US Environmental Protection Agency](#)
 Bipin Sivastava, [IBM](#)

Copyright © 2013 W3C® MIT, ERCIM, Keio. All Rights Reserved. W3C liability disclaimers and document use rules apply.

Abstract

This document is a glossary of terms defined and used to describe Linked Data, and its associated vocabularies and [Best Practices](#). This document published by the [W3C Government Linked Data Working Group](#) as a Working Group Note and the general public better understand publishing structured data using [Linked Data Principles](#).

Status of This Document

This section describes the status of this document at the time of its publication. Other documents may supersede this document. A list of current W3C publications and the latest revision of this technical report can be found in the [W3C Publications](#) page.

This document was published by the [Government Linked Data Working Group](#) as a Working Group Note. If you wish to make comments regarding this document, please send them to public-gld-comments@w3.org ([subscribe](#), [unsubscribe](#), [archive](#)).

Publication as a Working Group Note does not imply endorsement by the W3C Membership. This is a draft document and may be updated, replaced or obsolete by other documents at any time. It is inappropriate to cite this document as a standard.

This document was produced by a group operating under the [5 February 2004 W3C Patent Policy](#). W3C maintains a [public list of any patent disclosures](#) made in connection with the deliverables of the group; that page also includes information about the [dissemination of any W3C copyright](#) and the [list of any W3C copyright](#).

Table of Contents

1. 5 Star Linked Open Data
2. Apache License
3. API
4. CC-BY-SA License
5. Closed World
6. Connection
7. Conneg
8. Content Negotiation
9. Controlled Vocabulary
10. Comma Separated Values (CSV)

<https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html#linked-open-data>
 Accessed 17 May 2021.

FEATURED COLLECTION
 The archaeological site of the Roman fort of *Hadrian's Wall*, with over 17 miles of excavated masonry, Roman fortifications and buildings with over 100,000 artefacts and information from the site.

NEWS
 From the 1st of April, the set of the website will be updated at the current site.

SEARCH
 The AOS database is a central point of digital heritage that has a lot to offer you. We have a lot to offer you. We have a lot to offer you. We have a lot to offer you.

NEWS
 Following our Heritage Futures Programme for Digital Heritage.

<https://archaeologydataservice.ac.uk/>

2009 RESULTS (PAGE 2 OF 104)

- Site Data From an Archaeological Excavation at *Hadrian's Wall* (2009-2010)
- The Site of the Former *St. Luke's College*, Topham Road, Exeter 2009-2010
- Photography of *Excavations* 1997-2006
- Images from a Historic Building Recording Survey of *Blacksmiths Shop* (2009)
- Images and records from an evaluation and excavation of *Fibra Barn Green*, Merton, Devon 2014
- Images from a Historic Building Recording at *Snibston Hall Farm*, Snibston, Malton 2014
- Stratigraphy of the *Neolithic Architecture Group* (2011)
- Site and Post-Excavation Data from an Archaeological Evaluation at *Madhouse* (2011)
- Site and Post-Excavation Data from an Archaeological Evaluation at *Long Gaf Lands* (2011)
- Site and Post-Excavation Data from an Archaeological Evaluation at *Wantage* (2011)
- Images from an Archaeological Watching Brief at *South of Hanson Drive*, Reading, Berkshire 2013
- Data from a *Geophysical Survey* of *Chapel Lane*, (2011)
- Images from an Archaeological Trial Trench Evaluation on *Esboart Road*, Gloucester 2012
- Site Images and Data from an Archaeological Evaluation at *Land off Peppercorn*, Essex 2012
- Images from an Archaeological Watching Brief at the *Church of St. Thomas* of *Canterbury* (2012)
- Images from a Historic Building Recording at *New Trees Farm*, New Trees Lane, Bursledon, South Yorkshire 2012
- Site Images from an Archaeological Excavation on *Land East of Habbott Road*, Kirby Cotes, Essex 2012
- Site Images from an Archaeological Watching Brief at *St. Peter's Church*, West Berkshire 2012
- Images from a Historic Building Recording at *51 Starling Road*, Haverhill 2012
- Site Images from an Archaeological Watching Brief on *Land at Bitcher Park Hill*, Teasdale, Devon 2012-2019

<https://archaeologydataservice.ac.uk/archive/archives.xhtml>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 67

67

Explore the data set to establish

- The kind of data,
- Its nature,
- Its quality,
- Its architecture,
- What aspects of data to model such as concepts and relationships,
- Rights Issues.

Not a complex example....

<https://archaeologydataservice.ac.uk/archive/archives.xhtml>

Birte Bruggmann (2004) Glass Beads from Anglo-Saxon Graves [data-set].
 York: Archaeology Data Service [distributor]
<https://doi.org/10.5284/1000232>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 68

68

Attribution 4.0 International (CC BY 4.0)

You are free to:

- Share** — copy and redistribute the material in any medium or format.
- Adapt** — remix, transform, and build upon the material for any purpose, even commercially.

Under the following terms:

- Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation.

- Creative Commons Zero is a good model (CC0): no copyright, public domain
- Encourage use and re-use, including for commercial purposes
- Require attribution to primary and secondary sources
- Main usual exclusions: no endorsement, association, warranty, liability
- Need special terms for “raw” data collected under confidentiality agreements

© Seamus Ross, FI at Uoff, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 69

69

Title	Image	Description
Attribution CC BY		This license lets others distribute, remix, tweak, and build upon the work, even commercially, as long as they credit the creator for the original creation. This is the most flexible and accommodating of the available Creative Commons licenses. Recommended for maximum dissemination and use of licensed materials.
Attribution-NonCommercial CC BY-NC		This license lets others remix, tweak, and build upon the work non-commercially, as long as they credit the creator and license their new creations under the identical terms. This license is often compared to “copyleft” free and open source software licenses. All new works based on yours will carry the same license, so any derivatives will also allow commercial use. This is the license used by Wikipedia, and is recommended for materials that would benefit from incorporating content from Wikipedia and similarly licensed projects.
Attribution-NonCommercial-ShareAlike CC BY-NC-SA		This license lets others remix, tweak, and build upon your work non-commercially, as long as they credit the creator and license their new creations under the identical terms.
Attribution-NonCommercial-NoDerivs CC BY-NC-ND		This license is the most restrictive of the six main licenses, only allowing others to download your works and share them with others as long as they credit the creator, but they can't change them in any way or use them commercially.

License Text and Icons by Creative Commons Organization and is licensed under a Creative Commons Attribution 4.0 License

Seven regularly used licenses [\[edit \]](#)

Icon	Description	Acronym	Allows Remix culture	Allows commercial use	Allows Free Cultural Works	Meets 'Open Definition'
	Freeing content globally without restrictions	CC0	Yes	Yes	Yes	Yes
	Attribution alone	BY	Yes	Yes	Yes	Yes
	Attribution + ShareAlike	BY-SA	Yes	Yes	Yes	Yes
	Attribution + Noncommercial	BY-NC	Yes	No	No	No
	Attribution + NoDerivatives	BY-ND	No	Yes	No	No
	Attribution + Noncommercial + ShareAlike	BY-NC-SA	Yes	No	No	No
	Attribution + Noncommercial + NoDerivatives	BY-NC-ND	No	No	No	No

Image from https://en.wikipedia.org/wiki/Creative_Commons_license

70

YALE CENTER FOR BRITISH ART

OPEN DATA AND DATA SERVICES TERMS OF USE

These Terms of Use apply to any use or user of the data and data services made available by the Yale Center for British Art. Yale University's standard Terms of Use also apply. If you do not agree to these Terms in full, do not access or use the Center's data or data services.

Privacy

The Center's servers collect information on the use of its sites and services, including but not limited to visiting IP addresses, referring URLs, and visited URLs. The Center's sites and services may use Google Analytics or comparable tools to collect statistics on usage, and may not collect. Please see Yale University's Privacy Policy for further information.

Conditions

The Center strives to make its data and data services (including but not limited to APIs, feeds and Linked Data Services) openly available in order to provide the maximum of access to, and the best use of its resources. As for all the Center's open-access digital and data services, we need feedback from anyone without seeking explicit permission from the Center. However, in order to ensure the availability and usefulness of the Center's data and data services, the following conditions apply:

- Openness:** We encourage free and open use and reuse of the Center's data. Therefore, you may not distribute or sublicense the Center's data under more restrictive terms than those set forth here. For reuse or distribution, include a link to these terms.
- Attribution:** Acknowledge the Yale Center for British Art by including the statement "Data Source: Yale Center for British Art" and linking to the Center's Collections Data Sharing technology information page as illustrated. When using linked data, the URL for the object being referenced (example linked) should be provided. However, you may NOT imply that the Center and/or Yale University endorse you or your use of the Center's data and data services in any way.
- Transparency and Courtesy:** Be clear about who you are and how your services will interact with our data services. Where you are using a non-browser user agent, your User-Agent header should include details of your software application and your contact details including an e-mail address and URL. Where you make, use, or distribute software that makes requests of the Center's services, you should ensure that your software will not perform excessive requests.
- Updates:** You are responsible for checking regularly for updates to our data and data services, ensuring that your applications continue to function appropriately with our data services, and using the most up-to-date data provided.
- Legal:** It is your responsibility to ensure that your use is compliant with legal requirements in any applicable jurisdictions.

Users further acknowledge the following:

- Digital Resources:** Data and data services may include links to digital resources, including image files. The availability of a link to a digital resource does not imply or convey permission to download, distribute, or use the link or image file, (however, digital resources of works in the public domain with no other known restrictions, as identified in the public domain with no other known restrictions, as identified in the data, may be used freely. See [Using Images](#) for more information.)
- Excessive Use:** The Center reserves the right to limit or block any user or client for excessive use. The definition of excessive includes, but is not limited to, any use that may have a detrimental effect on the service, whether intentional or non-intentional, and is entirely at the Center's discretion.
- Availability:** The Center reserves the right to revise these Terms at any time, as well as to modify or terminate any products, services, software, applications, or features on the site at any time and without notice.

THE DATA AND DATA SERVICES ARE PROVIDED "AS IS," WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NON-INFRINGEMENT. THE YALE CENTER FOR BRITISH ART IS NOT LIABLE FOR ANY ERRORS OR OMISSIONS, AND SHALL NOT BE LIABLE FOR ANY LOSS, INJURY, OR DAMAGE OF ANY KIND CAUSED BY ITS USE.

