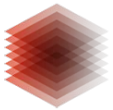
Integrating data and analysis

On bridging data publishers and computational environments

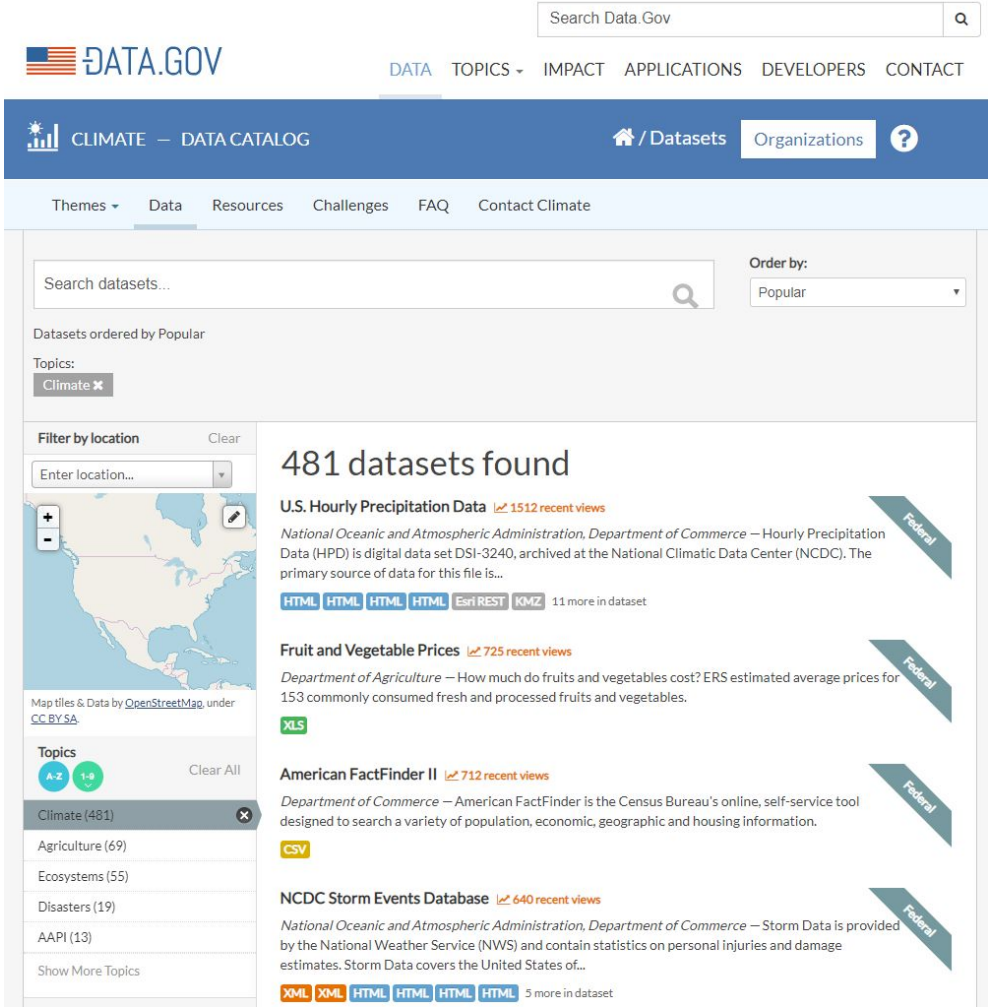
Markus Stocker (@envinf), Uwe Schindler, Robert Huber

