

From Observational Data to Information IG (OD2I IG)

Markus Stocker (@envinf)

TIB Leibniz Information Centre for Science and Technology

On behalf of the OD2I Team

tinyurl.com/y9tuzvsa

Tour de Table
(time permitted)

Agenda

- Brief introduction to OD2I IG
- Update on activities since P11
- The OD2I reference conceptualization
- Conceptualizing data to information in a cloud infrastructure
- Discussion

OD2I IG

- Primary data are interpreted for their meaning in determinate contexts
 - Primary data can be observational, experimental, simulation
 - Contexts relevant to science, industry, or society generally

OD2I IG

- Primary data are interpreted for their meaning in determinate contexts
 - Primary data can be observational, experimental, simulation
 - Contexts relevant to science, industry, or society generally
- Within a context
 - Primary data are uninterpreted
 - Data interpretation results in meaningful data
 - Meaningful data is information

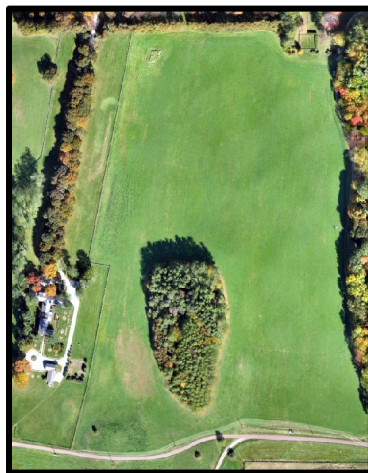
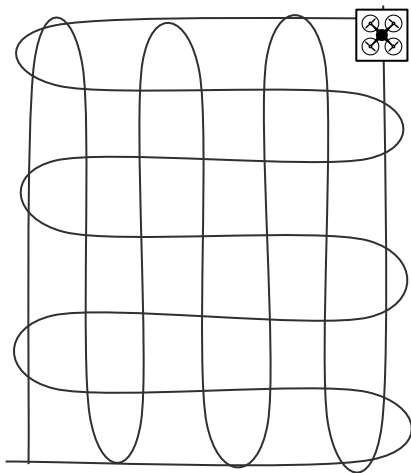
OD2I IG

- Primary data are interpreted for their meaning in determinate contexts
 - Primary data can be observational, experimental, simulation
 - Contexts relevant to science, industry, or society generally
- Within a context
 - Primary data are uninterpreted
 - Data interpretation results in meaningful data
 - Meaningful data is information
- Primary data thus evolve to become contextually meaningful information
 - Information about the natural and human worlds of interest

Examples

Scientific Unmanned Aircraft Systems

- Observational data: Multispectral Imagery
- Information: Manure Nutrient Management and Biomass Estimations
- Activity: Evaluation of agricultural soil climate change mitigation potential



Equation 7. Emission Equation for Direct N₂O Emissions from Agricultural Soils

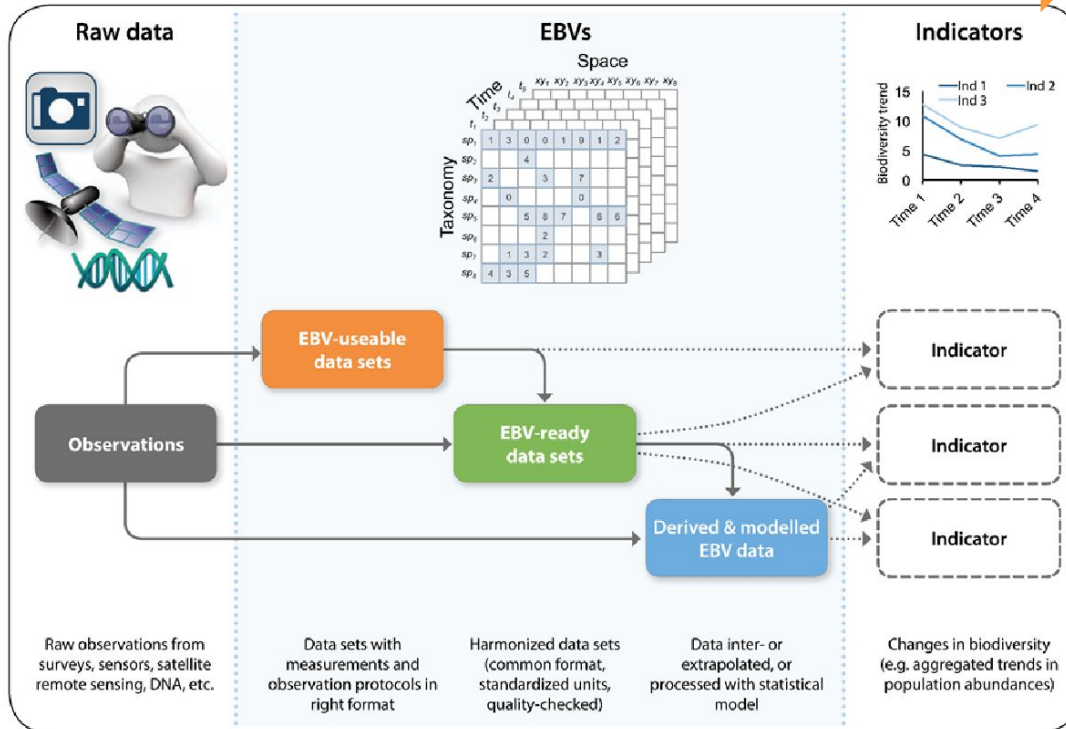
$$\text{Emissions (MMTCO}_2\text{E)} = \text{Total N} \times \text{fraction unvolatilized (0.9 synthetic or 0.8 organic)} \times 0.01 \text{ (kg N}_2\text{O-N/ kg N)} \times 44/ 28 \text{ (Ratio of N}_2\text{O to N}_2\text{O-N)} \times 298 \text{ (GWP)} \div 1,000,000,000 \text{ (kg/ MMTCO}_2\text{E)}$$



Agriculture	1.0	0.96	0.92	0.84	0.84	0.84
Enteric Fermentation	0.59	0.56	0.53	0.50	0.53	0.53
Manure Management	0.12	0.14	0.15	0.17	0.16	0.16
Agricultural Soils	0.29	0.26	0.24	0.17	0.15	0.15
TOTAL GROSS EMISSIONS	8.11	8.86	9.34	8.23	8.11	8.27
<i>Change relative to 1990</i>	-	+9%	+15%	+1.5%	0%	+2%

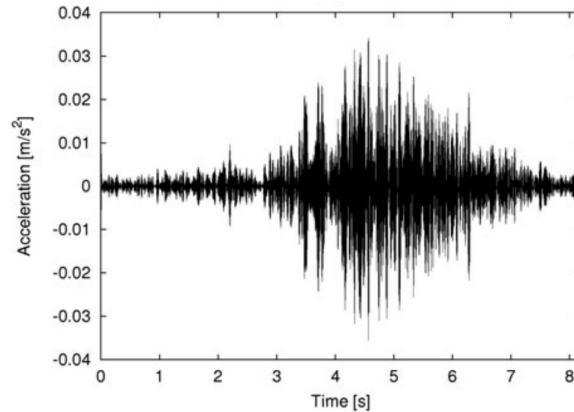
Essential Biodiversity Variables

Increasing information value



Intelligent Transportation Systems

- Observational data: Road pavement vibration
- Information: Descriptions of vehicles, their type, speed and driving direction
- Activity: Machine learning classification of vibration patterns