YALE CENTER FOR BRITISH ART

<https://britishart.yale.edu/open-data-and-data-services-terms-use>

71

RDFa (Resource Description Framework in Attributes)
--a mechanism "...to markup existing human-readable Web page content to express machine-readable data"

W3C

RDFa 1.1 Primer - Third Edition
Rich Structured Data Markup for Web Documents
W3C Working Group Note 17 March 2015

This version:
<http://www.w3.org/TR/2015/NOTE-rdfa-primer-20150317/>

Latest published version:
<http://www.w3.org/TR/rdfa-primer/>

Latest editor's draft:
<http://www.w3.org/2010/02/rdfa/sources/rdfa-primer/Overview.src.html>

Previous version:
<http://www.w3.org/TR/2013/NOTE-rdfa-primer-20130822/>

Editors:
Ivan Herman ivan@w3.org
Ben Adida Creative Commons: ben@adida.net
Manu Sporny Digital Bazaar: msporny@digitalbazaar.com
Mark Birbeck webBackPlane.com: mark.birbeck@webbackplane.com

Please check the [errata](#) for any errors or issues reported since publication.

This document is also available in this non-normative format: [diff to previous version](#)

[Copyright](#) © 2010-2015 [W3C](#)[®] [MIT](#) [ERCIM](#) [Keio](#) [Bell Labs](#); [W3C liability](#) [trademarks](#) [and](#) [document use rules](#) apply.

Abstract

The last couple of years have witnessed a fascinating evolution: while the Web was initially built predominantly for human consumption, web content is increasingly consumed by machines which expect some amount of structured data. Sites have started to identify a page's title, content type, and preview image to provide appropriate information in a user's newsfeed when she clicks the "Like" button. Search engines have started to provide richer search results by extracting fine-grained structured details from the Web pages they crawl. In turn, web publishers are producing increasing amounts of structured data within their Web content to improve their standing with search engines.

A key enabling technology behind these developments is the ability to add structured data to HTML pages directly: RDFa (Resource Description Framework in Attributes) is a technique that allows just that: it provides a set of markup attributes to augment the visual information on the Web with machine-readable hints. In this Primer, we show how to express data using RDFa in HTML, and in particular how to mark up existing human-readable Web page content to express machine-readable data.

This document provides only a Primer to RDFa 1.1. The complete specification of RDFa, with further examples, can be found in the RDFa 1.1 Core [\[rdfa-core\]](#), RDFa Lite [\[rdfa-lite\]](#), XHTML+RDFa 1.1 [\[xhtml-rdfa\]](#), and the HTML5+RDFa 1.1 [\[html-rdfa\]](#) specifications.

<https://www.w3.org/TR/rdfa-primer/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 72

72

RDF-Based Semantics

Document title: **RDF-Based Semantics (Open SPARQL)**

Editor: **Michael Schuster** (FZI Research Center for Information Technology)

Contributors (alphabetical order): **Jeremy Carroll**, **MP Jones** at **TheOpenLink**, **John Heintzel**, **W3C/OWL**, **Peter F. Patel-Schreiber**, **Resource Communications**

Abstract

The OWL 2 Web Ontology Language (formally OWL 2) is an ontology language for the Semantic Web with formally defined modeling, OWL 2 ontologies provide classes, properties, individuals, and data in **Formalisms** are generally recognized as RDF documents. The OWL 2 Document Overview describes the overall state of OWL 2, and should be read before other OWL 2 documents.

This document defines the RDF-compatible model-theoretic semantics of OWL 2.

Status of this Document

Copyright © 2008-2020 W3C® (MIT®, ERCIM®, Keio®, All Rights Reserved. W3C, liability, trademarks and document use if rules apply).

Contents (tree)

- 1 Introduction (Informative)
- 2 Ontologies
 - 2.1 Syntax
 - 2.2 Content of Ontologies (Informative)
- 3 Vocabulary
 - 3.1 Standard Prefixes
 - 3.2 Vocabulary Names
 - 3.3 Qualified Names
 - 3.4 Facet Names
- 4 Interpretations
 - 4.1 Change Maps
 - 4.2 Vocabulary Interpretations
 - 4.3 Satisfaction, Consistency and Entailment
 - 4.4 Paths of the Urniverse
 - 4.5 Class Expressions
- 5 Semantic Conditions
 - 5.1 Semantic Conditions for the Paths of the Urniverse
 - 5.2 Semantic Conditions for the Vocabulary Classes
 - 5.3 Semantic Conditions for the Vocabulary Properties
 - 5.4 Semantic Conditions for Simple Constraints
 - 5.5 Semantic Conditions for Expressions

Web Ontology Language (OWL), https://www.w3.org/2007/OWL/wiki/RDF-Based_Semantics

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 73

73

Terse RDF Triple Language

RDF 1.1 Turtle

Terse RDF Triple Language

W3C Recommendation 25 February 2014

This version: <http://www.w3.org/TR/2014/REC-turtle-20140225/>

Latest published version: <http://www.w3.org/TR/turtle/>

Test suite: <http://www.w3.org/TR/2014/NOTE-rdf11-testcases-20140225/>

Implementation report: <http://www.w3.org/2013/TurtleReports/index.html>

Previous version: <http://www.w3.org/TR/2014/PR-turtle-20140225/>

Editors: [Eric Prud'hommeaux](#), [W3C](#), [Gavin Carothers](#), [Lex Machina, Inc.](#)

Authors: [David Beckett](#), [Tim Berners-Lee](#), [W3C](#), [Eric Prud'hommeaux](#), [W3C](#), [Gavin Carothers](#), [Lex Machina, Inc.](#)

Please check the [errata](#) for any errors or issues reported since publication.

The English version of this specification is the only normative version. Non-normative [translations](#) may also be available.

Copyright © 2008-2014 W3C® (MIT, ERCIM, Keio, Beihang). All Rights Reserved. W3C, liability, trademarks and document use if rules apply.

Abstract

The Resource Description Framework (RDF) is a general-purpose language for representing information in the Web. This document defines a textual syntax for RDF called Turtle that allows an RDF graph to be completely written in a compact and natural text form, with abbreviations for common usage patterns and datatypes. Turtle provides levels of compatibility with the N-Triples [N-TRIPLES] format as well as the triple pattern syntax of the SPARQL W3C Recommendation.

Status of This Document

This section describes the status of this document at the time of its publication. Other documents may supersede this document. A list of current W3C publications and the latest revision of this technical report can be found in the [W3C Technical Reports Index](http://www.w3.org/TR/) at <http://www.w3.org/TR/>.

This document is a part of the RDF 1.1 document suite. The document defines Turtle, the Terse RDF Triple Language, a concrete syntax for RDF [RDF11-CONCEPTS].

This document was published by the [RDF Working Group](#) as a Recommendation. If you wish to make comments regarding this document, please send them to public-ietf-comments@w3.org ([subscribe](#), [archives](#)). All comments are welcome.

<https://www.w3.org/TR/turtle/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 74

74

**PRIVACY: LINKED
(?OPEN) DATA**

- ❑ Release of data as open and for linking that can be routed back to particular individuals is a risk to the individuals and society
- ❑ De-identification of data is a way to protect privacy
- ❑ Re-identification is a process to use data to reconstruct individuals in de-identified data
- ❑ Application of de-identification techniques and processes
- ❑ Re-identification risk management procedures using a framework
- ❑ Do it and then test that it has been done right

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 75

75

**LINKED OPEN DATA
AND DE-
IDENTIFICATION**

‘...techniques ... intended to remove identifying information from a dataset while retaining some utility in the remaining data’

Simson L. Garfinkel, 2015, De-Identification of Personal Information, NISTIR 8053, p. 6,
<http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 76

76

DATA CONTAIN DIFFERENT CLASSES OF IDENTIFIERS --

- **Direct identifiers**
 - Names, SIN, email, home address, hospital id number,
- **Quasi-identifiers (indirect identifiers)**
 - Date of birth, postcode, gender
- **Biometric**
 - Physiological (e.g., face, iris, ear, fingerprint)
 - Behaviourial (e.g., voice, gait, gesture, lip-motion)
- **Soft biometric**
 - Height, weight, eye colour, silhouette, age, gender race, moles, tattoos, birthmarks, scars
- **Contextual identifiers**
 - Text, Speech, dress, hairstyle, socio-political or environmental

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 77

77

Received: 18 December 2020 | Revised: 27 April 2021 | Accepted: 28 April 2021
DOI: 10.1002/aii2.23

LETTER WILEY

Heritage connector: A machine learning framework for building linked open data from museum collections

Kalyan Dutia | John Stack

Science Museum Group, London, UK

Correspondence
Kalyan Dutia and John Stack, Science Museum Group, London, UK.
Email: kalyan.dutia@psm.ac.uk; john.stack@sciencemuseum.ac.uk

Funding information
Arts and Humanities Research Council

Abstract

As with almost all data, museum collection catalogues are largely unstructured, variable in consistency and overwhelmingly composed of thin records. The form of these catalogues means that the potential for new forms of research, access and scholarly enquiry that range across multiple collections and related datasets remains dormant. In the project *Heritage Connector: Transforming text into data to extract meaning and make connections*, we are applying a battery of digital techniques to connect similar, identical and related objects within and across collections and other publications. In this article, we describe a framework to create a Linked Open Data knowledge graph from digital museum catalogues, perform record linkage to Wikidata, and add new entities to this graph from textual catalogue record descriptions (information retrieval). We focus on the use of machine learning to create these links at scale with a small amount of labelled data, and models which are small enough to run inference on datasets the size of museum collections on a mid-range laptop or a small cloud virtual machine. Our method for record linkage against Wikidata achieves 85%+ precision with the Science Museum Group (SMG) collection, and our method for information retrieval is shown to improve NER performance compared with pretrained models on the SMG collection with no labelled training data. We publish open-source software providing tools to perform these tasks.

