Semantic Data Science

Markus Stocker
TIB Leibniz Information Centre for Science and Technology

@envinf
https://orcid.org/0000-0001-5492-3212

Objective

Make the case that **scientists** and research infrastructures **are elements of scientific knowledge infrastructures** and that there is an urgent need in such infrastructures to **ensure that the meaning of data is made explicit**.

Schedule

Time	Content
08:00 - 09:00	Part I: The conceptual and technology framework
09:00 - 09:30	Part II: A knowledge infrastructure in atmospheric physics as a case in point
09:30 - 10:15	Part III: Hands-on assignment
10:15 - 10:30	Part IV: Discussion

Part I

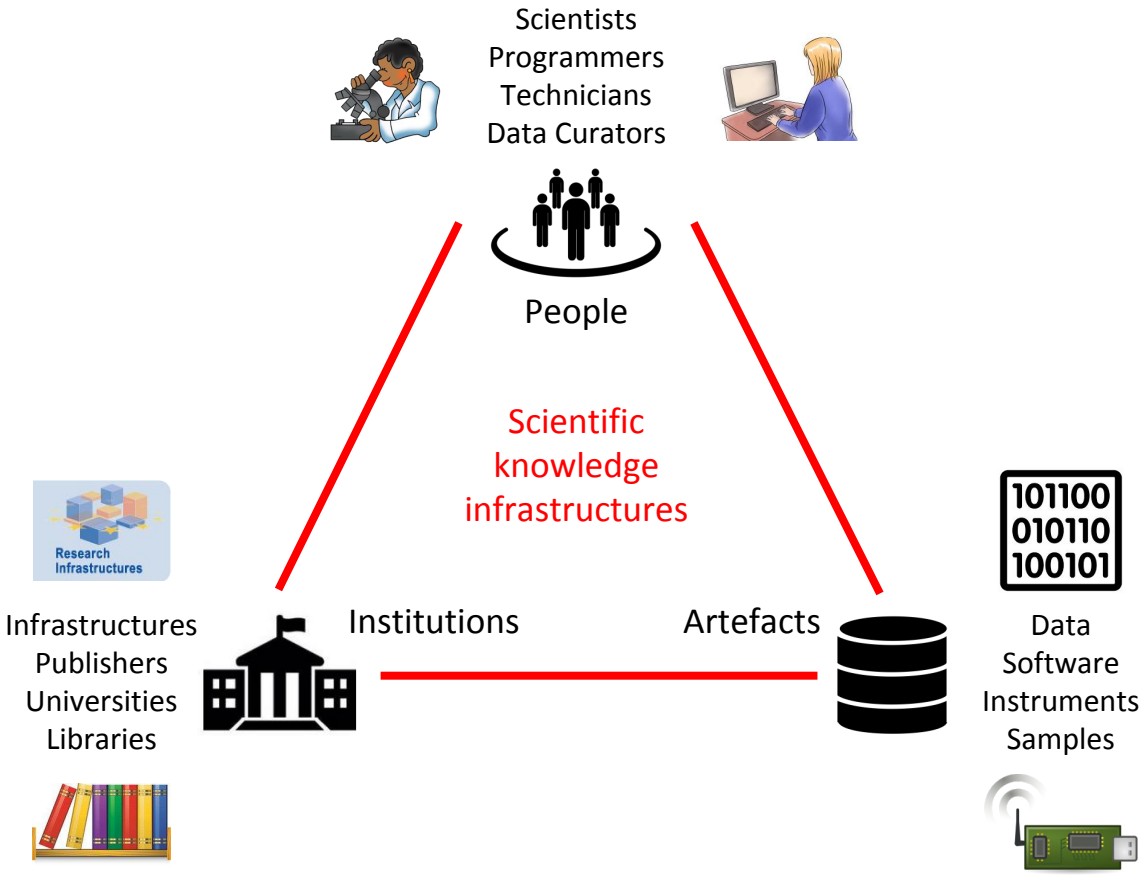