OD2I IG

- Advance understanding for how observational data evolve to information
- Primary focus on research data and the scientific domain
- Advance systems in their support to capture meaning
- Information rather than data, or data and their meaning
- Be a global platform for advancing this subject matter

OD2I IG

- Started at P8 in Denver with a BoF
- Endorsed IG at P11 in Berlin
- BoF meetings in between
- Collected and presented use cases
- Networking with other RDA IGs/WGs
- Initial work on a OD2I Reference Conceptualization

Since Berlin (P11)

- Regular monthly conference calls
 - One Europe-Americas friendly
 - More recently, one Europe-Australasia friendly
- Discussions and a some concrete outcomes
 - OD2I Reference Conceptualization
 - Networking with
 - Virtual Research Environments IG
 - Small Unmanned Aircraft Systems' Data IG
 - Brokering Framework WG
 - Joint sessions, e.g. with VRE IG tomorrow, 9:30 (Tsodilo B1)
 - Joint publication with some IG members



ISDE 11 11th INTERNATIONAL
SYMPOSIUM ON
DIGITAL EARTH
Florence (Italy), September 24 – 27, 2019

Digital Earth in a transformed Society



[GENERAL INFORMATION](#)



[CALL FOR SPECIAL SESSIONS](#)



[CALL FOR ABSTRACTS](#)

[COMMITTEES](#)

[PROGRAM AT A GLANCE](#)

[SPEAKERS](#)



[CALL FOR SPONSORSHIPS AND EXHIBITORS](#)

[REGISTRATION](#)

CALL FOR ABSTRACTS

[READ MORE](#)

DIGITAL EARTH IN A TRANSFORMED SOCIETY

<http://www.digitalearth2019.eu/>

Challenges

- Pathfinding
- Defining and refining the scope
- Identify priorities
- Attract members
- Obtain new use cases