Kalyan Dutia and John Stack, 2021, "Heritage Connector: A machine learning framework for building linked open data from museum collections," *Applied AI Letters*, 3 May 2021, <https://doi.org/10.1002/aii2.23>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 78

78

RESOURCE DESCRIPTION FRAMEWORK (HEREAFTER RDF)

- ❑ RDF a mechanism to make "statements"
- ❑ These RDF Statements consist of
 - ❑ subject predicate/property/relation object
- ❑ Perhaps it is easiest to visualize this as nodes and links
- ❑ "Jean-Claude Gardin is a French Archaeologist"

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 79

79

RESOURCE DESCRIPTION FRAMEWORK (HEREAFTER RDF)

Jean-Claude GARDIN, « Les applications de la mécanographie dans la documentation archéologique », Bulletin des bibliothèques de France (BBF), 1960, n° 1-3, p. 5-16 (https://bibliothebnf.fr/consulter/bbf-1960-01-0005-001)

- ❑ RDF Representations are Adaptable, Extensible, and Architecturally Flexible
- ❑ Adding Nodes and Links Feasible
- ❑ Predicates have embedded semantic meaning
- ❑ Jean-Claude Gardin is a French Archaeologist
- ❑ is different than
 - ❑ Jean-Claude Gardin founded the Centre Mécanographique de Documentation Archéologique
 - ❑ Centre Mécanographique de Documentation Archéologique was part of French National Center for Scientific Research

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 80

80

UNIFORM RESOURCE IDENTIFIERS (URIs) EXPRESS RDF STATEMENTS

- ❑ Establishing Uniform Resource Identifiers (URI)
- ❑ Sustainable and Resilient URIs
- ❑ Consider anticipating change by implementing versioning
 - ❑ Addresses problem of change/evolution in Vocabularies
 - ❑ Reflects richer understanding of the world
- ❑ Perhaps even define your URIs to reflect physicality, data, and virtuality
- ❑ Distinction between real world entities and Web
- ❑ Ensure URIs can be resolved
 - ❑ Our URI for Jean-Claude Gardin when resolved provides information in form of properties, classes, and provenance.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

81

81

Dutch Ships and Sailors

Home Project Dutch maritime historical database Semantic web

Dutch Ships and Sailors provides an infrastructure for maritime historical datasets. Using connecting data through semantic web technology, it brings together datasets related to recruitment and shipping in the East India trade during 17th century and in the shipping of the northern provinces of the Netherlands (mainly 18th century). For the northern provinces, the database contains data on the personnel recruited, the ships, and other variables (Observation: Nederlandse Handelmaatschappij). For the VOC, the database includes certain data on the recruitment of personnel in the Dutch Republic (HCC: Overzicht van de reizen van schepen naar en van de Oost-Indische Compagnie, the making of ships and crew composition in Asian waters (Geneva: Zeevaartmuseum).

Dutch Ships and Sailors was created in a Dutch project. It is hosted by Huggens ING in collaboration with VU University Amsterdam, the International Institute of Social History and Universiteitsmuseum Amsterdam.

Check out the [overview of the infrastructure](#).

More information on the Dutch Ships and Sailors project can be found [here](#).

An inventory of Dutch Maritime Historical Datasets can be found [here](#).

As part of a follow-up to the DASH-funded project *European Maritime Data* has made accessible to all of photographs of the original archival records of the Nederlandse Handelmaatschappij Database and enriched them with links to the Dutch Ships and Sailors.

Huggens ING

<https://dutchshipsandsailors.nl/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

82

82

Places Admin Repository Query Help Login

Dutch Ships and Sailors (DSS) Semantic Layer

On this page, you find a live version of the Semantic Layer made for the [Dutch Ships and Sailors](#) project.



Viewing the data

You are now looking at the ClioPatris user interface to the DSS data, which allows you to browse the data. One way to start is by [browsing the RDF Schemas](#) loaded. By selecting a graph, you can then click through to see basic statistics or download the graphs in multiple formats. You can also use the search field in the top right corner to search for resources with matching labels. The search field has semantic auto-completion to allow you to select specific search terms.

For more complex queries, we have a number of interfaces to the SPARQL endpoint (For example the great [YASGUI](#)). On the [DSS Queries](#) page, we have listed a number of interesting example queries. Lastly, the provenance of the data can be visualized using [proteowl](#).

To get a quick overview of the project and this demonstrator you can watch the five-minute video below:

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 83

83

Places Admin Repository Query Help Login

Load local file
Load from HTTP

Named graphs list

URI	Triples	Modified	Persistency
http://sparql.org/call/RemoveTriples/vocvoc:opt.nl.gz	19,104,514	Fri Jun 30 17:00:11 2015	⊙
http://sparql.org/call/ClearRepository/vocvoc:asd.nl.gz	2,328,820	Fri Jun 30 17:00:13 2015	⊙
http://sparql.org/collections/ds/vocvoc:mbh_data.nl.gz	1,296,641	Fri Jun 30 17:00:10 2015	⊙
http://sparql.org/collections/ds/vocvoc:voocopy_2_dss.nl.gz	1,128,416	Fri Jun 30 17:00:10 2015	⊙
http://sparql.org/collections/ds/vocvoc:bez.nl.gz	636,533	Fri Jun 30 17:00:11 2015	⊙
http://www.protonames.org/protonames-2SL.nl	309,678	Fri Jun 30 17:00:09 2015	⊙
http://e-culture.museum.nl/mv/ksd/attached/named.rdf	264,968	Fri Jun 30 17:00:08 2015	⊙
http://sparql.org/collections/ds/vocvoc:mbh_2_kk.nl	192,951	Fri Jun 30 17:00:10 2015	⊙
file://home/vdeboer/src/ClioPatris/dss/dss_costerone/arch.nl	154,031	Fri Jul 3 16:39:44 2015	⊙
http://sparql.org/collections/ds/dss/dss_data.nl	149,357	Fri Jun 30 17:00:09 2015	⊙
file://home/vdeboer/src/ClioPatris/dss/dss_costerone/vfb.nl	128,509	Fri Jul 3 16:39:44 2015	⊙
http://sparql.org/collections/ds/dss/ezemvcc/ezemvcc_data.nl	111,306	Fri Jun 30 17:00:10 2015	⊙
http://www.protonames.org/protonames.nl.as_skos.nl	42,814	Fri May 1 12:19:14 2015	⊙
file://home/vdeboer/src/ClioPatris/dss/dss_costerone/arch/Sama/As.nl	35,351	Fri Jul 3 16:39:44 2015	⊙
http://sparql.org/collections/ds/vocvoc:mbh_ship_sameta.nl	33,435	Fri Jun 30 17:00:10 2015	⊙
file://home/vdeboer/src/ClioPatris/dss/dss_costerone/vfb/Sama/As.nl	14,705	Fri Jul 3 16:39:44 2015	⊙
http://sparql.org/collections/ds/dss/voocopy/voocopy_dss.nl	12,851	Fri Jun 30 17:00:11 2015	⊙
http://sparql.org/collections/ds/dss/del/frnl.nl	8,426	Tue Apr 7 10:18:02 2015	⊙
http://sparql.org/collections/ds/dss/dss_data_thes_gren.nl	7,034	Fri Jun 30 17:00:09 2015	⊙
http://sparql.org/collections/ds/dss/dss_data_links.nl	5,449	Fri Jun 30 17:00:09 2015	⊙
http://sparql.org/collections/ds/dss/ezemvcc/ezemvcc_2_dss.nl	5,303	Fri Jun 30 17:00:09 2015	⊙
http://sparql.org/collections/ds/dss/ezemvcc/ezemvcc_gren_dss.nl	4,059	Tue Apr 7 10:18:02 2015	⊙
http://www.protonames.org/ontology_v2_2_1.rdf	2,895	Fri Jun 30 17:00:09 2015	⊙
http://sparql.org/collections/ds/dss/all_dss/2_protonames.nl	2,527	Fri Jun 30 17:00:09 2015	⊙
http://sparql.org/collections/ds/vocvoc:mbh_thes_places.nl	2,273	Fri Jun 30 17:00:10 2015	⊙
http://www.w3.org/ns/proxy	1,789	Mon Apr 29 17:46:07 2013	⊙
http://sparql.org/collections/ds/wildervank.nl	1,536	Tue Apr 7 10:18:02 2015	⊙
http://sparql.org/ds/home/schema.rdf	857	Fri Jun 23 10:48:38 2015	⊙
http://www.w3.org/ns/proxy#1.1/emptyname	631	Tue Jan 14 20:16:42 2014	⊙
http://sparql.org/collections/ds/ezemvcc/ezemvcc_dss_gren_places.nl	591	Fri Jun 30 17:00:10 2015	⊙
http://sparql.org/collections/ds/vocvoc:mbh_thes_ranzen.nl	585	Fri Jun 30 17:00:10 2015	⊙
http://www.w3.org/ns/proxy#2000/02/22rdf	490	Wed Jun 21 13:22:25 2017	⊙
http://sparql.org/collections/ds/dss/ezemvcc/ezemvcc_schema.nl	418	Tue Apr 7 10:18:02 2015	⊙
http://sparql.org/collections/ds/dss/voocopy_schema.nl	337	Fri Jun 30 17:00:12 2015	⊙
http://sparql.org/collections/ds/dss/ezemvcc/ezemvcc_cst.nl	273	Fri Jun 30 17:00:09 2015	⊙
file://home/vdeboer/src/ClioPatris/dss/ezemvcc/ezemvcc_base/ckn.rdf	256	Fri Apr 27 10:13:25 2015	⊙
http://sparql.org/collections/ds/vocvoc:tasks_and_shiptypes_1.nl	245	Fri Jun 30 17:00:09 2015	⊙
file://home/vdeboer/src/ClioPatris/dss/dss_costerone/schemaMap_chk.nl	241	Fri Jul 3 16:39:45 2015	⊙
http://sparql.org/collections/ds/ezemvcc/ezemvcc_schema.nl	232	Fri Jun 30 17:00:10 2015	⊙
file://home/vdeboer/src/ClioPatris/dss/dss_costerone/vfb/schema.nl	215	Fri Jul 3 16:39:45 2015	⊙
file://home/vdeboer/src/ClioPatris/dss/dss_costerone/vfb/schema.nl	209	Tue Aug 27 17:14:08 2015	⊙
http://sparql.org/collections/ds/vocvoc:mbh_thes_generated.nl	196	Fri Jun 30 17:00:10 2015	⊙

https://semanticweb.cs.vu.nl/dss/browse/list_graphs

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 84

84

```

@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix wgs: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ns9: <http://www.geonames.org/ontology#> .
@prefix dcterms: <http://purl.org/dc/terms/> .

# Named toplevel resources (20)

ns9:Feature
  a rdfs:Class ,
    owl:Class ;
  rdfs:comment "A geographical object uniquely defined by its
geonames id."@en ;
  rdfs:label "Feature"@en ;
  rdfs:subClassOf _:bn1 ,
    _:bn2 ,
    wgs:SpatialThing ,
    skos:Concept .

ns9:alternateName
  a rdf:Property ,
    owl:DatatypeProperty ;
  rdfs:domain ns9:Feature ;
  rdfs:label "alternateName" ;
  rdfs:range rdfs:Literal ;
  rdfs:subPropertyOf skos:altLabel .
    
```

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 85

85

Summary information for graph "http://www.w3.org/2002/07/owl"

Source URL:	file:///home/vdeboer/src/ChioPatrizia/ChioPatrizia.rdf#base/owl2.stl	Search this graph
# entities:	438	
# predicates:	13	
# subjects:	78	
# named subjects:	0	
# referenced classes:	8	

Local view for "http://www.w3.org/2002/07/owl"