The conceptual and technology framework

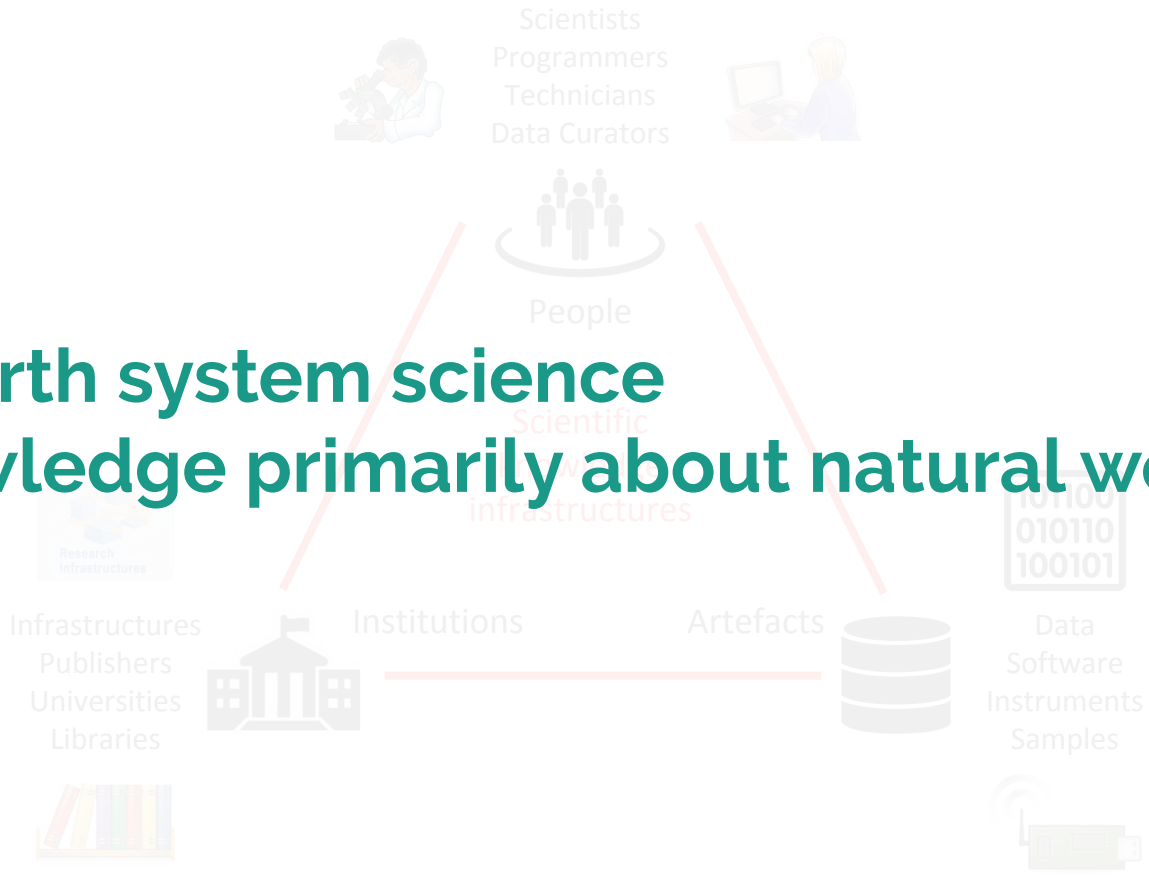
The conceptual framework

Knowledge infrastructures

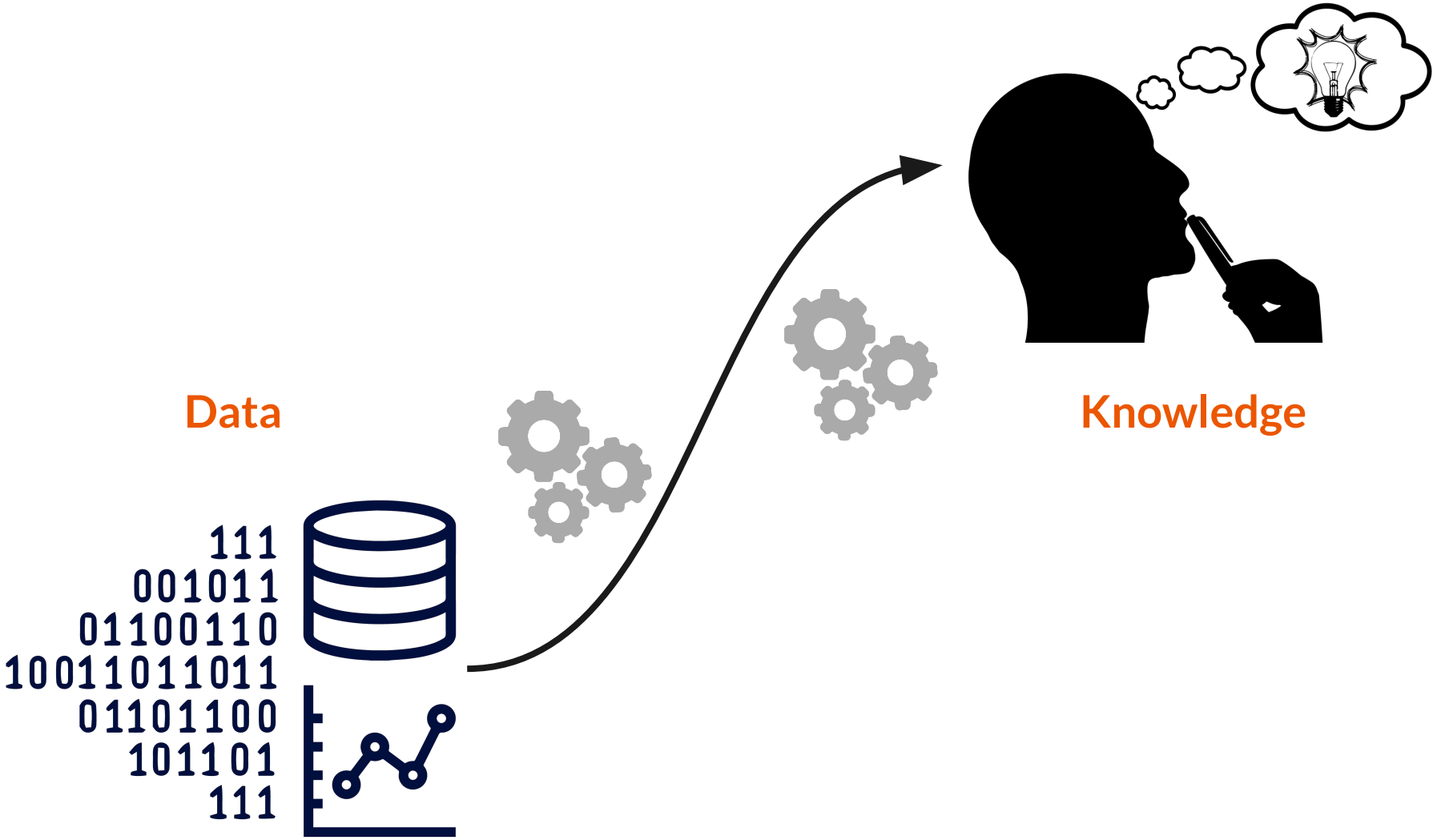
Robust networks of people, artifacts, and institutions that generate, share, and maintain specific knowledge about the human and natural worlds.

