A Missing Link from Data to Knowledge: Infrastructure that Curate the Meaning of Data

Markus Stocker(1), Markus Fiebig(2), Alex Hardisty(3)

(1) German National Library of Science and Technology (TIB)
(2) NILU - Norsk Institutt for Luftforskning
(3) Cardiff University

markus.stocker@tib.eu | @envinf

Introduction

- Researchers are essential on the “road” from data to knowledge
- By interpreting data, they determine their contextual meaning [1]
- Thereby generating information - *meaningful* data

[1] Aamodt, A., Nygård, M. (1995). Different roles and mutual dependencies of data, information, and knowledge – An AI perspective on their integration, Data & Knowledge Engineering, vol. 16, no. 3, pp. 191-222. https://doi.org/10.1016/0169-023X(95)00017-M

Data, Information, Knowledge

- Logical progression has been described as “fairytale” [1]
- Indeed, information is represented as data in systems

[1] Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge Journal of the American Society for Information Science and Technology, Wiley Subscription Services, Inc., A Wiley Company, 58, 479-493. https://doi.org/10.1002/asi.20508

Data, Information, Knowledge

- Logical progression has been described as “fairytale” [1]
- Indeed, information is represented as data in systems
- In what sense, then, do we progress from data to information, knowledge?

[1] Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge Journal of the American Society for Information Science and Technology, Wiley Subscription Services, Inc., A Wiley Company, 58, 479-493. https://doi.org/10.1002/asi.20508

Data, Information, Knowledge

- Logical progression has been described as “fairytale” [1]
- Indeed, information is represented as data in systems
- In what sense, then, do we progress from data to information, knowledge?
- Perhaps from *primary* data (observational, experimental, computational [2])
- That are uninterpreted in determinate context
- Through data interpretation activity, resulting in information
- Contextually meaningful (well-formed truthful [3]) data

[1] Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge Journal of the American Society for Information Science and Technology, Wiley Subscription Services, Inc., A Wiley Company, 58, 479-493. https://doi.org/10.1002/asi.20508
[2] Borgman, C. L. (2007). Scholarship in the Digital Age: Information, Infrastructure, and the Internet. MIT University Press.
[3] Floridi, L. (2011). The Philosophy of Information. Oxford University Press.

Consider

- Advanced research infrastructures that curate primary data do exist
  - Examples in most scientific domains
  - ICOS, ACTRIS, NEON to name a few in earth and environmental science
  - CERN, ELIXIR, CESSDA, CLARIN to name a few in other domains

Consider

- Advanced research infrastructures that curate primary data do exist
  - Examples in most scientific domains
  - ICOS, ACTRIS, NEON to name a few in earth and environmental science
  - CERN, ELIXIR, CESSDA, CLARIN to name a few in other domains
- Weak integration of researcher data interpretation with infrastructures
  - Download of published data is surely the predominant paradigm
  - Even though download is *considered harmful* [1]
  - *Infrastructural disconnect*: Data use does not occur *on* research infrastructures

[1] Atkinson, M.P. et al. (2018). Download considered harmful – provokes – a flexible federation framework. (In Preparation)