Predicate	Value (sorted: default)
dc:title	"The OWL 2 Schema vocabulary (OWL 2)"
rdfs:type	owl:Ontology
rdfs:comment	" This ontology partially describes the built-in classes and properties that together form the basis of the RDF/XML syntax of OWL 2. The content of this ontology is based on Tables 6.1 and 6.2 in Section 6.4 of the OWL 2 RDF-Based Semantics specification, available at http://www.w3.org/TR/owl2-rdf-based-semantics/ . Please note that those tables do not include the different annotations (labels, comments and rdfs:isDefinedBy links) used in this file. Also note that the descriptions provided in this ontology do not provide a complete and correct formal description of either the syntax or the semantics of the introduced terms (please see the OWL 2 recommendations for the complete and normative specifications). Furthermore, the information provided by this ontology may be misleading if not used with care. This ontology SHOULD NOT be imported into OWL ontologies. Importing this file into an OWL 2 DL ontology will cause it to become an OWL 2 Full ontology and may have other, unexpected, consequences. "
rdfs:isDefinedBy	< http://www.w3.org/TR/owl2-annotation-sec4/#/ > < http://www.w3.org/TR/owl2-rdf-based-semantics/ > < http://www.w3.org/TR/owl2-syntax/ >
owl:imports	< http://www.w3.org/2000/01/rdf-schema# > < http://www.w3.org/2002/07/owl# >
rdfs:seeAlso	< http://www.w3.org/TR/owl2-rdf-based-semantics/#table-axiomatic-classes > < http://www.w3.org/TR/owl2-rdf-based-semantics/#table-axiomatic-properties >
owl:versionInfo	"Date: 2009-11-15 10:54:13.4"
owl:versionIRI	The OWL 2 Schema vocabulary (OWL 2)

If properties reside in the graph <http://www.w3.org/2002/07/owl>

Persistence information

This graph has no associated persistence files

Actions

Show this graph as Turtle

https://semanticweb.cs.vu.nl/dss/browse/list_graph?graph=http%3A//www.w3.org/2002/07/owl

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 86

86

UNIFORM RESOURCE IDENTIFIERS (URIs) EXPRESS RDF STATEMENTS

- ❑ **NOTE:** ALL URLs can be described as URIs, but not vice versa
 - ❑ Explanation – URIs identify things uniquely whereas URLs also capture their location
 - ❑ URI for Jean-Claude Gardin, https://dbpedia.org/page/Jean-Claude_Gardin
 - ❑ URL (and URI) for Jean-Claude Gardin, https://fr.wikipedia.org/wiki/Jean-Claude_Gardin
- ❑ Uniform Resource Identifiers (URIs) express RDF statements
 - ❑ <https://viaf.org/viaf/54147973/rdf.xml> Jean-Claude Gardin
- ❑ Intended for representing "data" about web resources (e.g., title)
- ❑ Representing "things" that can be identified, reasoned about, but are not on the Web
- ❑ Essential to understand the data set or sets you release as L(O)D
- ❑ Concepts represented in controlled vocabularies or ontologies
- ❑ Attributes often encapsulate underlying concepts
- ❑ Model relationships

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 87

87

Use HTTP URIs
To benefit from and increase the value of the World Wide Web, governments and agencies **SHOULD** provide HTTP URIs as identifiers for their resources. There are many benefits to participating in the existing network of URIs, including linking, caching, and indexing by search engines. As stated in [howto-ldp], HTTP URIs enable people to "look-up" or "dereference" a URI in order to access a representation of the resource identified by that URI. To benefit from and increase the value of the World Wide Web, data publishers **SHOULD** provide URIs as identifiers for their resources.

Provide at least one machine-readable representation of the resource identified by the URI
In order to enable HTTP URIs to be "dereferenced", data publishers have to set up the necessary infrastructure elements (e.g. TCP-based HTTP servers) to serve representations of the resources they want to make available (e.g. a human-readable HTML representation or a machine-readable Turtle). A publisher may supply zero or more representations of the resource identified by that URI. However, there is a clear benefit to data users in providing at least one machine-readable representation. More information about serving different representations of a resource can be found in [COLURIS].

A URI structure will not contain anything that could change
It is good practice that URIs do not contain anything that could easily change or that is expected to change like session tokens or other state information. URIs should be stable and reliable in order to maximize the possibilities of reuse that Linked Data brings to users. There must be a balance between making URIs readable and keeping them more stable by removing descriptive information that will likely change. For more information on this see [Architecture of the World Wide Web: URI Persistence](#).

URI Opacity
The Architecture of the World Wide Web [webarch], provides best practices for the treatment of URIs at the time they are resolved by a Web client. Agents making use of URIs **SHOULD NOT** attempt to infer properties of the referenced resource. URIs **SHOULD** be constructed in accordance with the guidance provided in this document to ensure ease of use during development and proper consideration to the guidelines given herein. However, Web clients accessing such URIs **SHOULD NOT** parse or otherwise read into the meaning of URIs.

<https://www.w3.org/TR/ld-bp/#HTTP-URIS>

```

@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

# Named toplevel resources (1)

<http://www.w3.org/2003/g/data-view#namespaceTransformation>
  a rdf:Property ;
  rdfs:domain owl:Ontology ;
  rdfs:label "namespaceTransformation" ;
  rdfs:range rdfs:Resource .
    
```

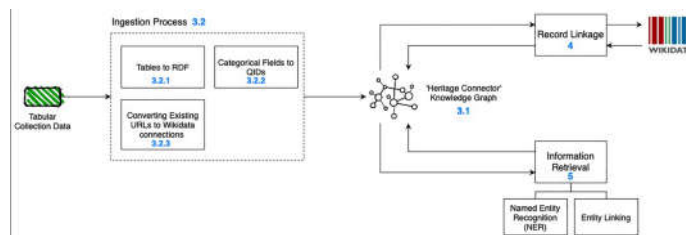
© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 88

88

UNIFORM RESOURCE IDENTIFIERS (URIs) EXPRESS RDF STATEMENTS

- ❑ Getting from your data to Linked (Open) Data
- ❑ Choose appropriate ontologies/vocabularies
 - ❑ As the distinction is often unclear:
 - ❑ "A bookseller may want to integrate data coming from different publishers. The data can be imported into a common RDF model, eg, by using converters to the publishers' databases. However, one database may use the term "author", whereas the other may use the term "creator". To make the integration complete, an extra definition should be added to the RDF data, describing the fact that the relationship described as "author" is the same as "creator". This extra piece of information is, in fact, a vocabulary (or an ontology), albeit an extremely simple one." From: <https://www.w3.org/standards/semanticweb/ontology>
- ❑ Determine which data will create value through Linking and making accessible
- ❑ There will be a need to "pull" the data from the resource (e.g., SQL Query)
- ❑ Select software to transform your data into Linked Data

95



Kaylan Dutia and John Stack, 2021, "Heritage Connector: A machine learning framework for building linked open data from museum collections," Applied AI Letters, 3 May 2021, <https://doi.org/10.1002/aill.2.23>

96

As Dominik Lukas, Claudia Engel and Camilla Mazzucato, 2018, "Towards a Living Archive: Making Multi Layered Research Data and Knowledge Generation Transparent", *Journal of Field Archaeology*, 43:sup1, S19-S30, DOI: [10.1080/00934690.2018.1516110](https://doi.org/10.1080/00934690.2018.1516110). Figure 2 on page S25.

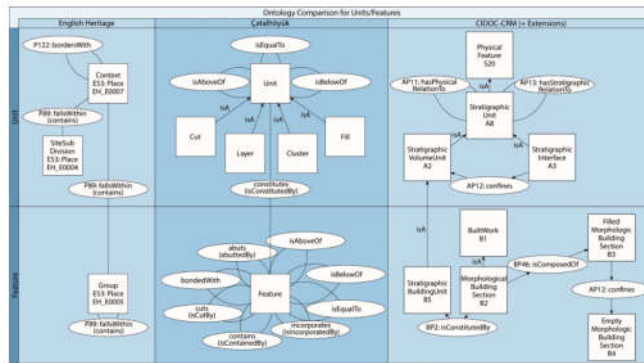


Figure 2 Mapping of Catalhöyük concepts for different contexts (unit, feature) to the English Heritage (EH-CRM) and CIDOC Conceptual Reference Models (CIDOC-CRM, including presently available extensions).

97

As Dominik Lukas, Claudia Engel and Camilla Mazzucato, 2018, "Towards a Living Archive: Making Multi Layered Research Data and Knowledge Generation Transparent", *Journal of Field Archaeology*, 43:sup1, S19-S30, DOI: [10.1080/00934690.2018.1516110](https://doi.org/10.1080/00934690.2018.1516110) explain in a discussion of the Çatalhöyük Research Project and its (living) archive. Figure 3, S27.

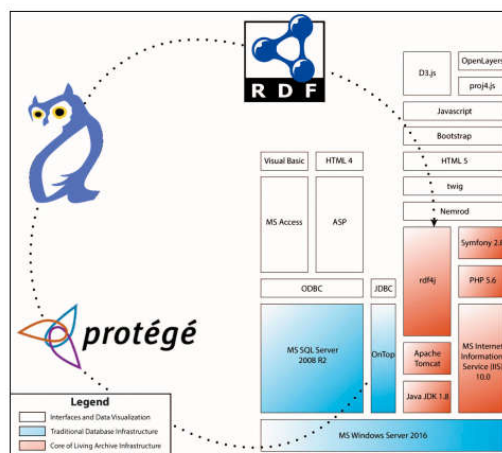


Figure 3 Transformation of data from the original relational database into semantic triples. Protégé (with the plugin OnTop) is used to develop the ontology (owl) and to export the semantic triples in rdf/xml-format for ingestion into the rdfj triple store. The RDF logo was originally designed by Bill Schwappacher. The OWL logo and the RDF logo are both donated to the W3C (World Wide Web Consortium). Contents from the W3C are licensed under the W3C Document license <http://www.w3.org/Consortium/Legal/2002/copyright-documents-20021231>

98

The screenshot shows the 'Introduction' page of the Ontop website. The page title is 'Introduction' and the logo 'ontop' is visible in the top left. The text describes Ontop as a Virtual Knowledge Graph system that supports the content of arbitrary relational databases as knowledge graphs. It mentions that Ontop translates SPARQL queries expressed over the knowledge graphs into SQL queries executed by the relational data sources. The page also includes sections for 'Versions' and 'Main features'. The URL 'https://ontop-vkg.org/' is displayed at the bottom of the page content.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 99

99

The screenshot shows the article page for 'The enslaved ontology: Peoples of the historic slave trade' in the 'Journal of Web Semantics'. The article is published by Elsevier. The authors listed are Cogan Shimizu, Pascal Hitzler, Quinn Hirt, Dean Rehberger, Seila Gonzalez Estrecha, Catherine Foley, Alicia M. Sheill, Walter Hawthorne, Jeff Mixter, Ethan Watrall, Ryan Carty, and Duncan Tarr. The abstract states: 'We present the Enslaved Ontology (V1E) which was developed for integrating data about the historic slave trade from diverse sources in a one case driven by historians. Ontology development followed modular ontology design principles as derived from ontology design patterns application best practices and the eXtreme Design Methodology. Ontology content focuses on data about historic persons and the events records from which this data can be taken. It also incorporates provenance modeling and some temporal and spatial aspects. The ontology is available as serialized in the Web Ontology Language OWL, and carries modularization annotations using the Ontology Pattern Language (OPL). It is available under the Creative Commons CC BY 4.0 license.' The article is dated 2020.