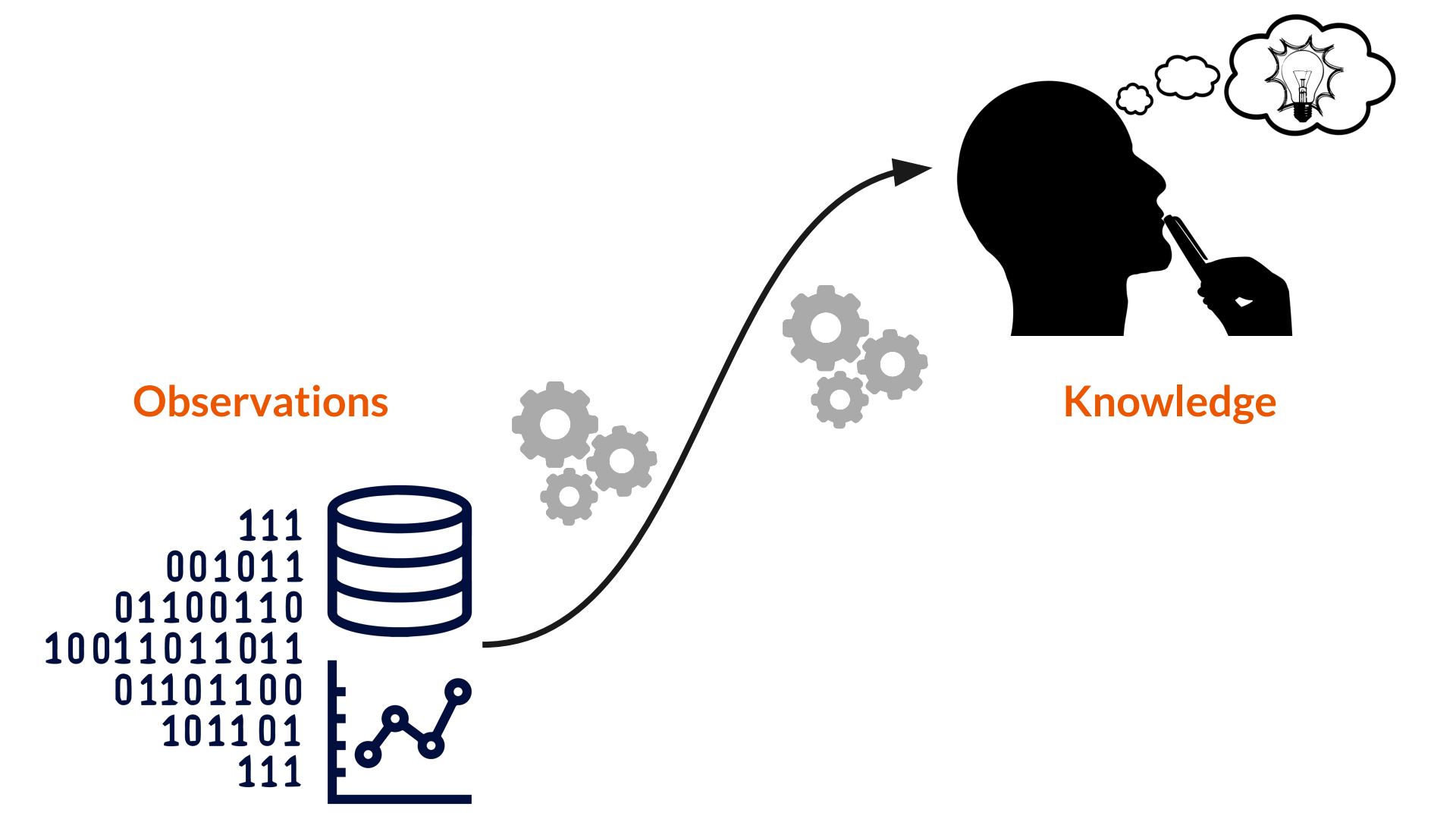
-- Paul Edwards (2010). A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming, p. 17