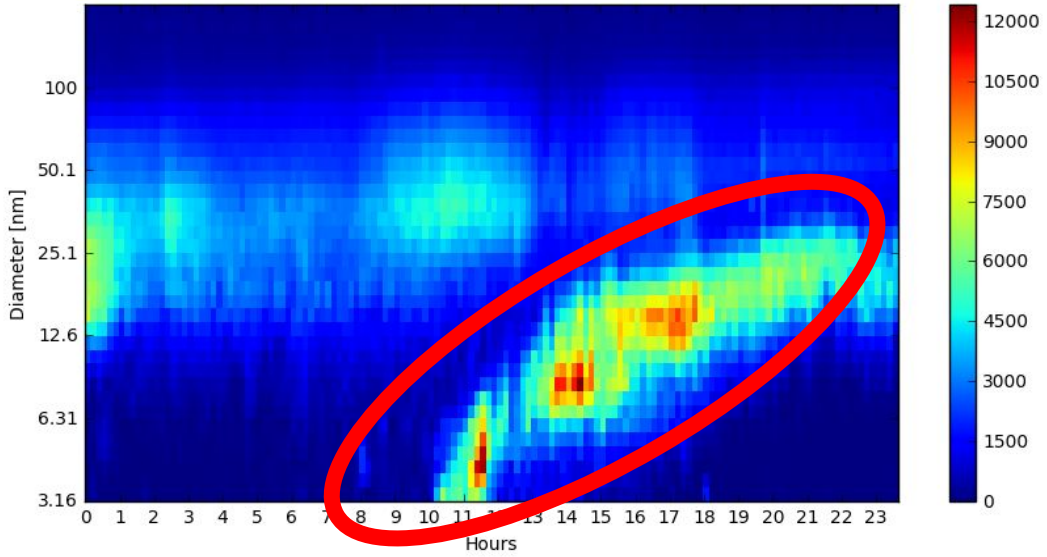
Consider

- Advanced research infrastructures that curate primary data do exist
  - Examples in most scientific domains
  - ICOS, ACTRIS, NEON to name a few in earth and environmental science
  - CERN, ELIXIR, CESSDA, CLARIN to name a few in other domains
- Weak integration of researcher data interpretation with infrastructures
  - Download of published data is surely the predominant paradigm
  - Even though download is *considered harmful* [1]
  - *Infrastructural disconnect*: Data use does not occur *on* research infrastructures
- Too often, meaning is *lost in translation* when information is represented
  - For instance, raster image is two-dimensional array of integers in systems
  - The meaning of those integers is implicit, requires re-interpretation
  - Information not represented using a language for knowledge representation

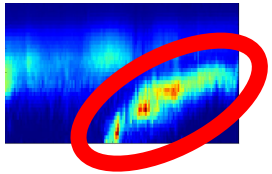
[1] Atkinson, M.P. et al. (2018). Download considered harmful – provokes – a flexible federation framework. (In Preparation)

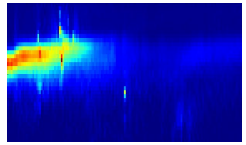
Example

Research Infrastructures



Particle size distribution
primary (observational) data





//

?

Information about new
particle formation events

Information about NPFE: Data, actually!

734546 0 1 0 0
734547 1 0 0 0
734550 0 0 0 1
734551 0 0 1 0

MATLAB datenum
Class Ia
Class Ib
Class II
Non Event

2011-07-04,NE
2011-07-04,1
2011-07-05,3
2011-07-06,BD

Date
Class 0-4 | Label NE,BD

NE = Non Event
BD = Bad Data

04/07/2009,733958,2
05/07/2009,733959,0
06/07/2009,733960,1
08/07/2009,733962,3

Date
MATLAB datenum
Class 1,2 | Label 0,3

Class 1 = Class Ia, Ib
Class 2 = Class II
0 = Non Event
3 = Undefined

MATLAB datenum
Class Ia
Class Ib
Class II
Non Event

734546 0 1 0 0

734547 1 0 0 0

734550 0 0 0 1

734551 0 0 1 0

Not FAIR

(meta)data

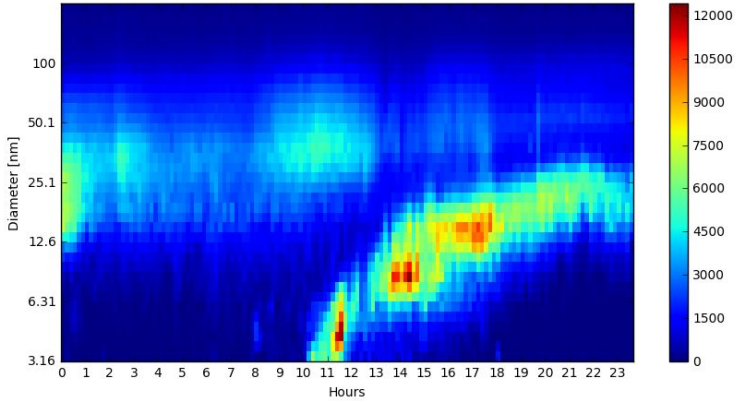
Proposal

Data Use *as a Service*

- Deep integration of researcher community data use with infrastructure
- Reuse software implementation for data interpretation (Jupyter)
- Semantic technologies to represent information, data and meaning
- Ensure information is acquired by information infrastructure

```
from smear.datafetcher import fetchdata
from smear.dataplotter import plotdata

# Fetch and plot concentration data for the given time and location
# from SmartSMEAR, https://avaa.tdata.fi/web/smart
plotdata(fetchdata('2013-04-04', 'Hyytiälä'))
```