Cogan Shimizu, Pascal Hitzler, Quinn Hirt, Dean Rehberger, Seila Gonzalez Estrecha, Catherine Foley, Alicia M. Sheill, Walter Hawthorne, Jeff Mixter, Ethan Watrall, Ryan Carty, Duncan Tarr, 2020, "The enslaved ontology: Peoples of the historic slave trade," *Journal of Web Semantics*, V 63, <https://doi.org/10.1016/j.websem.2020.100567>.

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 100

100

Page: [unreadable]

RDF-Based Semantics

[View Source | Contents]

Document title:
 OWL 2 Web Ontology Language
 RDF-Based Semantics (Second Edition)

Editor:
 Michael Schuster¹, FZI Research Center for Information Technology

Contributors (alphabetical order):
 Jeremy Carroll, HP Labs at Yahoo! Research
 [unreadable], HP Labs
 Peter F. Patel-Schneider, InnoCentives

Abstract
 The OWL 2 Web Ontology Language, informally OWL 2, is an ontology language for the Semantic Web with formally defined modeling. OWL 2 ontologies provide classes, properties, individuals, and data in. Particulars are generally described as RDF documents. The OWL 2 Document Overview describes the overall state of OWL 2, and should be read before other OWL 2 documents.
 This document defines the RDF-compatible model-theoretic semantics of OWL 2.

Status of this Document
 Copyright © 2002-2020 (W3C®) (MIT®, ECSCM®), All Rights Reserved. W3C, liability®, trademark® and document use® rules apply.

Contents (toc)

- 1 Introduction (Informative)
- 2 Ontologies
 - 2.1 Syntax
 - 2.2 Content of Ontologies (Informative)
- 3 Vocabulary
 - 3.1 Standard Profiles
 - 3.2 Vocabulary Terms
 - 3.3 Disjoint Names
 - 3.4 Facet Names
- 4 Interpretation
 - 4.1 Semantic Maps
 - 4.2 Vocabulary Interpretations
 - 4.3 Satisfaction, Consistency and Entailment
 - 4.4 Paths of the Universe
 - 4.5 Class Expressions
- 5 Semantic Conditions
 - 5.1 Semantic Conditions for the Paths of the Universe
 - 5.2 Semantic Conditions for the Vocabulary Classes
 - 5.3 Semantic Conditions for the Vocabulary Properties
 - 5.4 Semantic Conditions for Disjoint Classifications
 - 5.5 Semantic Conditions for Expressions

Web Ontology Language (OWL), https://www.w3.org/2007/OWL/wiki/RDF-Based_Semantics

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 101

101

Vocabulary	Main use	Sample relations
Dublin Core	Documents	creator date rights
FOAF (Friend of a friend)	People and relationships	firstName familyName knows
SKOS (Simple Knowledge Organization System)	Thesauri	broader narrower prefLabel
OWL (Web Ontology Language)	Ontologies	sameAs

Table 1: Sample relations from vocabularies or ontologies used in RDF triples

□ *Lyne Da Sylva, 2018. "Towards linked data: Some consequences for researchers in the social sciences and humanities". *Proceedings of the Association for Information Science and Technology*, 55(1), 94–103. <https://doi.org/10.1002/prs.2018.14505501011> page 96

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 102

102

The screenshot shows the Linked Open Vocabularies (LOV) website. On the left, a circular network of 754 vocabularies is displayed, with prominent nodes for 'vann', 'foaf', 'skos', 'dcterms', and 'dce'. On the right, a detailed view of the 'DCMI Metadata Terms (dcterms)' vocabulary is shown, including a list of terms and a smaller version of the network diagram.

Linked Open Vocabularies, <https://lov.linkeddata.es/dataset/lov>

<https://lov.linkeddata.es/dataset/lov/vocabs/dcterms>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 103

103

The screenshot shows the 'About LOV' page on the Linked Open Vocabularies website. The page title is 'In LOV at a glance ...'. The content explains that LOV stands for Linked Open Vocabularies, derived from LOD (Linked Open Data). It describes how vocabularies describe and link data on the web, providing definitions for classes and properties. The page also lists features such as Vocabulary Documentation, Data Access, Vocabulary Search Engine, Application Ecosystem, and Applications using LOV.

<https://lov.linkeddata.es/dataset/lov/about>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 104

104

<https://lov.linkeddata.es/dataset/lov/about>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

105

105

<https://www.getty.edu/research/tools/vocabularies/lov/>

“URIs (Uniform Resource Identifiers): A description of the full set of URIs for the Getty Vocabularies is available at”
https://www.getty.edu/research/tools/vocabularies/lov/aat_semantic_representation.pdf

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

106

106

PLEIADES

Home Places Credits Participate Blog Documentation Downloads

38,360 Places **34,505** Names **41,138** Locations

Recently Modified Resources

About Pleiades

Pleiades gives scholars, students, and enthusiasts worldwide the ability to use, create, and share historical geographic information about the ancient world in digital form. At present, Pleiades has extensive coverage for the Greek and Roman world, and is expanding into Ancient Near Eastern, Byzantine, Celtic, and Early Medieval geography.

The most recently modified resources are shown in the map at left.

All published content is accessible to everyone under open license. To join and contribute new or improved content, please see [Welcome to Pleiades](#).

For a complete listing of editors, content contributors, and financial supporters, please see the [credits page](#).

Search [Search Site] 14,868
Advanced Search...

News

Maintenance 1 December 2020
Nov 25, 2020

Pleiades Datasets 2.3 released (12 October 2020)
Oct 12, 2020

New Relationship Type: "crosses"
Sep 28, 2020

Pleiades Datasets 2.2 released (7 April 2020)
Apr 13, 2020

Pleiades Datasets 2.1 released
Feb 27, 2020

More news...

Copyright © Ancient World Mapping Center and Institute for the Study of the Ancient World. Sharing and reusing permitted under terms of the Creative Commons Attribution 3.0 License (cc-by). Please see our credits.

Powered by Pines & Pythons

<https://pleiades.stoa.org/home>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

107

107

The Linked Open Data Cloud

Browse Search a dataset Diagnostics Subscribers About

Pleiades (1/201)

★★★★

About this dataset

Pleiades is a gazetteer for ancient world places operated by NYU's Institute for the Study of the Ancient World and supported by the US National Endowment for the Humanities. It is derived originally from the Barrington Atlas of the Greek and Roman World and currently adds new resources. Pleiades includes "35,000+ ancient places" (ca. 1000+ ancient names and name variants with time periods) "12,000+ ancient locations" (Permanent, call URIs for these resources (<http://pleiades.stoa.org/id/urn:cts:stoa:pleiades:place> for example) "Maps and KML, GeoJSON, Turf, and RDF/XML variants of resources (<http://pleiades.stoa.org/geo/geojson> for example) "Spatial queries" (Query tabular results of locations, names, and place data and available at <http://datafind.org/geo/geojson/pleiades/geo>) There is a detailed FAQ (linked) published with the theme. (linked) <http://pleiades.stoa.org/downloads> (<http://pleiades.stoa.org/geo/geojson/pleiades/geo>) **##** Description of geodata from the <http://pleiades.stoa.org> - This KML file, for use with Google Earth and other compatible systems, contains basic coordinate and name information for ancient Greek and Roman place names (published by the Mapping Editors of the Ancient World Mapping Center's Pleiades community) (<http://pleiades.stoa.org>). In this edition, it comprises place features in Latin, Pinyin, and Chinese, and includes corresponding coordinates and place names. It is a subset of the data available in the Pleiades dataset, and will add more information to the individual place name descriptions. - - - This content is original work of the staff of the Ancient World Mapping Center and members of the Pleiades Community. It is built in part on information that was compiled by the American Philological Association's Classical Atlas Project (1986-2002), which was used during development with the permission of the APhA. - - - Check out our efforts. Pleiades content may contain minor omissions. These should be assumed to be the responsibility of the project director, and not to reflect the quality and completeness of Classical Atlas Project data nor the opinions of the Atlas Project's compilers and editors. Pleiades will open to public participation in early 2018. At this time, it will be open to users to highlight and correct errors and omissions, and to update obsolete information. - - - Coordinate accuracy and precision are discussed at <http://pleiades.stoa.org/geo/geojson/pleiades/geo/accuracy> **##** Openness: OPEN "License: cc-by" (see bottom of <http://www.ancientworldmappingcenter.org/pleiades/geo/geojson/pleiades/geo>) **##** License: <http://creativecommons.org/licenses/by/4.0/> with details on <http://pleiades.stoa.org/terms/> with license: <http://creativecommons.org/licenses/by/4.0/>

Contact Details

Contact Point: Pleiades Project
Website: <http://pleiades.stoa.org/>

Download Links

Examples

- Example data from: <http://pleiades.stoa.org/geo/geojson/pleiades/geo> (Turkish)

Other downloads

- Download: <http://pleiades.stoa.org/geo/geojson/pleiades/geo> (Turkish)
- Download: <http://pleiades.stoa.org/geo/geojson/pleiades/geo> (Turkish)
- Download: <http://pleiades.stoa.org/geo/geojson/pleiades/geo> (Turkish)

Data Facts

Total size: 2,600,000+ triples

Namespace: <http://pleiades.stoa.org/id/> <http://pleiades.stoa.org/id/urn:cts:stoa:pleiades:place> <http://datafind.org/geo/geojson/pleiades/geo>

Links to Pleiades: 127+ triples

<https://lod-cloud.net/dataset/pleiades>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)

108

108

<https://pleiades.stoa.org/news/blog/pleiades-datasets-2-3>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 109

109

<https://pleiades.stoa.org/vocabularies/relationship-types>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 110

110

PLEIADES

Home | Places | Credits | Participate | Blog | Documentation | Downloads

You are here: Home — Vocabularies

Vocabularies

Created: Sean Gillis
Last modified: Dec 28, 2011 01:59 PM

Controlled vocabularies employed by Pleiades including time periods and place categories

- Association Certainty — by — last modified Jan 23, 2012 01:18 PM
- Level of certainty in association between places and locations or names
- Attention Confidence — by — last modified Dec 08, 2011 08:17 AM
- Level of confidence in temporal attestation
- Pleiades Vocabulary Language and Script — by Sean Gillis — last modified Dec 08, 2011 01:59 PM
- Language and script of the attested name. We use the ISO-639-1 code (http://en.wikipedia.org/wiki/ISO_639-1)
- Name Accuracy — by — last modified Dec 08, 2011 07:45 AM
- Level of transcription accuracy
- Name Completeness — by — last modified Dec 08, 2011 08:34 AM
- Level of transcription completeness
- Name Types — by — last modified Dec 09, 2011 02:03 AM
- The several types of appellations in Pleiades
- Time Periods
- Named time periods associated with a range of years
- Features (or Place) Categories
- Once we called these "types", but as they are not hierarchical and not exclusive we have changed them to "categories"
- Location Categories
- Once we called these "types", but as they are not hierarchical and not exclusive we have changed them to "categories"
- Connection Types
- Categories of relationships between places
- Archaeological Remains
- Categories of archaeological remains

Copyright © Ancient World Mapping Center and Institute for the Study of the Ancient World. Sharing and reusing permitted under terms of the CC-BY license.