markus.stocker@tib.eu, uschindler@pangaea.de, rhuber@uni-bremen.de

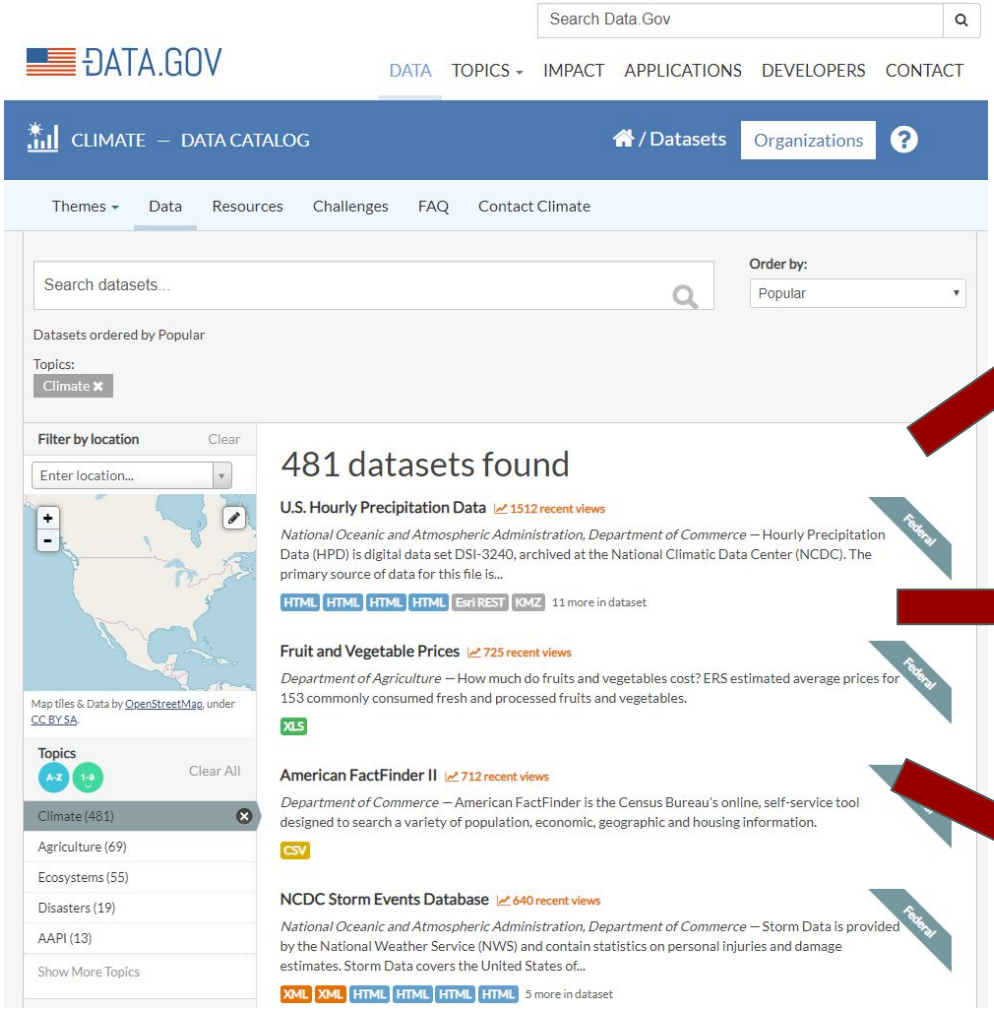


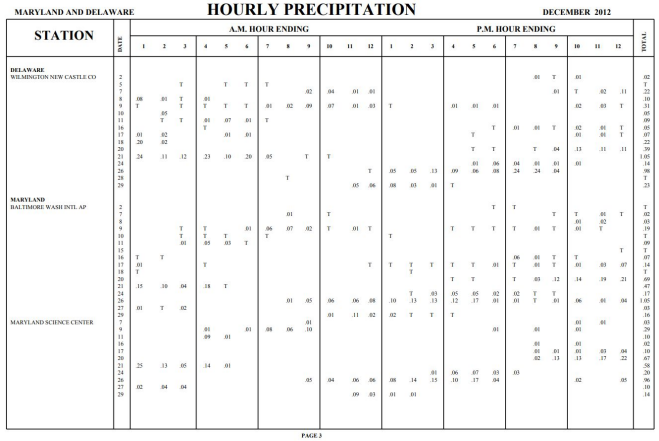
TIB LEIBNIZ INFORMATION CENTRE FOR SCIENCE AND TECHNOLOGY UNIVERSITY LIBRARY

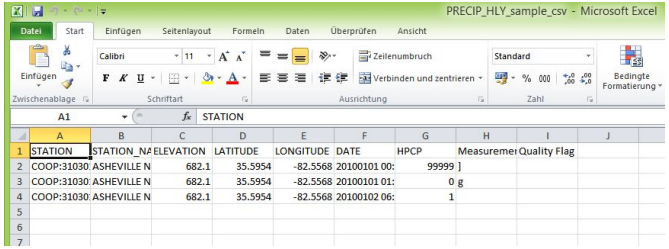
Universität Bremen



Catalogs were a great first step but ...