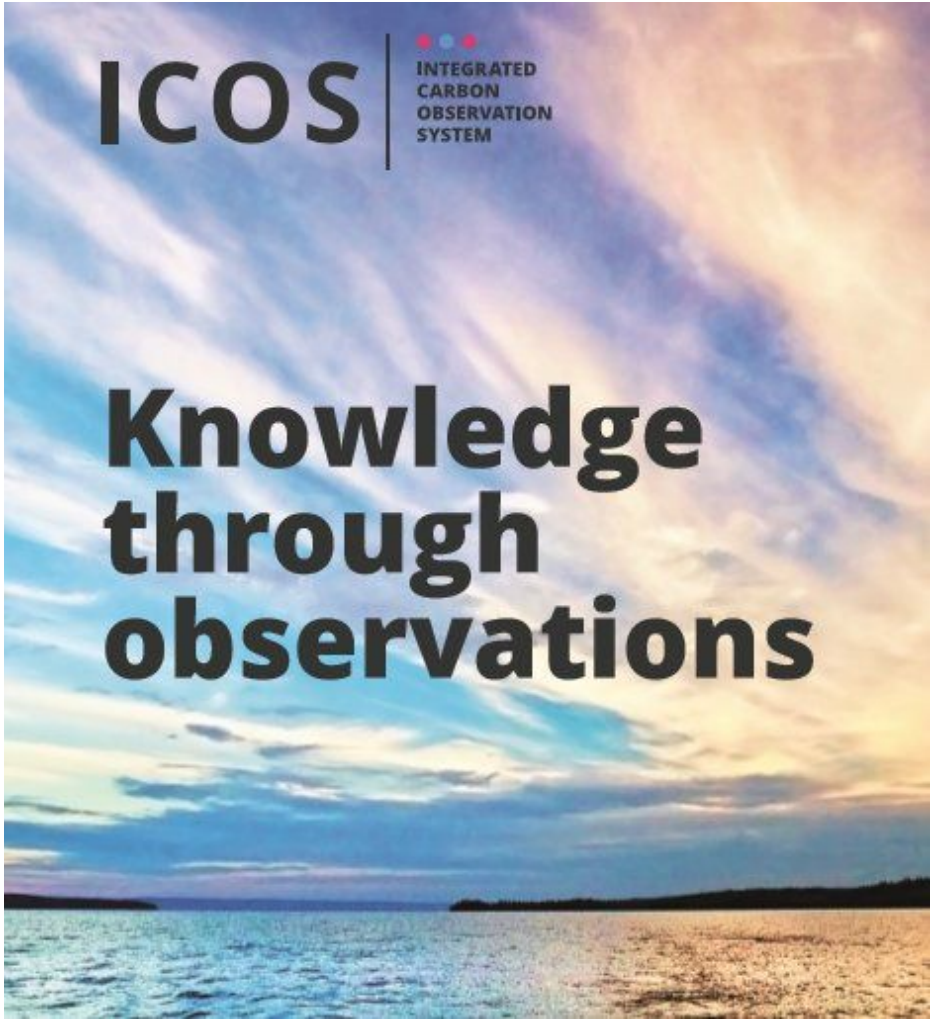

Scientists
Programmers
Technicians
Data Curators
People
Scientific knowledge infrastructures
Research Infrastructures
Infrastructures
Publishers
Universities
Libraries
Institutions
Artefacts
101100
010110
100101
Data
Software
Instruments
Samples