Powered by Drupal 6.31.0

<https://pleiades.stoa.org/vocabularies>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 111

111

PLEIADES

Home | Places | Credits | Participate | Blog | Documentation | Downloads

You are here: Home — How to cite Pleiades

How to cite Pleiades

Created: Sean Gillis
Last modified: Dec 28, 2011 01:59 PM

Instructions on citing Pleiades data, including a list of metadata fields and a list of links to related resources.

Copyright © Ancient World Mapping Center and Institute for the Study of the Ancient World. Sharing and reusing permitted under terms of the CC-BY license.

Powered by Drupal 6.31.0

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 112

112

PLEIADES
 Home Places Credits Participants Blog Resources/About Us Downloads

Persepolis/Parsa/Persai/Sat. Setun

Overview
 This is a new place. It has been added to the Pleiades database. You can view the details of this place on the map or in the list view.

Coordinates
 WGS84: 32° 02' 00.00" N, 52° 00' 00.00" E
 UTM: 49QKJ 611000 3448000 49QKJ 611000 3448000

Locations
 • 32° 02' 00.00" N, 52° 00' 00.00" E
 • 32° 02' 00.00" N, 52° 00' 00.00" E
 • 32° 02' 00.00" N, 52° 00' 00.00" E

Names
 • Persepolis/Parsa/Persai/Sat. Setun
 • Persai/Sat. Setun
 • Parsai/Sat. Setun
 • Persai/Sat. Setun
 • Parsai/Sat. Setun

Public Use
 Attribution: Pleiades.org
 License: CC BY-NC-SA
 URL: <https://pleiades.stoa.org/places/922695>

Related Content
 About Pleiades
 Pleiades.org

<https://pleiades.stoa.org/places/922695>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 113

113

URI: <https://pleiades.stoa.org/places/922695>

flickr Explore Prints Get Pro

Explore Trending Events

Tags **pleiades:findspot=922695**

All Photos Tagged pleiades:findspot=922695

5th-4th c. BCE
From Persa

<https://flickr.com/photos/tags/pleiades:findspot=922695>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 114

114

Received: 19 December 2020 | Revised: 27 April 2021 | Accepted: 28 April 2021
DOI: 10.1002/aii2.23

WILEY

LETTER

Heritage connector: A machine learning framework for building linked open data from museum collections

Kalyan Dutia | **John Stack**

Science Museum Group, London, UK

Correspondence
Kalyan Dutia and John Stack, Science Museum Group, London, UK.
Email: kalyan.dutia@smg.ac.uk; john.stack@sciencemuseum.ac.uk

Funding information
Arts and Humanities Research Council

Abstract

As with almost all data, museum collection catalogues are largely unstructured, variable in consistency and overwhelmingly composed of thin records. The form of these catalogues means that the potential for new forms of research, access and scholarly enquiry that range across multiple collections and related datasets remains dormant. In the project *Heritage Connector: Transforming text into data to extract meaning and make connections*, we are applying a battery of digital techniques to connect similar, identical and related objects within and across collections and other publications. In this article, we describe a framework to create a Linked Open Data knowledge graph from digital museum catalogues, perform record linkage to Wikidata, and add new entities to this graph from textual catalogue record descriptions (information retrieval). We focus on the use of machine learning to create these links at scale with a small amount of labelled data, and models which are small enough to run inference on datasets the size of museum collections on a mid-range laptop or a small cloud virtual machine. Our method for record linkage against Wikidata achieves 85%+ precision with the Science Museum Group (SMG) collection, and our method for information retrieval is shown to improve NER performance compared with pretrained models on the SMG collection with no labelled training data. We publish open-source software providing tools to perform these tasks.

Kalyan Dutia and John Stack, 2021, "Heritage Connector: A machine learning framework for building linked open data from museum collections," *Applied AI Letters*, 3 May 2021, <https://doi.org/10.1002/aii2.23>

Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)
115

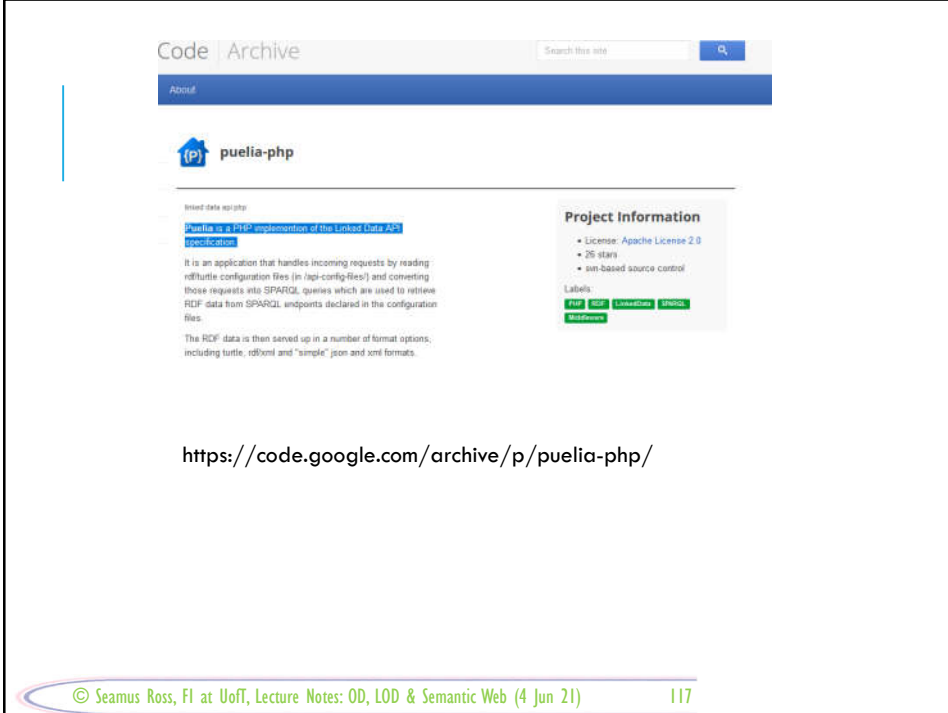
115

DEFINING ACCESS TO LINKED DATA RESOURCES

- Access to the Linked (Open) Data could be by navigating the graph
- Providing the RDF data for download ("data dump")
 - No Query ability, dependent upon user ability and experience, but facilitates reuse
- Access via a SPARQL Endpoint
 - Low reliability, slow, complex for uninitiated, but rich for expert
- We will discuss SPARQL and experiment with that
- Access via an "Insulated SPARQL" handle
 - For instance: "Puelia is a PHP implementation of the Linked Data API specification."

Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21)
116

116



Code Archive

About

puelia-php

Latest data snapshot:

puelia-php is a PHP implementation of the **Linked Data API** specification.

It is an application that handles incoming requests by reading rdflite configuration files (in /api-config-files/) and converting those requests into SPARQL queries which are used to retrieve RDF data from SPARQL endpoints declared in the configuration files.

The RDF data is then served up in a number of format options, including turtle, rdflib and "simple" json and xml formats.

Project Information

- License: Apache License 2.0
- 26 stars
- link-based source control

Labels: [API](#) [API](#) [LinkedData](#) [URIs](#) [URIs](#)

<https://code.google.com/archive/p/puelia-php/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 117

117

SPARQL QUERY LANGUAGE FOR SEMANTIC WEB

- SPARQL (SPARQL Protocol and RDF Query Language)
 - Query Language
 - SPARQL queries contain triple patterns.
 - Enable Quickly Navigating Linked Data
 - simply--"translates" interlinked graph data into tabular data represented as rows and columns
 - Engage with both structured and semi-structure data
 - 'Perform complex joins of disparate databases'*
 - Supports exploration of 'unknown relationships'*

*See SPARQL By Example: Tutorial, Lee Feigenbaum, 2009, <https://www.w3.org/2009/Talks/0615-qbe/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 118

118

<https://www.w3.org/wiki/SparqlEndpoints>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 119

119


<https://www.w3.org/wiki/VirtuosoUniversalServer>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 120

120

```

SELECT ?Archaeologist_Name
WHERE {
  ? Archaeologist_Name ?b
  <http://dbpedia.org/class/yago/WikicatItalianArchaeologists>.
}
    
```




© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 121

121

FROM ACTIVITY TO IMPACT

Table 1. Example of output-activity-outcome-impact for Liander.

	Activity	Output	Outcome	Impact
Liander	Releasing data as open data	Open small-scale consumption dataset	Energy apps based on open data	Contribution to energy conservation

Types of Effects

- External (e.g., users, intensity, planning and policy, increase in page views, data re-use)
- Internal (e.g., transaction costs, data quality, usage)
- Relational (e.g., bi-directional communication, brand image)

After Frederika Welle Donker, Bastiaan van Loenen, and Arnold K Bregt, 2016, "Open Data and Beyond," *International Journal of Geo-Information.*, 5(4), 48-; doi:10.3390/ijgi5040048 p 6

122

Summary of Best Practices

The following best practices are discussed in this document and listed here for convenience.

STEP #1 PREPARE STAKEHOLDERS:
Prepare stakeholders by explaining the process of creating and maintaining [Linked Open Data](#).

STEP #2 SELECT A DATASET:
Select a dataset that provides benefit to others for reuse.

STEP #3 MODEL THE DATA:
[Modeling Linked Data](#) involves representing data objects and how they are related in an application-independent way.

STEP #4 SPECIFY AN APPROPRIATE LICENSE:
Specify an appropriate open data license. Data reuse is more likely to occur when there is a clear statement about the origin, ownership and terms related to the use of the published data.

STEP #5 GOOD URIS FOR LINKED DATA:
The core of Linked Data is a well-considered URI naming strategy and implementation plan, based on [HTTP URIs](#). Consideration for naming objects, multilingual support, data change over time and persistence strategy are the building blocks for useful Linked Data.