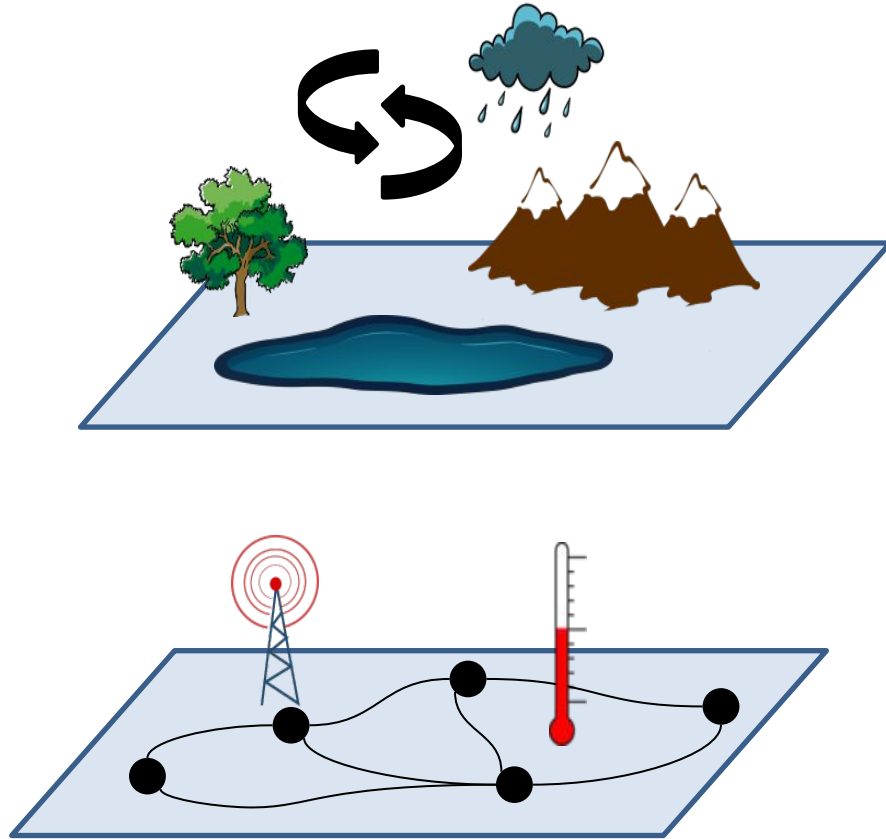
OD2I

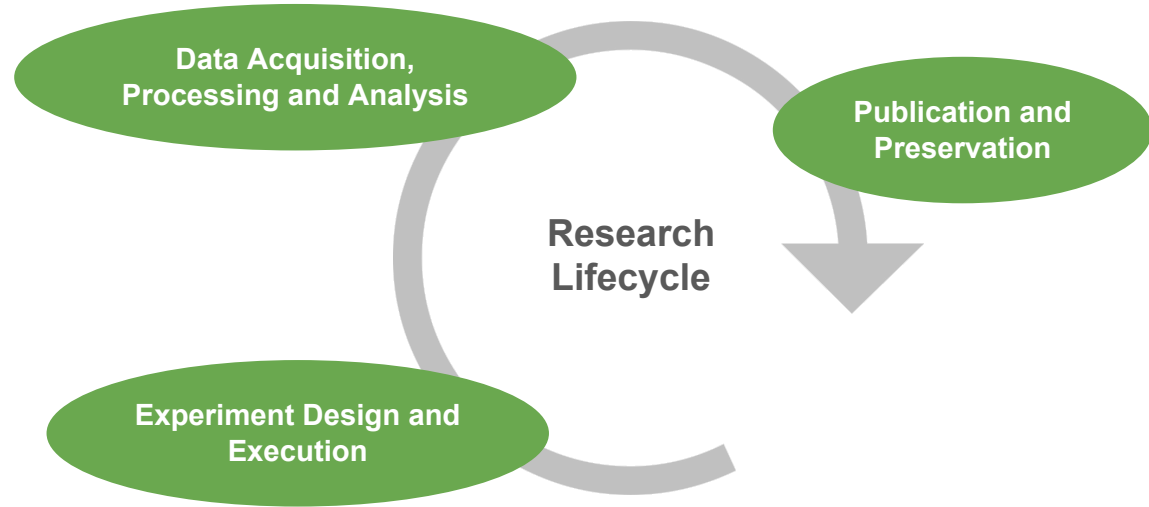
Reference Conceptualization

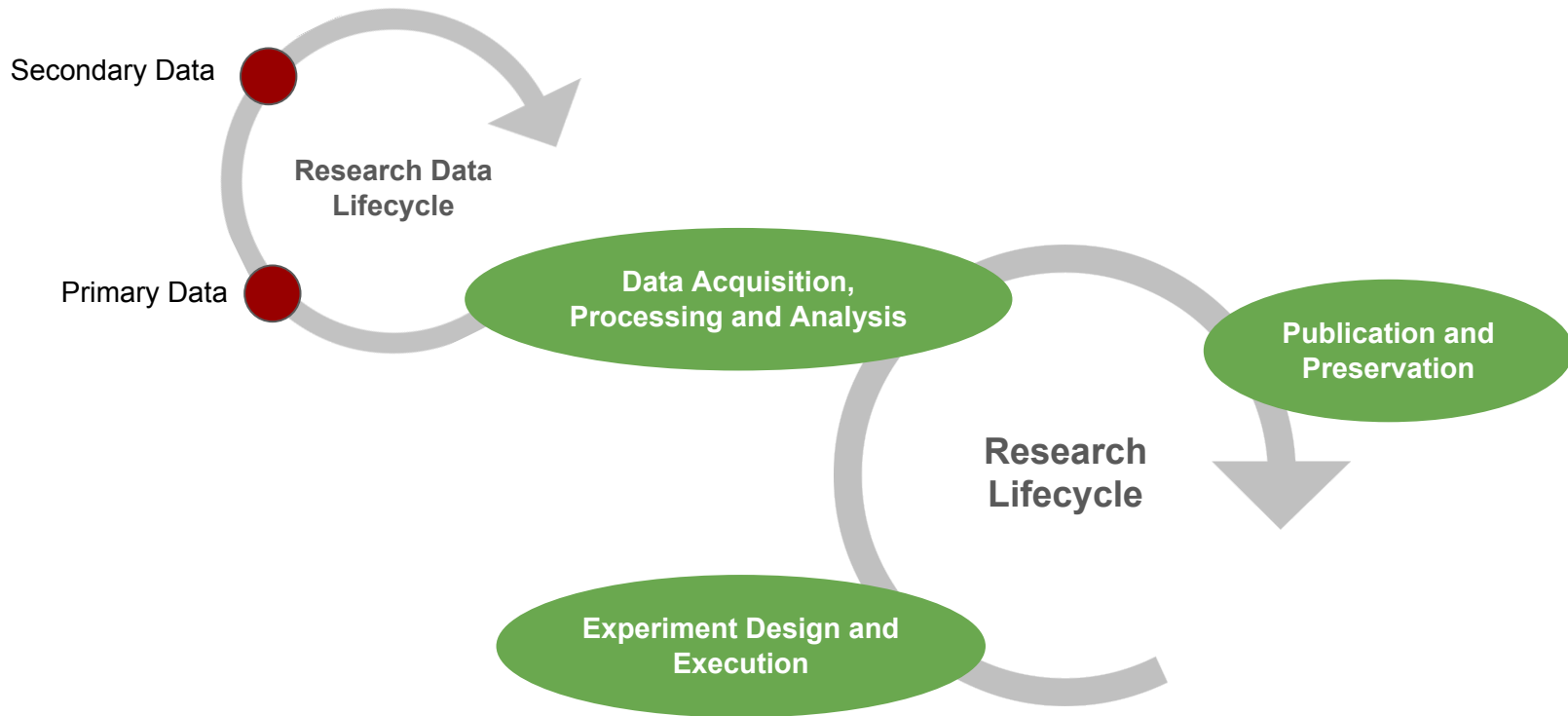
Information

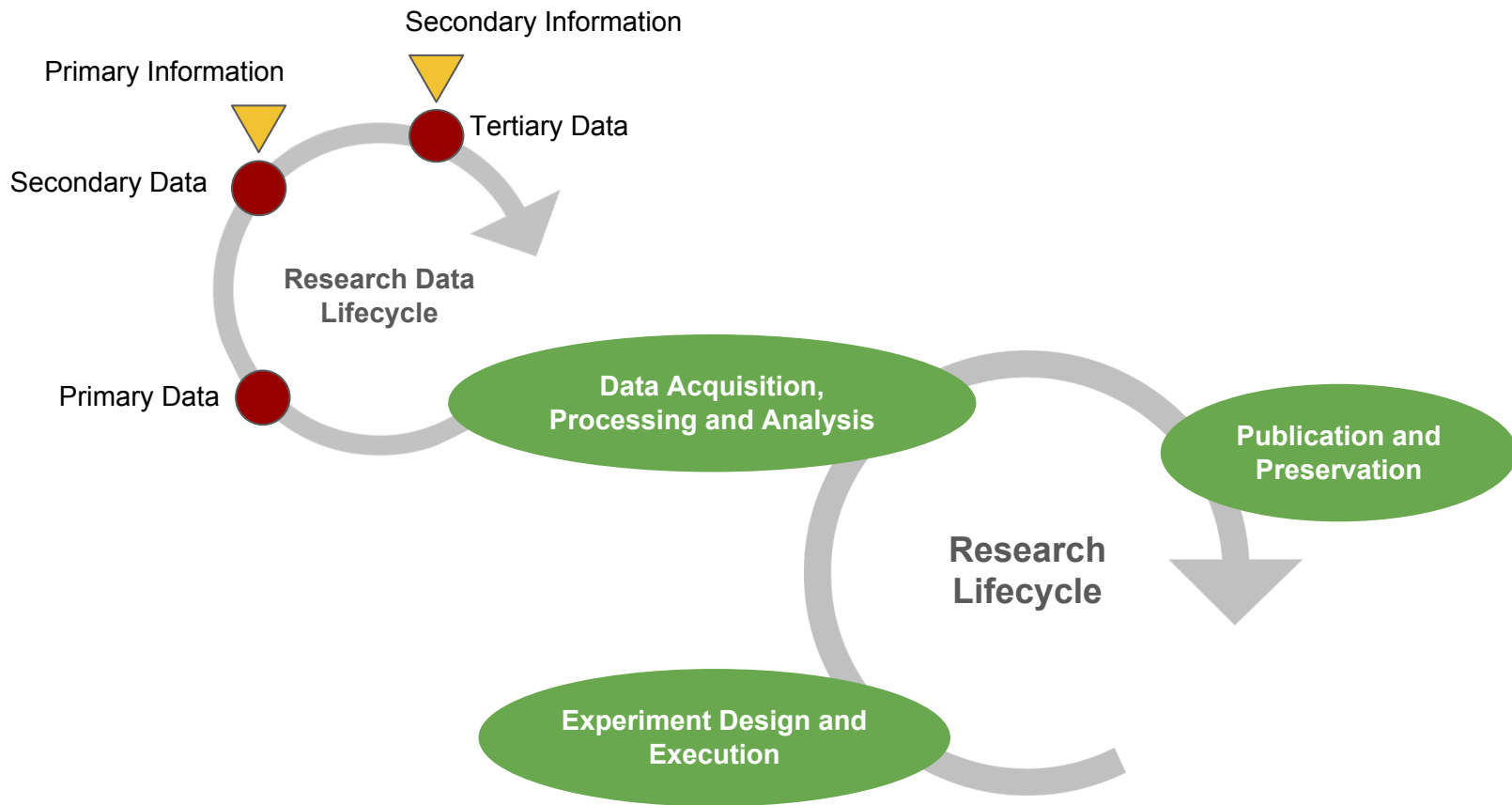


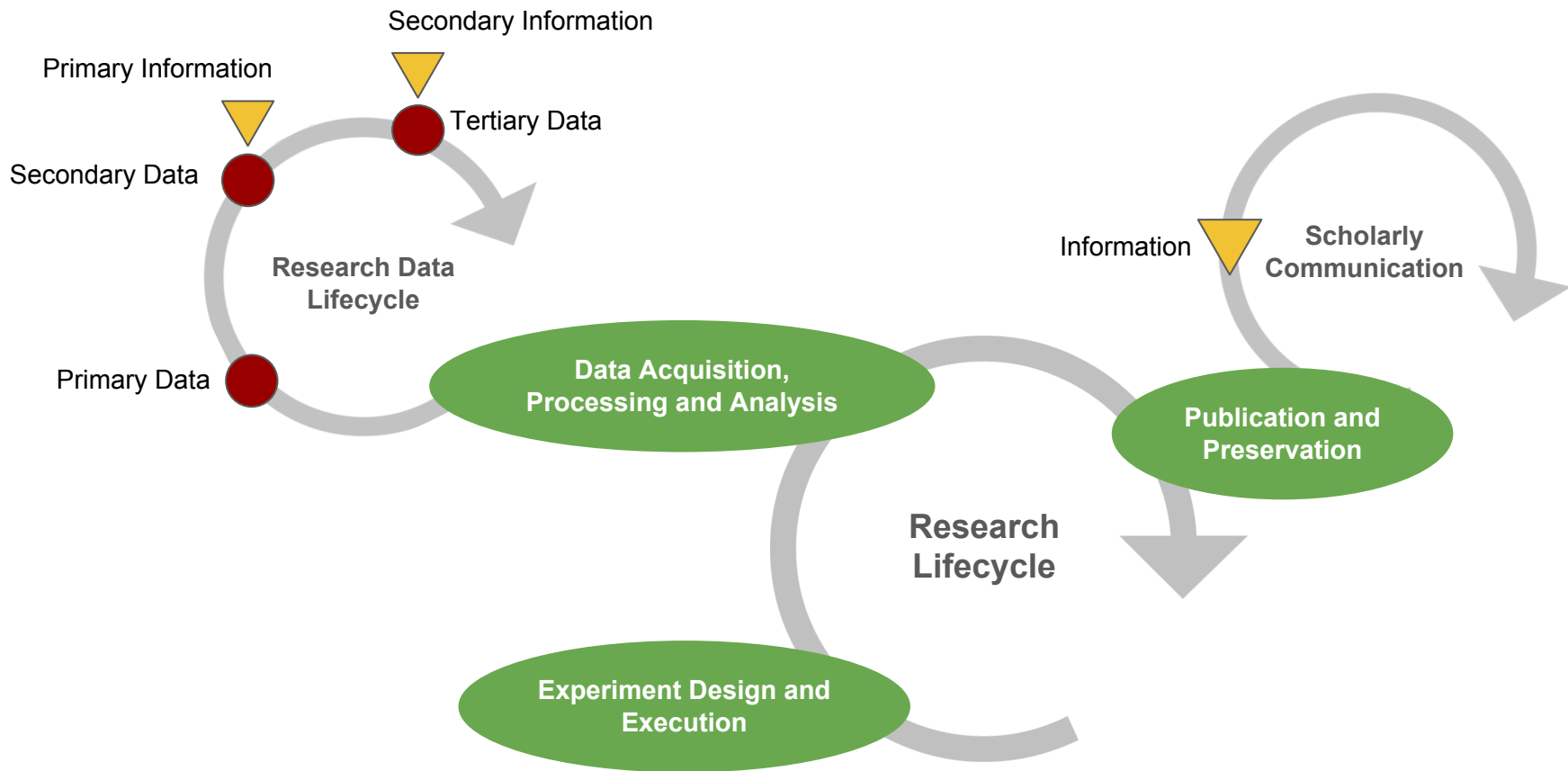
Observational Data

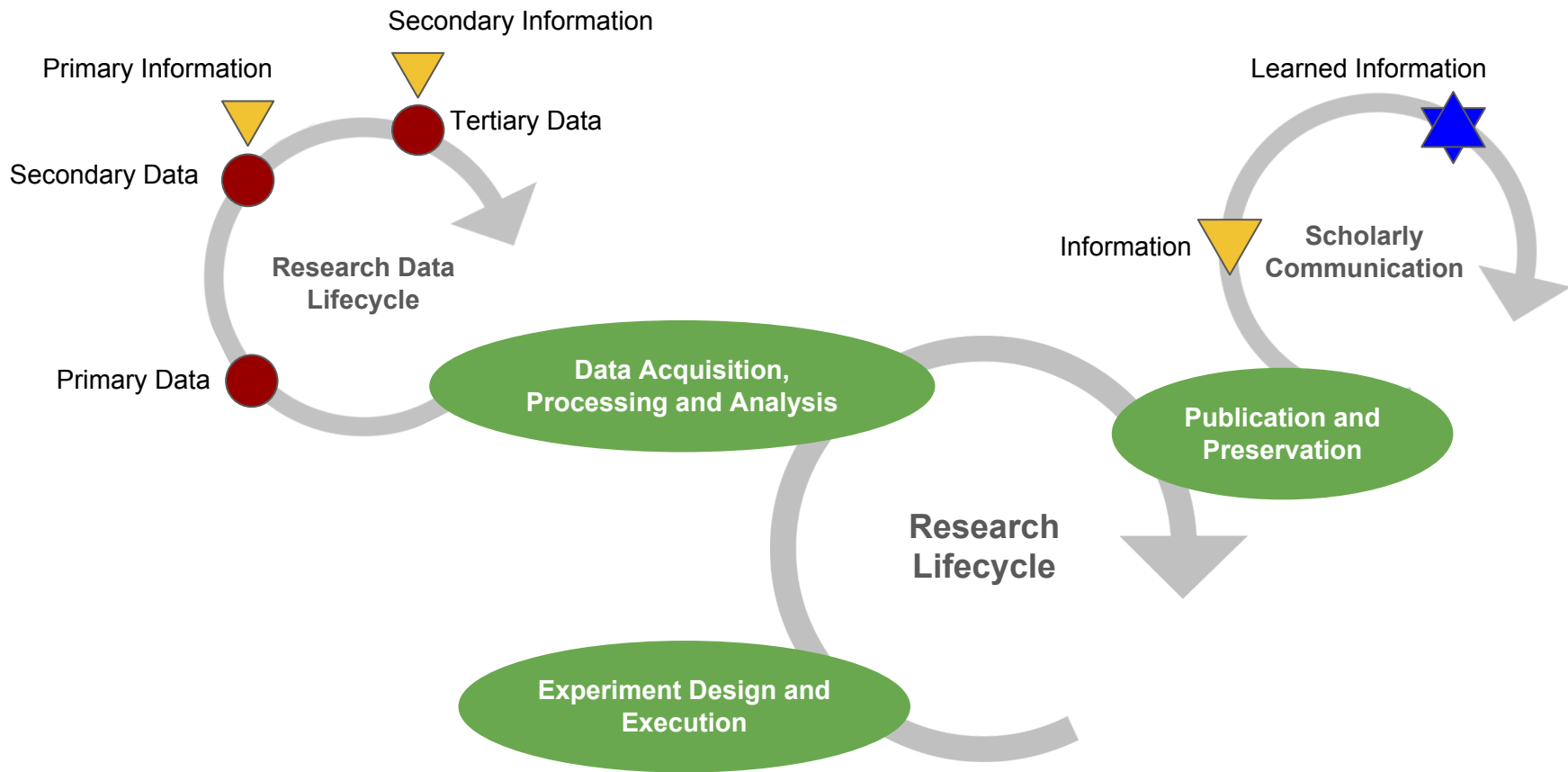








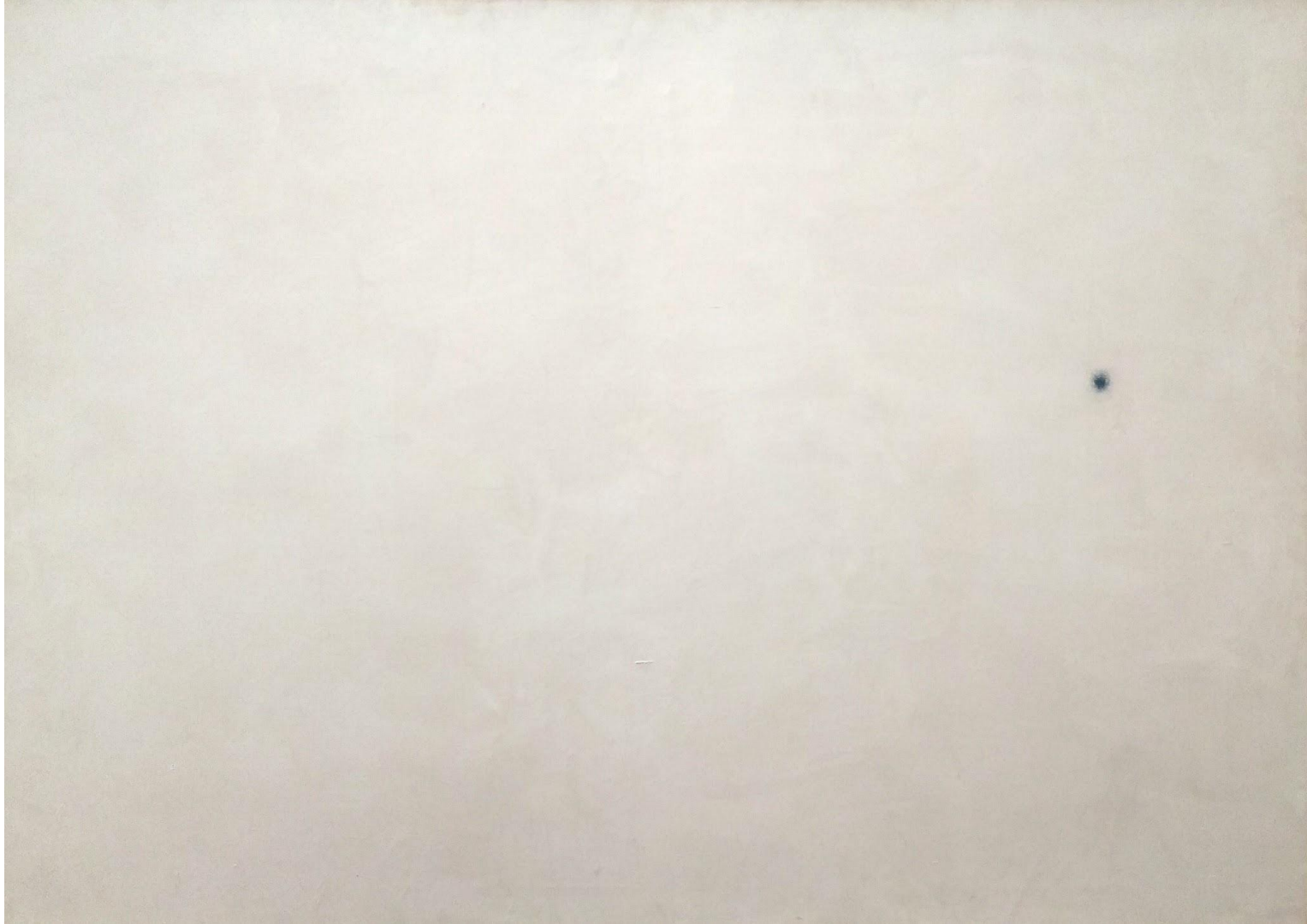




Definitions

Datum

Joan Miró
Landscape
(1968)



Datum

A datum is a putative [supposed] fact regarding some difference or lack of uniformity within some context

Floridi, L. (2011). The Philosophy of Information. Oxford University Press.

Primary and derivative data

- Primary data are the principal data stored, for example in a database
 - For instance, numerical values resulting from observation activities
 - Measurement data acquired from sensor networks
- Derivative data are data that are extracted from some (primary) data
 - Primary data used as indirect sources
 - About things other than those directly addressed by the primary data themselves