Search Data.Gov
DATA.GOV
DATA
TOPICS
IMPACT
APPLICATIONS
DEVELOPERS
CONTACT
CLIMATE — DATA CATALOG
/ Datasets
Organizations
Themes
Data
Resources
Challenges
FAQ
Contact Climate
Search datasets...
Order by:
Popular
Datasets ordered by Popular
Topics:
Climate
Filter by location
Clear
Enter location...
Map tiles & Data by OpenStreetMap, under CC BY SA
Topics
Clear All
Climate (481)
Agriculture (69)
Ecosystems (55)
Disasters (19)
AAPI (13)
Show More Topics
481 datasets found
U.S. Hourly Precipitation Data
1512 recent views
National Oceanic and Atmospheric Administration, Department of Commerce — Hourly Precipitation Data (HPD) is digital data set DSI-3240, archived at the National Climatic Data Center (NCDC). The primary source of data for this file is...
HTML HTML HTML HTML Esri REST KMZ 11 more in dataset
Federal
Fruit and Vegetable Prices
725 recent views
Department of Agriculture — How much do fruits and vegetables cost? ERS estimated average prices for 153 commonly consumed fresh and processed fruits and vegetables.
XLS
Federal
American FactFinder II
712 recent views
Department of Commerce — American FactFinder is the Census Bureau's online, self-service tool designed to search a variety of population, economic, geographic and housing information.
CSV
NCDC Storm Events Database
640 recent views
National Oceanic and Atmospheric Administration, Department of Commerce — Storm Data is provided by the National Weather Service (NWS) and contain statistics on personal injuries and damage estimates. Storm Data covers the United States of...
XML XML HTML HTML HTML HTML 5 more in dataset
Federal


MARYLAND AND DELAWARE
HOURLY PRECIPITATION
DECEMBER 2012
STATION
A.M. HOUR ENDING
P.M. HOUR ENDING
TOTAL
DELAWARE
WILMINGTON NEW CASTLE CO
MARYLAND
BALTIMORE WASH INTL AP
MARYLAND SCIENCE CENTER
PAGE 3


PRECIP_HLY_sample_csv - Microsoft Excel
Datei
Start
Einfügen
Seitenlayout
Formeln
Daten
Überprüfen
Ansicht
Einfügen
Zwischenablage
Schriftart
Ausrichtung
Zahl
Zeilenumbruch
Verbinden und zentrieren
Standard
Bedingte Formatierung
A1
STATION
STATION
STATION_NA
ELEVATION
LATITUDE
LONGITUDE
DATE
HPCP
Measureme
Quality Flag
COOP:31030
ASHEVILLE N
682.1
35.5954
-82.5568
20100101 00:
99999
20100101 01:
0
20100102 06:
1


Index of /pub/data/swdi/stormevents/csvfiles/
[parent directory]
Name
Size
Date Modified
legacy/
5/14/14, 2:00:00 AM
README
2.0 kB
5/14/14, 2:00:00 AM
Storm-Data-Export-Format.docx
23.3 kB
5/6/14, 2:00:00 AM
StormEvents_details-ftp_v1.0_d1950_c20170120.csv.gz
10.3 kB
1/20/17, 1:00:00 AM
StormEvents_details-ftp_v1.0_d1951_c20160223.csv.gz
11.7 kB
2/24/16, 1:00:00 AM
StormEvents_details-ftp_v1.0_d1952_c20170619.csv.gz
12.3 kB
6/19/17, 2:00:00 AM
StormEvents_details-ftp_v1.0_d1953_c20160223.csv.gz
21.3 kB
2/24/16, 1:00:00 AM
StormEvents_details-ftp_v1.0_d1954_c20160223.csv.gz
25.6 kB
2/24/16, 1:00:00 AM

We can do better ...

Curl

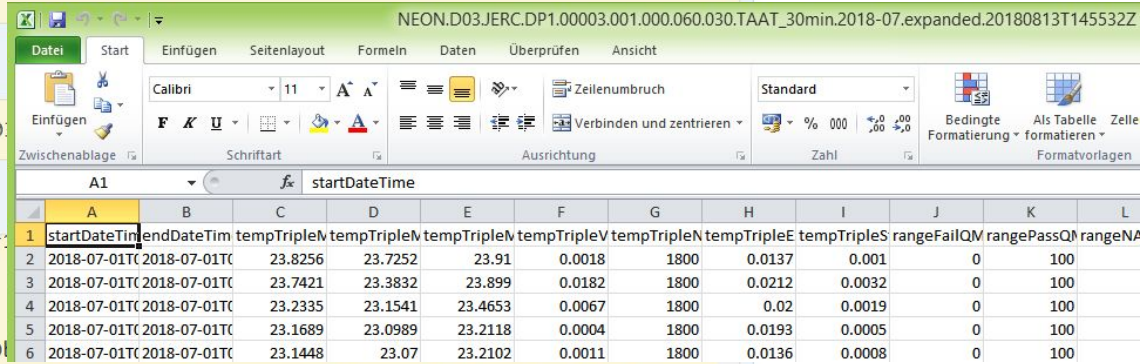
```
curl -X GET --header 'Accept: application/json' 'http://data.neonscience.org/api/v0/data/DP1.00003.001/JERC/2018-07'
```