STEP #6 USE STANDARD VOCABULARIES:
Describe objects with previously defined [vocabularies](#) whenever possible. Extend standard vocabularies where necessary, and create vocabularies (only when required) that follow best practices whenever possible.

STEP #7 CONVERT DATA:
Convert data to a Linked Data representation. This is typically done by script or other automated processes.

STEP #8 PROVIDE MACHINE ACCESS TO DATA:
Provide various ways for search engines and other automated processes to access data using standard Web mechanisms.

STEP #9 ANNOUNCE NEW DATA SETS:
Remember to [announce](#) new data sets on an authoritative domain. Importantly, remember that as a Linked Open Data publisher, an [implicit social contract](#) is in effect.

STEP #10 RECOGNIZE THE SOCIAL CONTRACT:
Recognize your responsibility in maintaining data once it is published. Ensure that the dataset(s) remain available where your organization says it will be and is maintained over time.

<https://www.w3.org/TR/ld-bp/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 123

123

Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery

Eero Hyonen
University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland and
Aalto University, Department of Computer Science, Finland
Semantic Computing Research Group (SeCo) (<http://seco.aalto.fi>)
E-mail: eero.hyonen@aalto.fi

Rolfes, Pascal (Helm), Kansas State University, Manhattan, KS, USA; Kryptosid, University of California, Santa Barbara, USA
Submitted version: Rafael Gonzalez, Stanford University, CA, USA; Peter Haas, semaphor GmbH, Wulfbell, Germany (On company review)

Abstract. This paper discusses a shift of focus in research on Cultural Heritage semantic portals, based on Linked Data, and outlines and proposes new directions of research. These generations of portals are identical. Ten years ago the research focus in semantic portal development was on data harmonization, aggregation, search, and browsing ("first generation systems"). At the moment, the rise of Digital Humanities research has started to shift the focus to providing the user with integrated tools for solving research problems in interactive ways ("second generation systems"). This paper discusses and argues that the next step toward "third generation systems" is based on Artificial Intelligence: future portals not only provide tools for the human to solve research problems but are used for finding research problems in the first place, for addressing them, and even for solving them automatically under the constraints set by the human researcher. Such systems should particularly be able to explain their reasoning, which is an important aspect in the current critical humanities research tradition. The second and third generation systems set new challenges for both computer scientists and humanities researchers.

Keywords: Digital Humanities, Linked Data, Semantic portals, Data analysis, Knowledge discovery

1. Introduction

Cultural Heritage (CH) has become a most active area of application of Linked Data and Semantic Web (SW) technologies [1]. Large amounts of CH content and metadata about it are available openly for research and public use based on collections in museums, libraries, archives, and media organizations. For example, data has been aggregated in large national and international repositories, web-services, and portals such as Europeana¹ and Digital Public Library of America², and forms a substantial part of DDFeels³ and Wikidata⁴.

The availability of Big Data has boosted the rapidly emerging new research area of Digital Humanities (DH) [2, 3] where computational methods are developed and applied to solving problems in humanities.

<http://europeana.eu>
<http://dpla.or>
<http://ddfeels.org>
<http://wikidata.org>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 124

124

SPARQL Query Editor About Tables

Default Data Set Name (Graph URI)
https://dbpedia.org

Query Text

Results Format: Turtle

Execute Query Reset

Execution timeout: 30000 milliseconds

Options

- Strict checking of void variables
- Strict checking of variable names used in multiple clauses but not logically connected to each other
- Suppress errors on wrong geometries and errors on geometrical operators (failed operations will return NULL)
- Log debug info at the end of output (has no effect on some queries and output formats)
- Generate SPARQL compilation report (instead of executing the query)

Copyright © 2021 OpenLink Software
Virtuoso version 08.03.3321 on Linux (x86_64-generic-linux-glibc25) Single Server Edition (61 GB total memory)

<https://dbpedia.org/sparql/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 125

125

Palladio: Visualize complex historical data with ease.

Start »

About

Palladio is a product of the historicaldata.stanford.edu group (July 2015 - June 2016). NetworkX is a graph data-driven toolkit for analyzing relationships across time. Our goal was to understand how to design graphical interfaces based on historical inquiry. We oriented the project around the development of a general purpose suite of visualization and analytical tools based on the principles created for the [Mapping the Republic of Letters](http://mappingtherepublicofletters.org) project, which examines the scholarly correspondences and networks of knowledge during the period 1500-1800.

Modeling Historical Data to Understand the Past

In 2016, students in Jeff Levent's CS468 Data Visualization class at Stanford produced an award-winning interactive visualization tool for the [Mapping the Republic of Letters](http://mappingtherepublicofletters.org) project that let historians explore 17th and 18th century correspondence in a novel way. Early modern communication networks previously only imagined in the minds of historians could now be seen graphically as pushing articles connecting the centers of Enlightenment thought. For the Stanford team of four data humanities faculty, tens of graduate students, and numerous international partners, the best-of-both-worlds was possible for producing knowledge. They used it as an example of data visualization transforming historical research. But it is the way the visualization helped to help us answer actual research questions that made us realize that we needed to become actively engaged in the design of both.

<http://hdlab.stanford.edu/palladio/about/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 126

126


Palladio: Visualize complex historical data with ease

[Start](#)

How to load data from a SPARQL endpoint

Data can be loaded from a SPARQL endpoint. The SPARQL language is used to query Linked Open Data (LOD) from various museums and heritage collections. An [invaluable tutorial](#) on loading data from a SPARQL endpoint can be found at [The Programming Historian](#).

To load data from a SPARQL endpoint, first click "Load data from a SPARQL endpoint Data!". You will be prompted to identify the path of the endpoint (for instance, the British Museum's public-facing SPARQL page is at <http://collection.britishmuseum.org/sparql>), but the SPARQL endpoint to allow Palladio to connect to the British Museum is "<http://collection.britishmuseum.org/sparql/join/>". An incomplete list of SPARQL endpoints is available [here](#).

SPARQL endpoint Data!

You can load data from a SPARQL endpoint by providing both the endpoint URL and a valid SPARQL query. After you run the endpoint query you will have the opportunity to validate your data and re-run the query if necessary before loading the data into Palladio.

SPARQL endpoint

Next, enter your SPARQL query. Correctly formatting SPARQL queries (including properly connecting multiple linked datasets) is outside the scope of this Tutorial; again, we highly recommend the Programming Historian's tutorial linked above, but when testing your queries it is a good idea to restrict the initial dataset to no more than 100 rows by concluding with "LIMIT 100" in your query.

After entering your SPARQL endpoint and query, click "Run Query". You will be given the chance to review your results before importing them by clicking "Load Data". (Note that for instance Palladio uses images for a gallery as the path to that image, not an image itself!)

<http://hdlab.stanford.edu/palladio/tutorials/loadSPARQL/>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 127

127

ResearchSpace: Map British Museum Data to CIDOC CRM

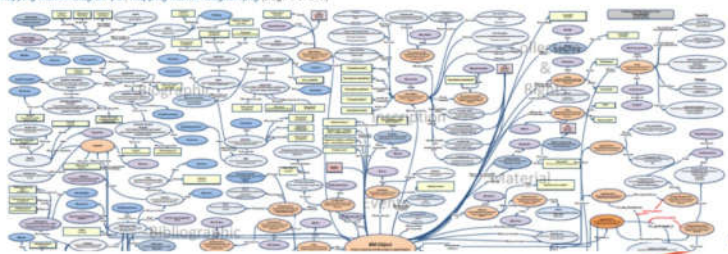
The Conceptual Reference Model Revealed
 Quality contextual data for research and engagement: A British Museum case study
 Dominic Oldham, Jordan Maternak, Vladimir Alexiev
 Version: Draft, 0.96, July 2013 (Confidential & Private – Limited Distribution for Discussion)


Contents: 359p

- 169: Main body, including discussion, illustrations and mapping diagrams
- 7p: Association Codes (see details at [BIM Association Mapping v2](#))
- 47p: Example Object Graph
- 134p: RDF/r configuration files (i.e. mapping implementation)

Overall Picture

[mapping-manual-diagram.pdf](#), [mapping-manual-diagram.png](#) (Page 9 of 359)






[Alexiev, https://www.slideshare.net/vladalexiev/museum-linked-open-data-ontologies-and-users-projects](https://www.slideshare.net/vladalexiev/museum-linked-open-data-ontologies-and-users-projects)

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 128

128

<https://hdlab.stanford.edu/palladio/tutorials/loadSPARQL/>



Palladio. Visualize complex historical data with ease

How to load data from a SPARQL endpoint

To load data from a SPARQL endpoint, first click "Load data from a SPARQL endpoint (Beta)". You will be prompted to identify the path of the endpoint (for instance, the British Museum's public-facing SPARQL endpoint is at <http://collection.britishmuseum.org/sparql>). An incomplete list of SPARQL endpoints is available here.

SPARQL endpoint (URL)

<https://collection.britishmuseum.org/sparql>

502 Bad Gateway

nginx/1.10.3 (Ubuntu)

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 129

129

The Orlando Women's Writing Linked Open Data Set (LOD) ★★

About this dataset

A linked open data set created from the original entries of the Orlando Project, an ongoing collaborative experiment in the use of computers to engage in women's literary history.

Contact Details

Contact Point: The CWRC Project

Download Links

Full Downloads

- Download (Weekly dump in n-quads triple + graph format)

SPARQL Endpoints

- SPARQL endpoint (SPARQL endpoint with all triples including ontologies. This endpoint is CORS compliant)

Examples

- Example Link (The resource will be resolved via content negotiation to RDF or HTML.)

Other downloads

- Download page (N-triples, gz compressed) (Raw dumps of the triples are located here)
- RDF Schema (All of the OWL files used by CWRC/Orlando, newest versions)
- Semantic Web Stories (The Orlando Project is an experiment in the integration of text and technology. It has designed and continues to enhance digital tools to harness the power of computers for critical literary and historical research. The project's constantly expanding and improving storehouse of knowledge about women's lives and writings, the Orlando website is based in order to be openly searchable and accessible by its encoding in a linked open data format.)
- and description (The Orlando Project is an experiment in the integration of text and technology. It has designed and continues to enhance digital tools to harness the power of computers for critical literary and historical research. The project's constantly expanding and improving storehouse of knowledge about women's lives and writings, the Orlando website is based in order to be openly searchable and accessible by its encoding in a linked open data format.)
- Mappings (Mappings are contained as part of the ontologies.)
- Download triples (Weekly data dump in n3 triple format)

Data Facts

Total size	3,459 triples
Links to ADT	127 triples
Links to ConfNames	4 triples
Links to LC3M	11 triples
Links to MapIn-west	3 triples
Links to ORCID	1 triples
Links to Wikipedia	87 triples
Links to vitar	15 triples

<https://lod-cloud.net/dataset/orlando-womens-writing-linked-open-data-set>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 130

130

The screenshot shows the CDC DATE website interface. It features a header with the CDC DATE logo and navigation tabs. The main content area is divided into several sections, each with a title and a brief description. The sections include:

- CDC**: The Foundation, Center for Data Science, CDC is a non-profit research institute for the study of disease and the development of public health programs.
- CDC-OUTREACH**: The Outreach program is a series of webinars, webinars, and other educational activities designed to help organizations and individuals understand the importance of data science in public health.
- CDC-OUTREACH**: The Outreach program is a series of webinars, webinars, and other educational activities designed to help organizations and individuals understand the importance of data science in public health.
- CDC-OUTREACH**: The Outreach program is a series of webinars, webinars, and other educational activities designed to help organizations and individuals understand the importance of data science in public health.
- CDC-OUTREACH**: The Outreach program is a series of webinars, webinars, and other educational activities designed to help organizations and individuals understand the importance of data science in public health.