Information

- An item σ is an instance of information if
 - σ consists of n data, $n \geq 1$
 - the data are well formed
 - the well-formed data are meaningful
 - the meaningful data are truthful

Data interpretation

- Activity carried out by an interpreter through which data becomes information
- Data are uninterpreted symbols with no meaning for the system concerned
- Interpretation occurs within a real-world context and for a particular purpose
- The interpreter thus determines the contextual meaning of data

Aamodt, A and Nygård, M. 1995. Different roles and mutual dependencies of data, information, and knowledge – An AI perspective on their integration. *Data & Knowledge Engineering*, 16(3): 191–222. DOI: [https://doi.org/10.1016/0169-023X\(95\)00017-M](https://doi.org/10.1016/0169-023X(95)00017-M)

Knowledge

- Learned information
- Information incorporated in an agent's reasoning resources
- Made ready for use within decision processes
- Output of learning processes

Aamodt, A and Nygård, M. 1995. Different roles and mutual dependencies of data, information, and knowledge – An AI perspective on their integration. *Data & Knowledge Engineering*, 16(3): 191–222. DOI: [https://doi.org/10.1016/0169-023X\(95\)00017-M](https://doi.org/10.1016/0169-023X(95)00017-M)

Data to information in a cloud infrastructure


A D4Science virtual research environment demonstrator in aerosol science

D4Science.org is an organisation offering a **Data Infrastructure** service and a number of **Virtual Research Environments**

Infrastructure Capacity

D4Science is a Data Infrastructure connecting **+4000 scientists in +50 countries**, integrating **+50 heterogeneous data providers**, executing **+55,000 models & algorithms/month**; providing access to over a **billion quality records** in repositories worldwide, with **99,8% service availability**.

D4Science hosts **+100 Virtual Research Environments** to serve the biological, ecological, environmental, social mining, culture heritage, and statistical communities world-wide.