Optional

```
from smear.datafetcher import fetchdata
from factory import assess

# Automated assessment for whether or not an event occurred
assess(fetchdata('2013-04-04', 'Hyytiälä'))
```

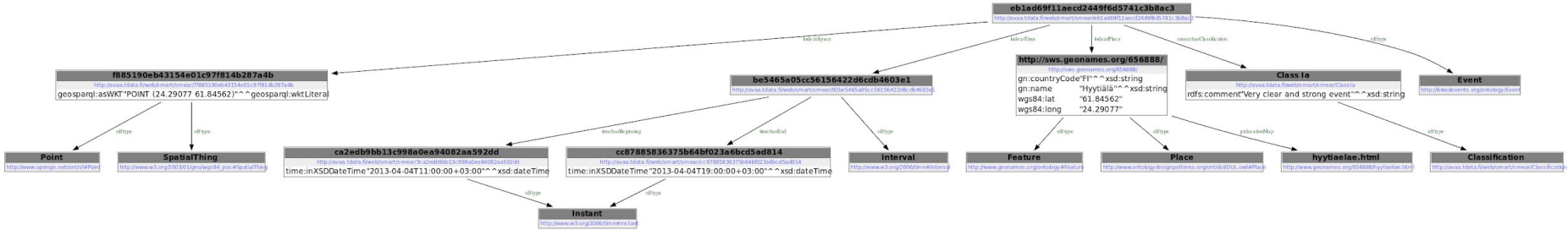
```
['Event']
```

```
from factory import record, event

# Record information about the new particle formation event
record(event('2013-04-04', 'Hyytiälä', '11:00', '19:00', 'Class Ia'))
```

```
[] a lode:Event ;
  smear:hasClassification smear:ClassIa ;
  lode:atPlace [
    a gn:Feature, DUL:Place ;
    gn:countryCode "FI"^^xsd:string ;
    gn:name "Hyytiälä"^^xsd:string
  ] ;
  lode:atTime [
    a time:Interval ;
    time:hasBeginning [ time:inXSDDateTime "2013-04-04T11" ] ;
    time:hasEnd [ time:inXSDDateTime "2013-04-04T19" ]
  ] ;
  lode:inSpace [
    a sf:Point, wgs84:SpatialThing ;
    geosparql:asWKT "POINT (24.29077 61.845629)"
  ] .
```

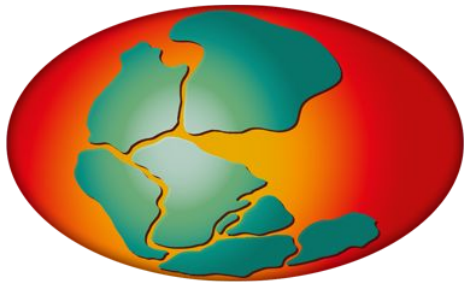
Information object
acquired by information infrastructure
representing data and meaning explicitly



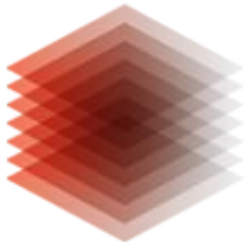
Discussion

Who operates the information infrastructure?