<http://dati.cdc.gov/indiceEN.html>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 131

131

The screenshot shows the LOD Navigator interface. It features a map of Italy with a network of nodes and edges representing the movements of Italian Shoah victims. The interface includes a search bar, a list of properties, and a table of data. The table is titled "Table 3: Overview of victims on the basis of the death_description property".

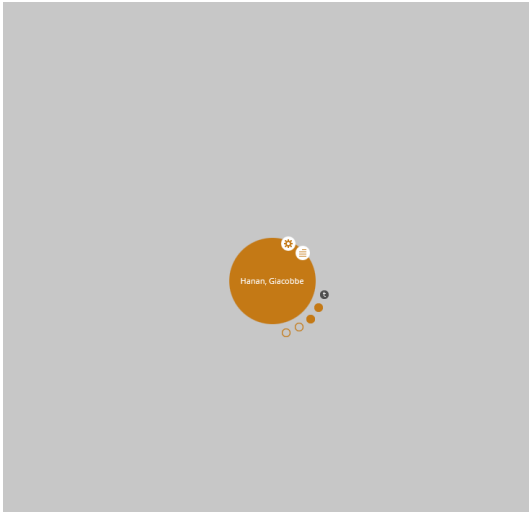
death_description	QUANTITY
Death in extermination camp	7,551
Survivors of Shoah	1,057
Death in massacre	202
Death in roundup	70
Unknown	28
Death en route to camp	15
Committal suicide	10
Death of hardships and privations	6
Killed in escape attempt	6
Missing	6
Killed during action	1
TOTAL	8,712

Table 3: Overview of victims on the basis of the death_description property

Figure 3: Screenshot of the interface: movements of Jews having a craft occupation and survived to the Shoah

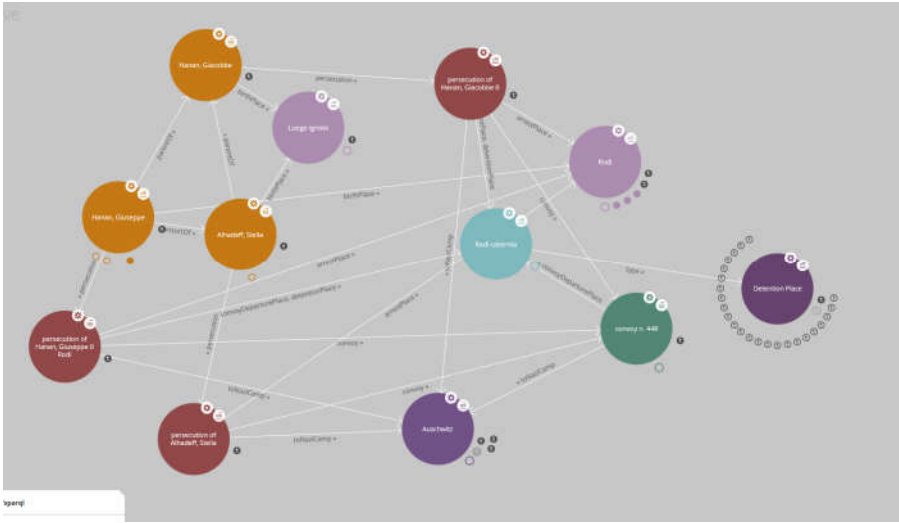
Rachele Sprugnoli, Moretti Giovanni, and Tonelli Sara, 2019, LOD Navigator: Tracing Movements of Italian Shoah Victims. *Umanistica Digitale*, 3(4). <https://doi.org/10.6092/issn.2532-8816/9050>

132



<http://dati.cdec.it/lodlive/?http://dati.cdec.it/lod/shoah/person/3423>


133



<http://dati.cdec.it/lodlive/?http://dati.cdec.it/lod/shoah/person/3423>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 134

134



CD EC

Hanan, Giacobbe

<http://dati.cdec.it/lod/shoah/person/3423>

rdfs:label Hanan, Giacobbe

skos:notation [Il libro della memoria: gli storie deportati dal'Italia, 1943-1945 / Liliana Picciotto; ricerca della Fondazione Centro di documentazione ebraica contemporanea. - Ed. 2002 altri nomi \(travati\) / Milano: Mursia, 2002, pp. 77-89](#)

cc:provenance [EB Digitali](#)

cc:type [ultima della Shoah](#)

cc:licenseURL [CC BY](#)

cc:attributionURL [Dati](#)

cc:attribution

cc:attributionDescription

cc:attributionDescription

cc:attributionName [Data](#)

cc:biography

foaf:familyName [Hanan](#)

foaf:gender [M](#)

cc:keywords [deportato dall'Italia](#)

foaf:firstName [Giacobbe](#)

cc:type [foaf:Person](#)

skos:parentTerm [http://dati.cdec.it/lod/shoah/person/3423](#)
[↳ parent term of Hanan, Giacobbe G](#)

cc:memberOf [http://dati.cdec.it/lod/shoah/area/Lugli_ignori](#)
[↳ Lugli Ignori](#)

skos:memberOf [http://dati.cdec.it/lod/shoah/area/Lugli_ignori](#)
[↳ Lugli Ignori](#)

skos:parentTerm [http://dati.cdec.it/lod/shoah/person/3423](#)
[↳ parent term of 2 instances](#)

<http://dati.cdec.it/lod/shoah/person/3423/html>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 135

135

CD EC

Salmoni, Renato

<http://dati.cdec.it/lod/shoah/person/7645>

rdfs:label [Salmoni, Renato](#)

skos:notation [Il libro della memoria: gli storie deportati dal'Italia, 1943-1945 / Liliana Picciotto; ricerca della Fondazione Centro di documentazione ebraica contemporanea. - Ed. 2002 altri nomi \(travati\) / Milano: Mursia, 2002, pp. 77-89](#)

cc:provenance [EB Digitali](#)

cc:licenseURL [CC BY](#)

cc:attributionURL [Dati](#)

cc:attribution

cc:attributionDescription

cc:attributionDescription

cc:attributionName [Data](#)

foaf:familyName [Salmoni](#)

foaf:gender [M](#)

cc:keywords [deportato dall'Italia](#)

foaf:firstName [Renato](#)

cc:type [foaf:Person](#)

skos:parentTerm [http://dati.cdec.it/lod/shoah/person/7645](#)
[↳ parent term of Salmoni, Renato / 1/12/1913](#)
[↳ parent term of Salmoni, Renato / 1/12/1913](#)

cc:memberOf [http://dati.cdec.it/lod/shoah/person/8871](#)
[↳ Salmoni, Dino](#)

cc:memberOf [http://dati.cdec.it/lod/shoah/person/7644](#)
[↳ Salmoni, Gilberto Raffaele](#)

cc:memberOf [http://dati.cdec.it/lod/shoah/person/9529472](#)
[↳ Anni Cento \(1900\)](#)

<http://dati.cdec.it/lod/shoah/person/7645/html>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 136

136

CONCLUSION

We have examined some research projects which take advantage of possibilities afforded by SW/LD technology. Our goal was to pinpoint aspects of the research methodology which are impacted by the technology. Our examination of these cases has revealed three main points of impact. The first is the epistemological foundations of the research; the focus is on individual entities in the research area, leading to what we have called the "atomization" of research. This in turn favours analytical skills in the researcher rather than his or her ability to synthesize and abstract away from specific phenomena. In addition, all research phenomena are reified – raised to a concrete status in their representation by URIs, regardless of their original characteristics. Secondly, and unsurprisingly, it is the data analysis phase which is affected the most; it relies more heavily on technical skills during data discovery and processing. It obscures the distinction between data and metadata, thus requiring even more analytical skills from the researcher to diligently differentiate between the two as necessary. Thirdly, we have shown how the model-building aspect of research is affected by SW/LD-driven approaches, in the sense that existing models of knowledge (i.e. existing documents) require a transcoding into this new format in order to facilitate further research.


The scale of the efforts to produce LD datasets is such that it cannot be ignored, and research in SSH must take it into account, as so much useful data are now available. Transformations to the conduct of research is inevitable. But it is important to understand how this shift may affect research methodology. What is needed then is new solutions for new challenges.

□ *Lyne Da Sylva, 2018. "Towards linked data: Some consequences for researchers in the social sciences and humanities". *Proceedings of the Association for Information Science and Technology*, 55(1), 94–103.
<https://doi.org/10.1002/pr2.2018.14505501011>, page 102

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 139

139

Overview



The *Open Science Framework* has been developed by the *Centre for Open Science* as a platform to facilitate collaboration between researchers and to facilitate open science practices throughout the entire project lifecycle. It is free to use and you can add collaborators from all over the world to work together on projects.

It facilitates collaboration through:

- controlled access to projects that can be set as at private, at public or a mixture between the two.
- providing a structure through which files can be managed, worked upon, and shared within teams.
- connection to other online platforms that can facilitate data generation, analysis, sharing, and publication.

It facilitates open science and reproducibility through:

- the ability to pre-register, and make public, protocols, analysis plans, research outputs, and datasets.
- the ability to generate persistent digital identifiers (Digital Object Identifiers (DOIs)) for datasets and pre-prints.
- the ability to license outputs and datasets to govern use by other researchers.
- the ability to archive datasets in a way that is consistent with University, funder, and journal data policies.

[Open Science Framework homepage](#)

[Open Science Framework FAQs](#)

[Open Science Framework Guides](#)

<https://library.bath.ac.uk/c.php?g=665432&p=4832004>

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 140

140

**FINAL
THOUGHTS
LINKED
(OPEN)
DATA**

- Concept of "Open"
- Transparency
- Reproducibility
- Interconnectedness of scholarship or business need
- Knowledge discovery
- Quality, Data, Links, Ontologies/Vocabularies
- Assessment and Validation

© Seamus Ross, FI at UofT, Lecture Notes: OD, LOD & Semantic Web (4 Jun 21) 141

141