Request URL

```
http://data.neonscience.org/api/v0/data/DP1.0000
```

Response Body

```
-Signature=f9da179fbed315a963d1f211b9d29dba23af
      },
      {
        "crc32": "f14d3818cdf8e83b675038d654e70
        "name": "NEON.D03.JERC.DP1.00003.001.000.060.030.TAAT_30min.2018-07.expanded.20180813T145532Z.csv",
        "size": "485127",
        "url": "https://neon-prod-pub-1.s3.data.neonscience.org/NEON.DOM.SITE.DP1.
.001/PROV/JERC/20180701T000000--20180801T000000/expanded/NEON.D03.JERC.DP1.00003.001.000.06
.TAAT_30min.2018-07.expanded.20180813T145532Z.csv?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20180919T161158Z&X-Amz-SignedHeaders=host&X-Amz-Expires=3600&X-Amz-Credential=pub-internal-read%2F20180919%2Fus-west-2%2Fs3%2Faws4_request&X-Amz-Signature=1558e7fb29fc42aeba66ebd39d097f34c88932598d1f33575cd3b7f645363daf"
      },
      {
        "crc32": "6c3ae69817dc5d9848971fe2bb020940",
```



	A	B	C	D	E	F	G	H	I	J	K	L
1	startDateTim	endDateTim	tempTripleM	tempTripleM	tempTripleM	tempTripleV	tempTripleN	tempTripleE	tempTripleS	rangeFailQM	rangePassQM	rangeNA
2	2018-07-01T0	2018-07-01T0	23.8256	23.7252	23.91	0.0018	1800	0.0137	0.001	0	100	
3	2018-07-01T0	2018-07-01T0	23.7421	23.3832	23.899	0.0182	1800	0.0212	0.0032	0	100	
4	2018-07-01T0	2018-07-01T0	23.2335	23.1541	23.4653	0.0067	1800	0.02	0.0019	0	100	
5	2018-07-01T0	2018-07-01T0	23.1689	23.0989	23.2118	0.0004	1800	0.0193	0.0005	0	100	
6	2018-07-01T0	2018-07-01T0	23.1448	23.07	23.2102	0.0011	1800	0.0136	0.0008	0	100	

Parameter(s):

#	Name	Short Name	Unit	Principal Investigator	Method	Comment
1	DEPTH, sediment/rock	Depth	m			Geocode
2	Alkenone, unsaturation index UK'37	UK'37		Müller, Peter J	Calculated from C37 alkenones (Prahl & Wakeham, 1987)	
3	Sea surface temperature, annual mean	SST (1-12)	°C	Müller, Peter J	Calculated from UK'37 (Prahl et al., 1988)	

License: Creative Commons Attribution 3.0 Unported

Size: 64 data points

Data

Download dataset as tab-delimited text *(use the following character encoding:* windows-1252: Windows Western)

1 Depth [m]	2 UK'37	3 SST (1-12) [°C]
0.03	0.667	18.50
0.08	0.656	18.10
0.13	0.652	18.00
0.18	0.663	18.40
0.23	0.652	18.00
0.28	0.656	18.10
0.33	0.658	18.20
0.38	0.662	18.30
0.43	0.656	18.10
0.48	0.658	18.20
0.53	0.660	18.30
0.58	0.675	18.70
0.63	0.669	18.50
0.68	0.664	18.40
0.73	0.658	18.20
0.78	0.652	18.00
0.83	0.643	17.80
0.88	0.614	16.90
0.93	0.614	16.90
0.98	0.603	16.60
1.03	0.607	16.70
1.08	0.600	16.50
1.13	0.605	16.60

PANGAEA.

