

CSV-LD: Spreadsheet-based Linked Data

Donny Winston

csv,conf,v6

2021-05-04T16:00Z/PT20M

Linked Data is a formal way to identify context within data

- **formal** – can be interpreted mechanically, via standards
- **context** – the concepts and relationships of the subject matter (ontology) and/or data structure (schema)
- **within** – within the same artifact (file/object) as the data

Expectations for **Linked Data**:

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 RDF
4. Include links to other URIs, so that they can discover more things

A JSON-LD document is a JSON document that includes its context

```
{ "@context": "https://json-ld.org/contexts/person.jsonld",  
  "@id": "http://dbpedia.org/resource/John Lennon",  
  "name": "John Lennon",  
  "born": "1940-10-09",  
  "spouse": "http://dbpedia.org/resource/Cynthia Lennon" }
```

Meanwhile, at <https://json-ld.org/contexts/person.jsonld> ...

```
{ "@context": { ...,  
  "xsd": "http://www.w3.org/2001/XMLSchema#",  
  "name": "http://xmlns.com/foaf/0.1/name",  
  "born": { "@id": "http://schema.org/birthDate", "@type": "xsd:date" },  
  "spouse": { "@id": "http://schema.org/spouse", "@type": "@id" },  
  ... }
```

"xsd:date" means "http://www.w3.org/2001/XMLSchema#date".

CSV on the Web (CSVW) helps you build sidecars for spreadsheets



Radosław Drożdżewski / CC-BY-SA-4.0





- Sidecar – a functional addition
 - Motorcycle sidecar: carry best friends
 - Kubernetes sidecar: support related work
 - ~~Unstructured README file~~
- Given mydata.csv:
 - Can serve JSON-LD sidecar at `mydata.csv-metadata.json`
 - Use vocabulary terms from <https://www.w3.org/ns/csvw#> to provide extra information

Linking and Packaging are complementary

- Linked Data: link out to context and other data.
- Data Package: link in to contained data.
- Example of packaging: the Frictionless Data specs
 - A container format, i.e. “Docker for data”, to describe and package a collection of data.
 - Describe the dataset (package spec), describe the structure (e.g. table schema spec / CSV data descriptor), and include the data resource (e.g. CSV file).
- Examples of linking:
 - embedded: @context field of JSON-LD
 - not embedded: -metadata.json sidecar, or link header, of CSVW

Barcode labels for columns

“header” for
delimiter-separated-values
file

				
	045930548112	611422350426	957982615339	276359624526
# methane molecule (in angstroms)				
C	0.000000	0.000000	0.000000	0.000000
H	0.000000	0.000000	0.000000	1.089000
H	1.026719	0.000000	0.000000	-0.363000
H	-0.513360	-0.889165	-0.889165	-0.363000
H	-0.513360	0.889165	0.889165	-0.363000

“Go to definition” for columns

header for
delimiter-separated-values
file

	A	B	C	D	
1	#formatVersion	https://ns.csv-ld.org/2021/05/csv-ld/core			
2	#vocabBase	https://ns.csv-ld.org/2021/05/csv-ld/demo#			
3	# methane molecule (in angstroms)				
4	atom	x	y	z	
5	C	0.000000	0.000000	0.000000	
6	H	0.000000	0.000000	1.089000	
7	H	1.026719	0.000000	-0.363000	
8	H	-0.513360	-0.889165	-0.363000	
9	H	-0.513360	0.889165	-0.363000	
10					

It is still a delimiter-separated-values file

	A	B	C	D
1	#formatVersion	https://ns.csv-ld.org/2021/05/csv-ld/core		
2	#vocabBase	https://ns.csv-ld.org/2021/05/csv-ld/demo#		

- First line:
`#formatVersion,https://ns.csv-ld.org/2021/05/core`
- A CSV-LD processor uses the first line to infer
 - `https://www.w3.org/ns/csvw#commentPrefix`
 - `https://www.w3.org/ns/csvw#delimiter`
- “#” is the default prefix for lines that are comments
- “,” is the default delimiter

“What is going on here?” “Follow your nose.”