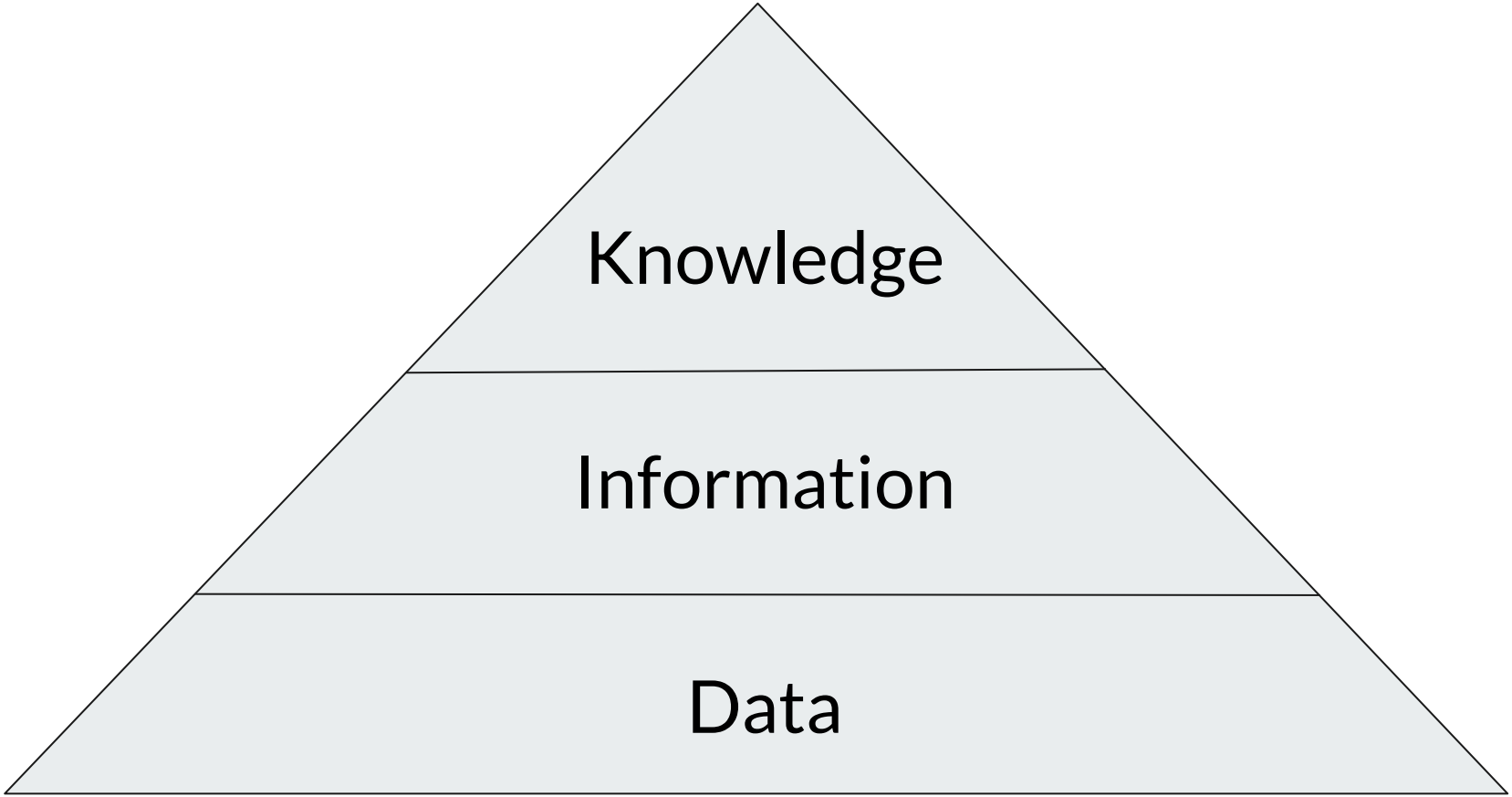


In Earth system science knowledge primarily about natural worlds


Data
111
001011
01100110
10011011011
01101100
101101
111
Knowledge


Observations
111
001011
01100110
10011011011
01101100
101101
111
Knowledge


ICOS
INTEGRATED
CARBON
OBSERVATION