DOI

https://doi.org/10.1594/PANGAEA.80968

```
ms@elephant:~$ curl -D- -H "Accept: text/tab-separated-values" -L https://doi.org/10.1594/PANGAEA.80968
HTTP/2 302
date: Wed, 19 Sep 2018 15:56:51 GMT
content-type: text/html;charset=utf-8
content-length: 183
set-cookie: __cfduid=db883aff5ac15916de1657fa2092dea341537372611; expires=Thu, 19-Sep-19 15:56:51 GMT; p
expires: Wed, 19 Sep 2018 16:13:47 GMT
location: https://data.datacite.org/10.1594%2FPANGAEA.80968
vary: Accept
expect-ct: max-age=604800, report-uri="https://report-uri.cloudflare.com/cdn-cgi/beacon/expect-ct"
server: cloudflare
cf-ray: 45cd3ea6687297b0-FRA

HTTP/2 303
date: Wed, 19 Sep 2018 15:56:52 GMT
content-type: text/html; charset=utf-8
location: https://doi.pangaea.de/10.1594/PANGAEA.80968
set-cookie: AWSALB=jCe/M/m2+PFFoaqgvMw8y4RAyEmMAJB2vryZHTWuI3hQSyfl6TyWheOEVDqN1+1rcixPn4Fy9FcSIwo4HhzIG
status: 303 See Other
cache-control: no-cache
vary: Accept-Encoding, Origin
x-request-id: c5e08e6a-7d1e-48e3-b7dd-586c580b1192
accept: text/tab-separated-values
x-runtime: 0.149856
x-powered-by: Phusion Passenger 5.3.4
server: nginx/1.14.0 + Phusion Passenger 5.3.4

HTTP/1.1 200 OK
Server: PANGAEA/1.0
Date: Wed, 19 Sep 2018 15:56:52 GMT
Transfer-encoding: chunked
Vary: Accept
Link: <https://doi.org/10.1594/PANGAEA.80968>;rel="cite-as", <https://doi.pangaea.de/10.1594/PANGAEA.809
://doi.pangaea.de/10.1594/PANGAEA.80968?format=citation_bibtex>;rel="describedby";type="application/x-bi
Content-disposition: attachment; filename=IOW226660-5_UK37_SST.tab
X-robots-tag: noindex,nofollow,noarchive
Content-type: text/tab-separated-values;charset=UTF-8
X-ua-compatible: IE=Edge
X-content-type-options: nosniff
Strict-transport-security: max-age=31536000

/* DATA DESCRIPTION:
Citation:       Mollenhauer, Gesine; Müller, Peter J (2002): UK37 and alkenone sea surface temperatures
        In supplement to: Mollenhauer, Gesine; Eglinton, Timothy I; Ohkouchi, Naohiko; Schneider, Ralph
tps://doi.org/10.1016/S0016-7037(03)00168-6
Related to:     Mollenhauer, Gesine (2002): Organic carbon accumulation in the South Atlantik Ocean: Sea
Project(s):     Geosciences, University of Bremen (GeoB) (URI: http://www.geo.uni-bremen.de/page.php?lan
Coverage:       LATITUDE: -24.108000 * LONGITUDE: 12.765000
        DATE/TIME START: 2000-08-08T00:00:00 * DATE/TIME END: 2000-08-08T00:00:00
        MINIMUM DEPTH, sediment/rock: 0.03 m * MAXIMUM DEPTH, sediment/rock: 1.58 m
Event(s):       IOW226660-5 (M48/2_359) * LATITUDE: -24.108000 * LONGITUDE: 12.765000 * DATE/TIME: 2000-
f,_1986)) * DEVICE: Gravity corer (Kiel type) (SL)
Parameter(s):   DEPTH, sediment/rock [m] (Depth) * GEOCODE
        Alkenone, unsaturation index UK'37 (UK'37) * PI: Müller, Peter J * METHOD: Calculated from C37 a
        Sea surface temperature, annual mean [°C] (SST (1-12)) * PI: Müller, Peter J * METHOD: Calculate
License:        Creative Commons Attribution 3.0 Unported (CC-BY)
Size:   64 data points
*/
Depth [m]       UK'37   SST (1-12) [°C]
0.03    0.667   18.50
0.08    0.656   18.10
0.13    0.652   18.00
0.18    0.663   18.40
0.23    0.652   18.00
0.28    0.656   18.10
0.33    0.658   18.20
0.38    0.662   18.30
0.43    0.656   18.10
0.48    0.658   18.20
```

1. curl -D- -H
   “Accept: text/tab-separated-values” -L
   https://doi.org/10.1594/PANGAEA.80968
2.
3.