	A	B	C	D	
1	#formatVersion	https://ns.csv-ld.org/2021/05/csv-ld/core			
2	#vocabBase	https://ns.csv-ld.org/2021/05/csv-ld/demo#			

Goals of `#formatVersion` directive on first line:

1. Get data consumer to learn more – “What is this format? Cool, a link...”
2. Help CSV-LD processors know what to expect

Reuse header rows with a “vocab-base”

- #vocabBase

	A	B	C	D	
1	#formatVersion	https://ns.csv-ld.org/2021/05/csv-ld/core			
2	#vocabBase	https://ns.csv-ld.org/2021/05/csv-ld/demo#			
3	# methane molecule (in angstroms)				
4	atom	x	y	z	
5	C	0.000000	0.000000	0.000000	
6	H	0.000000	0.000000	1.089000	

<https://ns.csv-ld.org/2021/05/csv-ld/demo#atom>

or

https://ns.csv-ld.org/2021/05/csv-ld/demo#entity_6428 with “atom” as label

Reuse vocabularies, succinctly

- #prefix

	A	B	C	D	E
1	#formatVersion	https://ns.csv-ld.org/2021/05/csv-ld/core			
2	#vocabBase	https://ns.csv-ld.org/2021/05/csv-ld/demo#			
3	#prefix	iupac:	http://www.example.org/ns/iupac#		
4	# methane molecule (in angstroms)				
5	iupac:atom	x	y	z	
6	C	0.000000	0.000000	0.000000	
7	H	0.000000	0.000000	1.089000	

<http://www.example.org/ns/iupac#atom> values should be valid IUPAC symbols for atomic elements

Include other metadata via RDF statements

	A	B	C	D	E		
1	#formatVersi	https://ns.csv-ld.org/2021/05/csv-ld/core					
2	#vocabBase	https://ns.csv-ld.org/2021/05/csv-ld/demo#					
3	#prefix	csvld:	https://ns.csv-ld.org/2021/05/csv-ld/core#				
4	#prefix	iupac:	http://www.example.org/ns/iupac#				
#, predicate, object	→	5	#	csvld:csvw	http://csvw.example.org/molXYZ.json		
#_...	→	6	# methane molecule (in angstroms)				
#id	→	7	#id	http://measurements.example.org/2003/CHEBI_16183			
		8	iupac:atom	x	y	z	
		9	C	0.000000	0.000000	0.000000	
		10	H	0.000000	0.000000	1.089000	