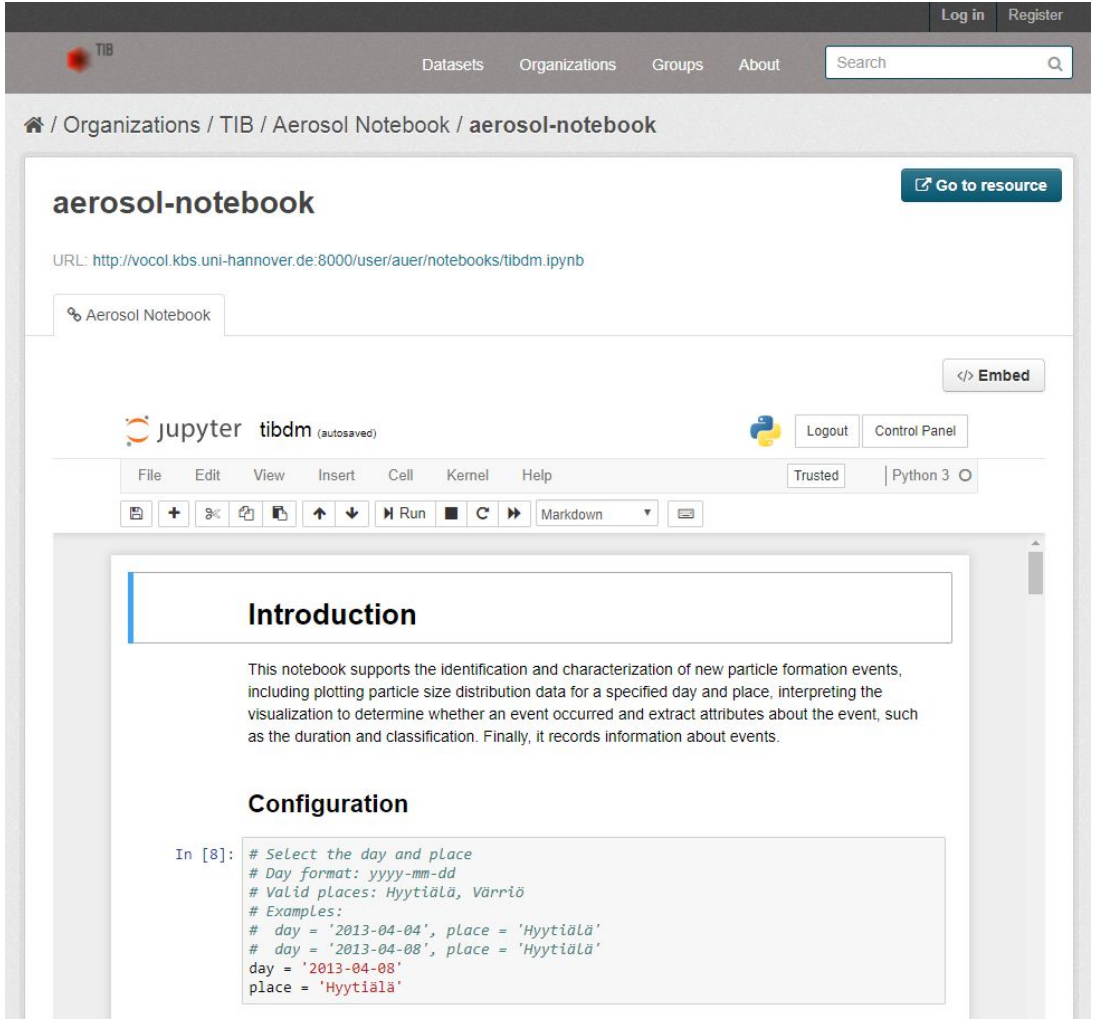

ACTRiS



PANGAEA.



TIB
LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK


Log in
Register
TIB
Datasets
Organizations
Groups
About
Search
/ Organizations / TIB / Aerosol Notebook / aerosol-notebook
aerosol-notebook
Go to resource
URL: http://vocol.kbs.uni-hannover.de:8000/user/auer/notebooks/tibdm.ipynb
Aerosol Notebook
Embed
jupyter tibdm (autosaved)
Logout
Control Panel
File Edit View Insert Cell Kernel Help
Trusted
Python 3
Run
Markdown
Introduction
This notebook supports the identification and characterization of new particle formation events, including plotting particle size distribution data for a specified day and place, interpreting the visualization to determine whether an event occurred and extract attributes about the event, such as the duration and classification. Finally, it records information about events.
Configuration
In [8]:
# Select the day and place
# Day format: yyyy-mm-dd
# Valid places: Hyytiälä, Värriö
# Examples:
# day = '2013-04-04', place = 'Hyytiälä'
# day = '2013-04-08', place = 'Hyytiälä'
day = '2013-04-08'
place = 'Hyytiälä'

RDA Interest Group

- From Observational Data to Information (OD2I IG)
- rd-alliance.org/groups/observational-data-information
- Recently endorsed
- Kick-off meeting at P11, Breakout 3

Takeaways

- From data to knowledge researchers are essential
- They determine the contextual meaning of data
- Deeper integration of research communities and infrastructure
- These elements of knowledge infrastructures
- Networks that generate and maintain knowledge