Depth [m]	UK'37	SST (1-12) [°C]
0.03	0.667	18.50
0.08	0.656	18.10
0.13	0.652	18.00
0.18	0.663	18.40
0.23	0.652	18.00
0.28	0.656	18.10
0.33	0.658	18.20
0.38	0.662	18.30
0.43	0.656	18.10
0.48	0.658	18.20

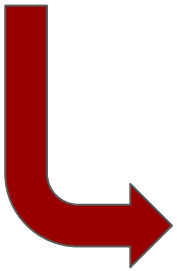
Observations

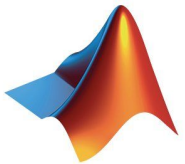
- DOI based access is great, should perhaps be LCD: Why?
- In practice, we struggle with large differences in data access
- Is there a good reason for this heterogeneity?
- APIs are great, but data are delivered to local hard drive as files
- What we really want is ...

… data loaded into a computational environment

http://data.neonscience.org/api/v0/data/DP1.00003.001/JERC/2018-07

https://doi.org/10.1594/PANGAEA.80968





Approaches

Still, data are not immediately processable

`getPackage()` can be used to pull a single zip file (all the data for a single data product by site by month combination) using the NEON API.

```
# Plant phenology observations from the Jornada LTER site, May 2017
getPackage(dpID = "DP1.10055.001", site_code = "JORN", year_month = "2017-05", package = "basic")
```

https://github.com/NEONScience/NEON-utilities/tree/master/neonUtilities

```
> library(pangaear)
> d <- pg_data("10.1594/PANGAEA.80968")
Downloading 1 datasets from 10.1594/PANGAEA.80968
Processing 1 files
> d[[1]]$data
# A tibble: 32 x 3
   `Depth [m]` `UK'37` `SST (1-12) [°C]`
         <dbl>   <dbl>             <dbl>
 1        0.03   0.667              18.5
 2        0.08   0.656              18.1
 3        0.13   0.652              18
 4        0.18   0.663              18.4
 5        0.23   0.652              18
 6        0.28   0.656              18.1
 7        0.33   0.658              18.2
 8        0.38   0.662              18.3
 9        0.43   0.656              18.1
10        0.48   0.658              18.2
# ... with 22 more rows
```

https://tibhannover.github.io/2018-07-09-FAIR-Data-and-Software/FAIR-remix-PANGAEA/

```python
from pandata.pandataset import PanDataSet

ds = PanDataSet('10.1594/PANGAEA.80968')
```

```python
ds.data[["Depth", "UK'37", "SST (1-12)"]]
```

	Depth	UK'37	SST (1-12)
0	0.03	0.667	18.5
1	0.08	0.656	18.1
2	0.13	0.652	18.0
3	0.18	0.663	18.4
4	0.23	0.652	18.0
5	0.28	0.656	18.1
6	0.33	0.658	18.2
7	0.38	0.662	18.3
8	0.43	0.656	18.1
9	0.48	0.658	18.2
10	0.53	0.660	18.3

PANGAEA.