[Learn More](#) 

Infrastructure Capacity

Access to a broad range of computational, storage and international data resources.

Resource Catalogue

Data discovery, accessing, analysis, and transformation in standard format.

Exploitation Models

Three simple exploitation models to meet your needs



Experiencing

D4Science offers a number of services and virtual research environments for users willing to acquire concrete understanding of the major features and capabilities.



Empowering

D4Science is supporting the operation of a large set of diverse Initiatives, Communities of Practice, and Projects by offering VREs and Services.



Powered by gCube

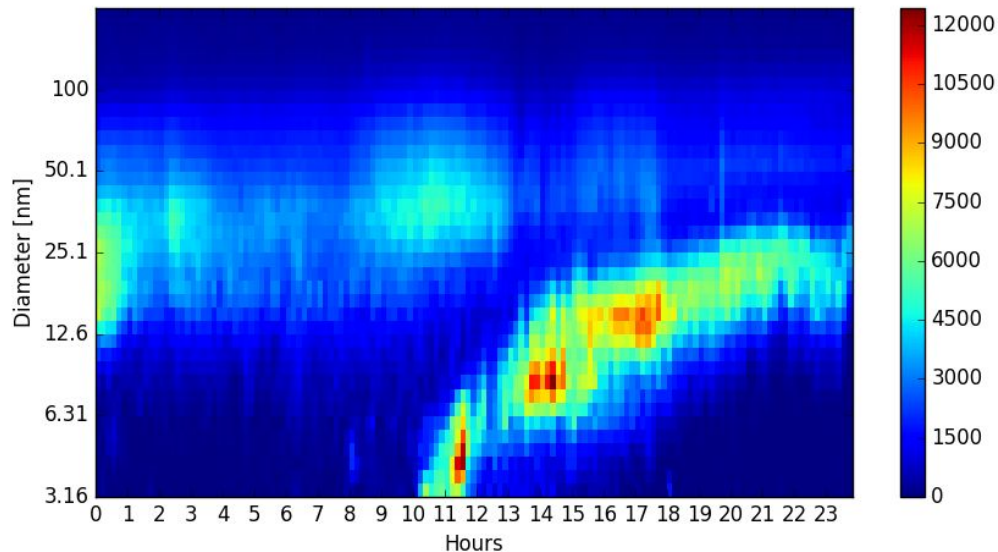
D4Science relies on the gCube software system, an open source system specifically conceived to build and operate Hybrid Data Infrastructures.

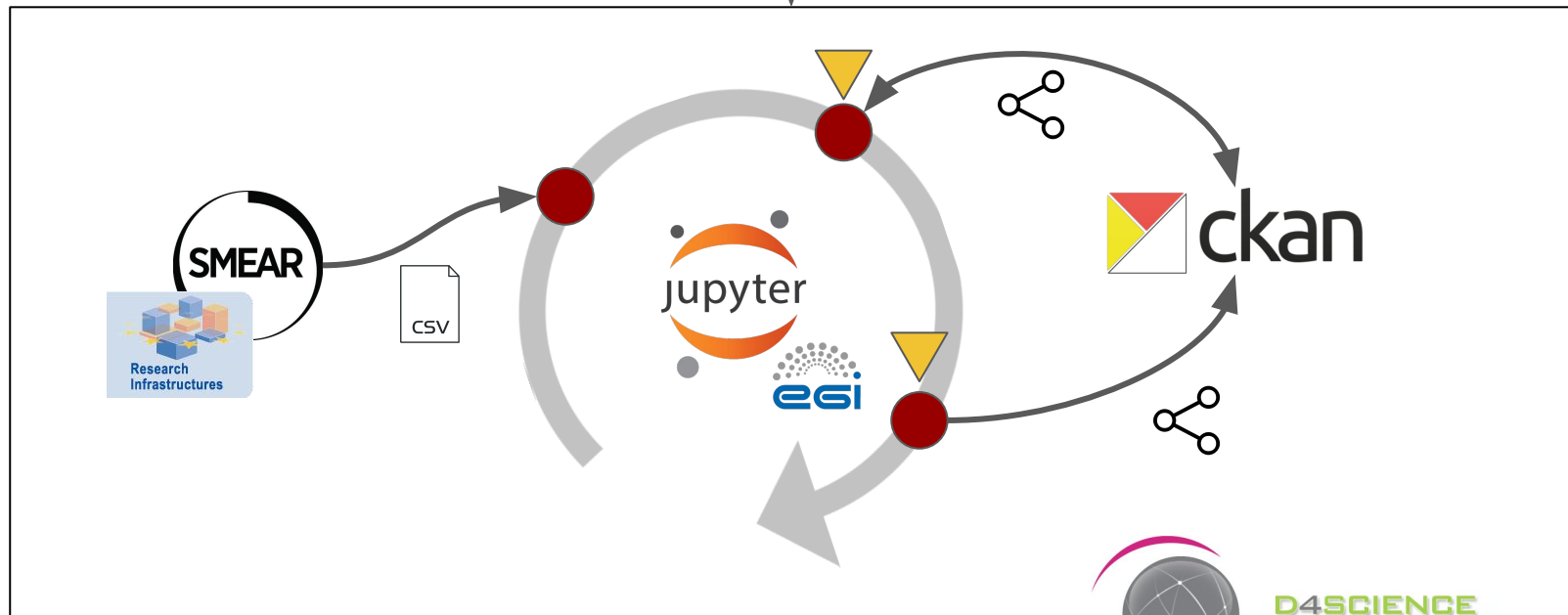
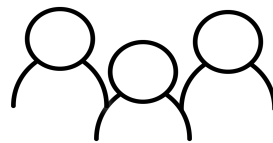


Use Case in Aerosol Science

Study of New Particle Formation Events

- Events whereby new particulate matter forms in the atmosphere
- Diameter size of particulate matter grows over time
- Aerosol scientists detect events by analysing observational data
- Events are described for their properties (e.g., duration)
- Relevant to climate change and respiratory health research





Virtual Research Environment

Jupyter @ EGI

File Edit View Run Kernel Hub Tabs Settings Help

Files

+

⚡ persistent

Name	Last Modified
classification.ipynb	9 days ago
processing.ipynb	a month ago
provenance.ipynb	a month ago