link to CSV on the Web (CSVW)
JSON-LD metadata

Vocabulary terms are resolvable HTTP URIs

- they need to be accessible on the web
- perhaps a data steward on your team can help
- or one in your organization

```
<!DOCTYPE html>
<html lang="en">
<head>
  <meta charset="UTF-8">
  <title>CSV-LD Core Vocabulary</title>
  <link rel="alternate" type="text/turtle" href="core.ttl">
  <style type="text/css">
    dt { font-weight: bold; text-decoration: underline dotted; }
  </style>
</head>
<body>
  <dl>
    <dt id="#formatVersion">formatVersion</dt>
    <dd>The version of the CSV-LD format that a processor should assume for this spreadsheet.</dd>
    <dt id="#vocabBase">vocabBase</dt>
    <dd>The "vocabulary base" for terms in this spreadsheet.</dd>
    <dt id="#prefix">prefix</dt>
    <dd>Defines a prefix so that common URI bases can be prefixes of term URIs.</dd>
    <dt id="#id">id</dt>
    <dd>The URI for this sheet.</dd>
    <dt id="#csvw">csvw</dt>
    <dd>The URI for JSON-LD CSVW metadata for this sheet.</dd>
  </dl>
</body>
</html>
```

```
@prefix csvld: <http://ns.csv-ld.org/2021/05/csv-ld/core#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .

<http://ns.csv-ld.org/2021/05/csv-ld/core> a owl:Ontology ;
  dc:title "CSV-LD Core Vocabulary" ;
  owl:versionIRI <http://ns.csv-ld.org/2021/05/csv-ld/core> .

csvld:formatVersion
  rdfs:label "formatVersion" ;
  rdfs:comment "The version of the CSV-LD format that a processor should assume for this spreadsheet." .

csvld:vocabBase
  rdfs:label "vocabBase" ;
  rdfs:comment ""The "vocabulary base" for terms in this spreadsheet."" .

csvld:prefix
  rdfs:label "prefix" ;
  rdfs:comment "Defines a prefix so that common URI bases can be prefixes of term URIs." .

csvld:id
  rdfs:label "id" ;
  rdfs:comment "The URI for this sheet." .

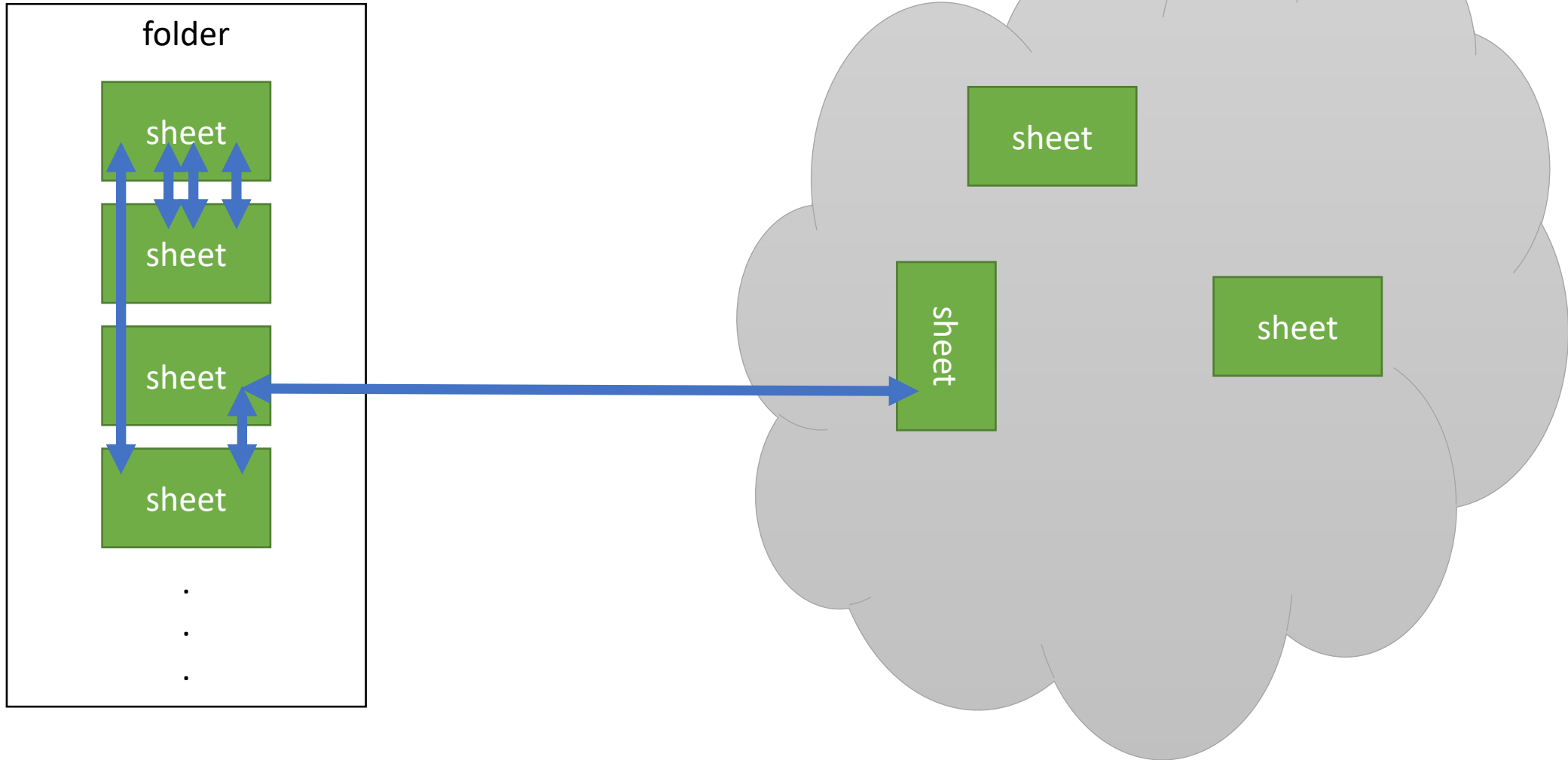
csvld:csvw
  rdfs:label "csvw" ;
  rdfs:comment "The URI for JSON-LD CSVW metadata for this sheet." .
```

HTML and RDF forms of
<https://ns.csv-ld.org/2021/05/csv-ld/core>
(source at <https://github.com/csv-ld/ns>)

Near term: CSV-LD to validate and to collect

- may validate columns in isolation
- may validate rows as “entities” with required and optional columns
 - columns may be sub-properties of the properties sought for an entity
 - “value kind” hierarchies
 - independent “generate” and “check” procedures to suggest and ensure valid values
 - entity validation may collect and consider all entities in the sheet
- may convert CSV-LD to JSON-LD or other RDF graph serializations

Someday: CSV-LD to unify and to discover



After this talk, come find me

- Near term: in the csv,conf Slack.
- <https://github.com/csv-ld/ns>. Apache 2.0. Raise issues!
- mail@donnywinston.com

- Materials Research Data Alliance (MaRDA)
 - <https://www.marda-alliance.org/>
 - Data Dictionaries Working Group (WG) – open discussion at <https://matsci.org/marda>
- National Microbiome Data Collaborative (NMDC)
 - <https://microbiomedata.org/>
 - <https://github.com/microbiomedata/>
- The [Recurse Center](#)

Thank you!

Questions?