python™

https://github.com/huberrob/panpython/

Easy for CSV/TSV but ...

```
@prefix dul: <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#> .
@prefix geosparql: <http://www.opengis.net/ont/geosparql#> .
@prefix gn: <http://www.geonames.org/ontology#> .
@prefix lode: <http://linkedevents.org/ontology/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sf: <http://www.opengis.net/ont/sf#> .
@prefix smear: <http://avaa.tdata.fi/web/smart/smear/> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://avaa.tdata.fi/web/smart/smear/2c3514176ca67a77a99292cbb4b6a3ae> a lode:Event ;
    smear:hasClassification smear:ClassIa ;
    lode:atPlace <http://sws.geonames.org/656888/> ;
    lode:atTime <http://avaa.tdata.fi/web/smart/smear/0cf796b1a1b4fb5563a52fb2b5ec6093> ;
    lode:inSpace <http://avaa.tdata.fi/web/smart/smear/7f885190eb43154e01c97f814b287a4b> .

<http://avaa.tdata.fi/web/smart/smear/0cf796b1a1b4fb5563a52fb2b5ec6093> a time:Interval ;
    time:hasBeginning smear:f72d5d2e62f9747161bb9fd127a64590 ;
    time:hasEnd smear:ffade79921356c06cbdcf1c1c8fdb4dc .

<http://avaa.tdata.fi/web/smart/smear/7f885190eb43154e01c97f814b287a4b> a sf:Point,
        wgs84:SpatialThing ;
    geosparql:asWKT "POINT (24.29077 61.84562)"^^geosparql:wktLiteral .

smear:ClassIa a smear:Classification ;
    rdfs:label "Class Ia"^^xsd:string ;
    rdfs:comment "Very clear and strong event"^^xsd:string .

smear:f72d5d2e62f9747161bb9fd127a64590 a time:Instant ;
    time:inXSDDateTime "2013-04-04T10:30:00+03:00"^^xsd:dateTime .

smear:ffade79921356c06cbdcf1c1c8fdb4dc a time:Instant ;
    time:inXSDDateTime "2013-04-04T12:00:00+03:00"^^xsd:dateTime .

<http://sws.geonames.org/656888/> a gn:Feature,
        dul:Place ;
    gn:countryCode "FI"^^xsd:string ;
    gn:locationMap <http://www.geonames.org/656888/hyytiaelae.html> ;
    gn:name "Hyytiälä"^^xsd:string ;
    wgs84:lat 6.184562e+01 ;
    wgs84:long 2.429077e+01 .
```

```
@prefix dul: <http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#> .
@prefix geosparql: <http://www.opengis.net/ont/geosparql#> .
@prefix gn: <http://www.geonames.org/ontology#> .
@prefix lode: <http://linkedevents.org/ontology/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sf: <http://www.opengis.net/ont/sf#> .
@prefix smear: <http://avaa.tdata.fi/web/smart/smear/> .
@prefix time: <http://www.w3.org/2006/time#> .
@prefix wgs84: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

<http://avaa.tdata.fi/web/smart/smear/2e3514176ca67a77a99292cbb4b6a3ae> a lode:Event ;
    smear:hasClassification smear:ClassIa ;
    lode:atPlace <http://sws.geonames.org/656888/> ;
    lode:atTime <http://avaa.tdata.fi/web/smart/smear/0cf796b1a1b4fb5563a52fb2b5ec6093> ;
    lode:inSpace <http://avaa.tdata.fi/web/smart/smear/7f885190eb43154e01c97f814b287a4b> .

<http://avaa.tdata.fi/web/smart/smear/0cf796b1a
    time:hasBeginning smear:f72d5d2e62f9747161b
    time:hasEnd smear:ffade79921356c06cbdcf1c1c

<http://avaa.tdata.fi/web/smart/smear/7f885190e
        wgs84:SpatialThing ;
    geosparql:asWKT "POINT (24.29077 61.84562)"

smear:ClassIa a smear:Classification ;
    rdfs:label "Class Ia"^^xsd:string ;
    rdfs:comment "Very clear and strong event"^^xsd:string .

smear:f72d5d2e62f9747161bb9fd127a64590 a time:Instant ;
    time:inXSDDateTime "2013-04-04T10:30:00+03:00"^^xsd:dateTime .

smear:ffade79921356c06cbdcf1c1c8fdb4dc a time:Instant ;
    time:inXSDDateTime "2013-04-04T12:00:00+03:00"^^xsd:dateTime .

<http://sws.geonames.org/656888/> a gn:Feature,
        dul:Place ;
    gn:countryCode "FI"^^xsd:string ;
    gn:locationMap <http://www.geonames.org/656888/hyytiaelae.html> ;
    gn:name "Hyytiälä"^^xsd:string ;
    wgs84:lat 6.184562e+01 ;
    wgs84:long 2.429077e+01 .
```

	beginning	end	classification	place	latitude	longitude
0	2007-05-18 12:30:00+03:00	2007-05-18 14:00:00+03:00	Class Ia	Hyytiälä	61.8456	24.2908
1	2011-04-19 09:00:00+03:00	2011-04-19 14:00:00+03:00	Class Ia	Hyytiälä	61.8456	24.2908
2	2013-04-04 10:00:00+03:00	2013-04-04 12:00:00+03:00	Class Ia	Hyytiälä	61.8456	24.2908

Observations

- Differences along dimensions of
  - Data Syntax (CSV, XML, RDF, just to name a few)
  - Exchange Protocols (HTTP but it is more complicated)
  - Programming Language (there are plenty)
- Developing libraries that cover these dimensions is expensive
- Is there an alternative?