Running

Commands

Cell Tools

Tags

```

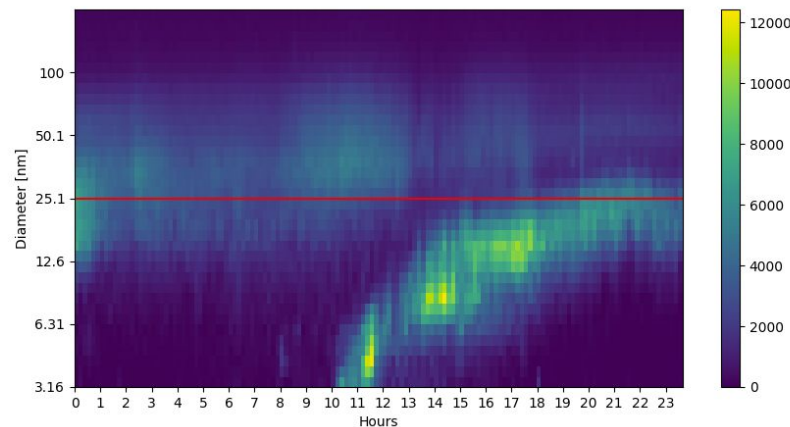
classification.ip:
+ ✂ 📄 📄 ▶ ■ ⌂ Code ▾

Python 3

In [5]: # Event days (Class Ia)
# 2007-04-15, 2007-05-05, 2007-05-18, 2007-10-19, 2008-02-19, 2009-03-19, 2009-03-22
# 2011-03-15, 2011-04-19, 2011-10-01, 2012-05-01, 2012-05-29, 2013-02-20, 2013-04-04
#
# Non Event days
# 2007-04-20, 2008-02-20, 2009-04-03, 2011-04-21, 2012-05-05, 2013-02-21

day = '2013-04-04'
place = 'Hyytiälae'

In [6]: image = plot(day, place)
visualize(image)
  
```

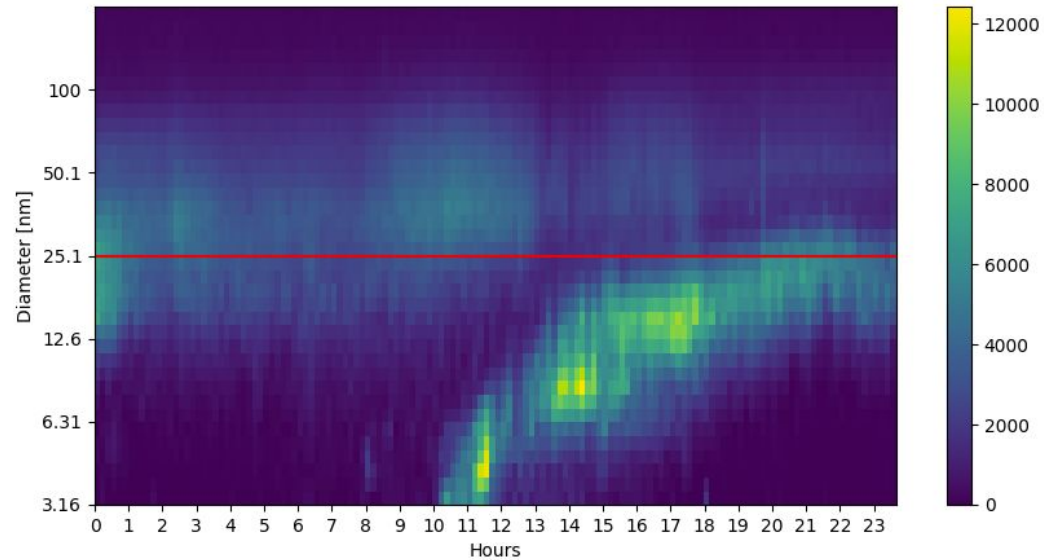


In [11]: record(day, place, '10:00', '12:00', 'Class Ia')

```
In [5]: # Event days (Class Ia)
# 2007-04-15, 2007-05-05, 2007-05-18, 2007-10-19, 2008-02-19, 2009-03-19, 2009-03-22
# 2011-03-15, 2011-04-19, 2011-10-01, 2012-05-01, 2012-05-29, 2013-02-20, 2013-04-04
#
# Non Event days
# 2007-04-20, 2008-02-20, 2009-04-03, 2011-04-21, 2012-05-05, 2013-02-21

day = '2013-04-04'
place = 'Hyytiäeläe'
```

```
In [6]: image = plot(day, place)
visualize(image)
```



```
In [11]: record(day, place, '10:00', '12:00', 'Class Ia')
```

Jupyter @ EGI



File Edit View Run Kernel Hub Tabs Settings Help

Files

persistent	
Name	Last Modified
classification.ipynb	a minute ago
processing.ipynb	a month ago
provenance.ipynb	a month ago

Commands

Cell Tools

Tabs

classification.ipynb X

processing.ipynb



```
df = pd.read_csv(StringIO(requests.get(outputUrl).content.decode('utf-8')),
                  dtype={'classification': 'str', 'place': 'str', 'latitude': 'float', 'longitude': 'float', 'uri': 'str'})
df.beginning = pd.to_datetime(df.beginning, utc=True).dt.tz_convert('Europe/Helsinki')
df.end = pd.to_datetime(df.end, utc=True).dt.tz_convert('Europe/Helsinki')
df.style.hide_columns(['uri'])
return df

def record(d):
    identifier = 'org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.PFRECORDDURATION'
    execution = wps.execute(identifier, [('value', str(d)), ('event_uris', ', '.join(df['uri'].tolist()))], output="non_deterministic_output")
```

Processing

The catalogued event descriptions can be read into a data frame, which is subsequently used to process event descriptions e.g., to compute average event durations or plot events on maps. Note that the system automatically translates the catalogued data into a data frame, which is easier to manipulate for data analysis.

In [3]: place = 'Hyttiaelae'

In [4]: # Note that this takes a couple of seconds; wait before you continue ...
df = read()

In [5]: df.style.hide_columns(['uri'])

Out[5]:

	beginning	end	classification	place	latitude	longitude
0	2005-05-12 10:00:00+03:00	2005-05-12 16:00:00+03:00	Class la	Hyttiala	61.8456	24.2908
1	2013-04-04 10:00:00+03:00	2013-04-04 12:00:00+03:00	Class la	Hyttiala	61.8456	24.2908

In [6]: # Mean event duration in hours [h]
d = (df.end - df.beginning).astype('timedelta64[h]').mean()

In [7]: d

Out[7]: 4.0

In []: record(d)

```
place = 'Hyytiaelae'
```

```
# Note that this takes a couple of seconds; wait before you continue ...  
df = read()
```

```
df.style.hide_columns(['uri'])
```

	beginning		end	classification	place	latitude	longitude
0	2005-05-12 10:00:00+03:00	2005-05-12 16:00:00+03:00		Class Ia	Hyytiälä	61.8456	24.2908
1	2013-04-04 10:00:00+03:00	2013-04-04 12:00:00+03:00		Class Ia	Hyytiälä	61.8456	24.2908

```
# Mean event duration in hours [h]  
d = (df.end - df.beginning).astype('timedelta64[h]').mean()
```

```
d
```

```
4.0
```

```
record(d)
```



ParticleFormation

This Virtual Research Environment is conceived to support the ENVR!plus Particle Formation use case. [read more](#)

Followers Items

0

3

 Follow

Organisations

ParticleFormation (3)

Types

There are no Types that match this search

Items [Activity Stream](#) [About](#)

Search items...



3 items found

Order by: Relevance ▾

NPFE plots at Hyytiälae

No Type

NPFE descriptions at Hyytiälae

No Type

NPFE mean durations

No Type

hyytiaelae-2013-04-04-plot

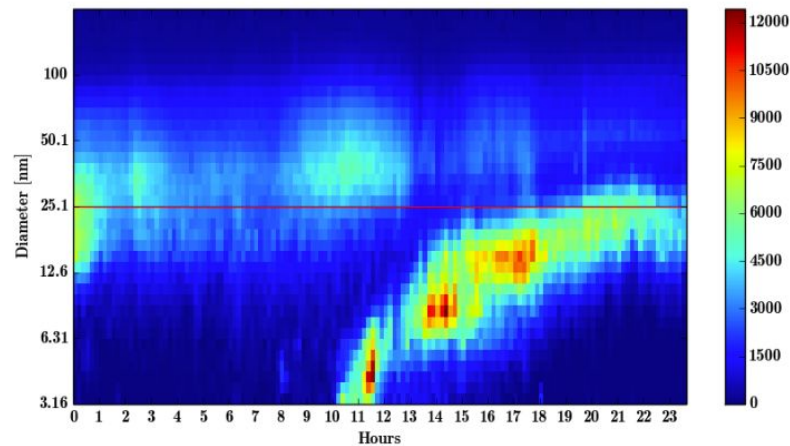
Manage

Go to resource

URL: <https://data.d4science.org/MzhkMUdQZmRrSkxOc2kzWHA0amdINIZTbW5yRWdGdUZhbWJQNStiS0N6Yz0>

Image

Embed



All Resources

hyytiaelae-2013-04-04- ...

hyytiaelae-2005-05-12- ...

Additional Information

Field	Value
Last updated	September 28, 2018
Created	September 28, 2018
Format	PNG
License	Academic Free License 3.0

hyytielae-2013-04-04-description

URL: <https://data.d4science.org/K0JMcUorTjJib1Bka0hVdDI4SmU5N21wQmpubHBJOXdhbWJQNStiIS0N6Yz0>

Text

Embed

```
@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 prov: <http://www.w3.org/ns/prov#> .
@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/smea/> .
@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/smea/2c3514176ca677a99292cbb4b6a3ae> a lode:Event,
    prov:Entity ;
    smear:hasClassification smear:ClassIa ;
    lode:atPlace <http://svs.geonames.org/656888/> ;
    lode:atTime <http://avaa.tdata.fi/web/smart/smea/316de4bbbe748e657b1ffeadb57f78a> ;
    lode:inSpace <http://avaa.tdata.fi/web/smart/smea/7f885190eb43154e01c97f814b287a4b> ;
    prov:wasDerivedFrom <http://data.d4science.org/bmszTnIrTkFVHzZHqmpGempQdXFVUTF0dkZlaHp6ZUJHbWJQNStiIS0N6Yz0-VLT> ;
    prov:wasGeneratedBy <http://avaa.tdata.fi/web/smart/smea/97b3650aa60cf864ea88ea44ec39d510> .

<http://avaa.tdata.fi/web/smart/smea/316de4bbbe748e657b1ffeadb57f78a> a time:Interval ;
    time:hasBeginning <http://avaa.tdata.fi/web/smart/smea/993547816203c9d7a724283dda2ccae7> ;
    time:hasEnd smear:ffade79921356c06cbdcf1c1c8fdb4dc .

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

<http://avaa.tdata.fi/web/smart/smea/97b3650aa60cf864ea88ea44ec39d510> a <http://purl.obolibrary.org/obo/OBI_0200111>,
    prov:Activity ;
    prov:endedAtTime "2018-09-28T14:59:27.177710+02:00"^^xsd:dateTime ;
    prov:startedAtTime "2018-09-28T14:59:27.177710+02:00"^^xsd:dateTime ;
    prov:used <http://data.d4science.org/bmszTnIrTkFVHzZHqmpGempQdXFVUTF0dkZlaHp6ZUJHbWJQNStiIS0N6Yz0-VLT> .

<http://avaa.tdata.fi/web/smart/smea/993547816203c9d7a724283dda2ccae7> a time:Instant ;
    time:inXSDDateTime "2013-04-04T10:00:00+03:00"^^xsd:dateTime .

smear:ClassIa a smear:Classification ;
    rdfs:label "Class Ia"^^xsd:string ;
    rdfs:comment "Class Ia is a classification of events that are generated by the system."^^xsd:string .
```

All Resources

hyytielae-2013-04-04-...

hyytielae-2005-05-12-...

Additional Information

Field	Value
Last updated	September 28, 2018

2018-11-01T181732-npfe-mean-duration

[Manage](#)
[Go to resource](#)

URL: <https://data.d4science.org/bUo0VjBHR0FnSGhWQ2FXeWFKblgwMVd4VGdxN0tLWEhHbWJQNSItS0N6Yz0>

[Text](#)
[Embed](#)

```
@prefix obo: <http://purl.obolibrary.org/obo/> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

obo:BFO_0000051 rdfs:label "has part" .

obo:IAO_0000004 rdfs:label "has measurement value" .

obo:IAO_0000039 rdfs:label "has measurement unit label" .

obo:OBI_0000293 rdfs:label "has_specified_input" .

obo:OBI_0000299 rdfs:label "has_specified_output" .

obo:OBI_0000312 rdfs:label "is_specified_output_of" .

<http://avaa.tdata.fi/web/smart/smeas/2c3514176ca67a77a99292cbb4b6a3ae> a obo:IAO_0000027 .

<http://avaa.tdata.fi/web/smart/smeas/b176682782ba99884a99259eb808e111> a obo:IAO_0000032,
    obo:OBI_0000679,
    prov:Entity ;
    obo:IAO_0000004 4.0 ;
    obo:IAO_0000039 obo:UO_0000032 ;
    obo:OBI_0000312 <http://avaa.tdata.fi/web/smart/smeas/dfdc52d59bfecb0b31075ebd29a9192e> ;
    prov:wasDerivedFrom <http://avaa.tdata.fi/web/smart/smeas/78549852a462fcdab086c5e6436700f> ;
    prov:wasGeneratedBy <http://avaa.tdata.fi/web/smart/smeas/dfdc52d59bfecb0b31075ebd29a9192e> .

<http://avaa.tdata.fi/web/smart/smeas/d9f51744177086cd05a822898a3ba0f8> a obo:IAO_0000027 .

obo:IAO_0000032 rdfs:label "scalar measurement datum" .

obo:IAO_0000100 rdfs:label "data set" .

obo:OBI_0000679 rdfs:label "average value" .

obo:OBI_0200079 rdfs:label "arithmetic mean calculation" .

obo:UO_0000003 rdfs:label "time unit" .

- - - - -
- - - - -
```

All Resources

2018-09-28T131252-npfe ...

2018-10-23T095719-npfe ...

2018-10-23T102059-npfe ...

Additional Information

Field	Value
Last updated	November 1, 2018

Advantages

- Syntactic and semantic homogeneity of derivative data across researchers
- Systematic acquisition of derivative data in infrastructure
- Semantics of derivative data are explicit (and machine readable)

Discussion

- General comments
- Reference conceptualization
- D4Science implementation of the aerosol use case
- New use cases
- Work plan until P13
- Work plan beyond P13