SYSTEM
Knowledge
through
observations



https://en.wikipedia.org/wiki/DIKW_pyramid

It’s more complicated

21

21°C

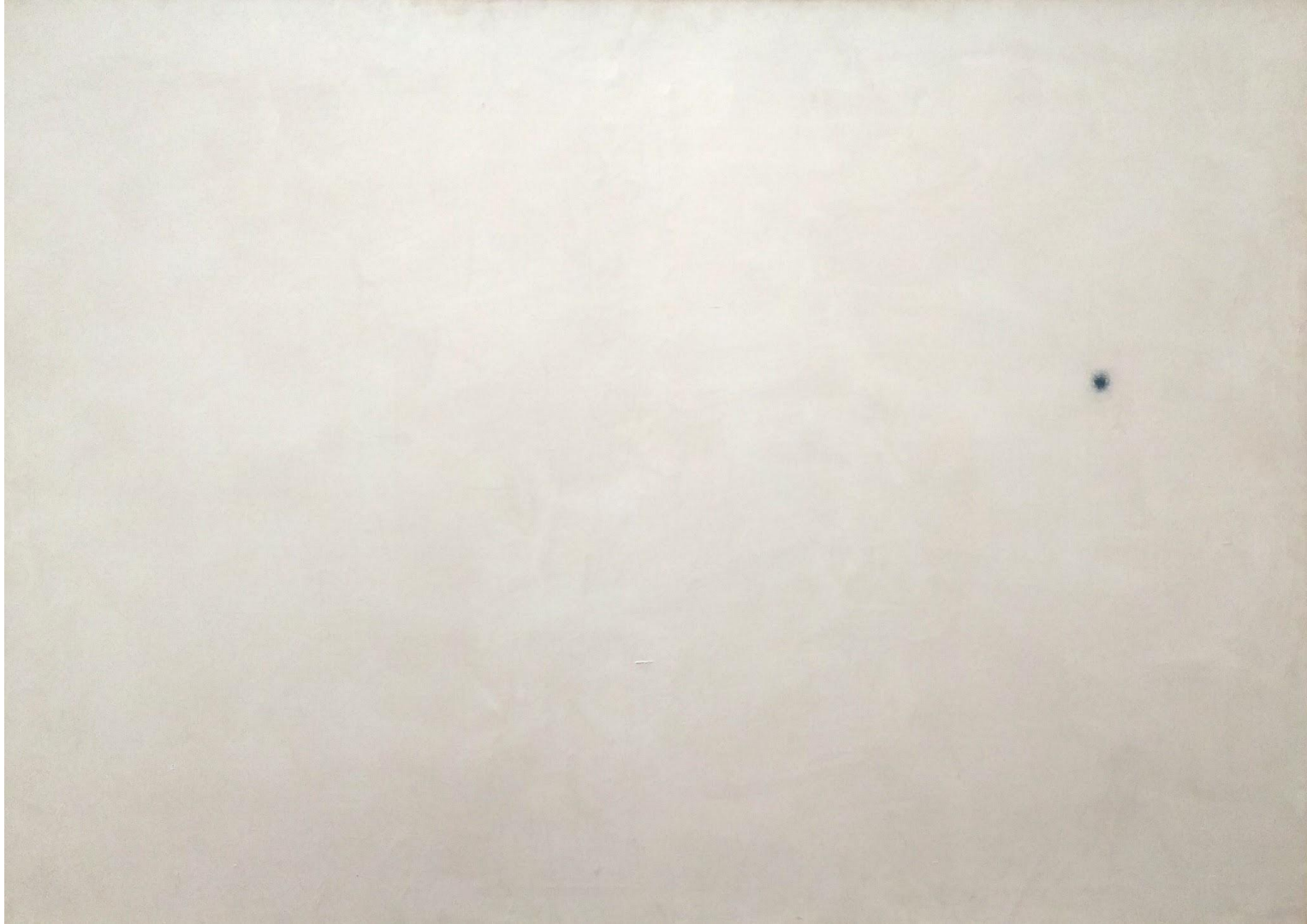
The temperature in my living room is 21°C

My thermometer observes that the temperature in my living room is 21°C

This is an observation
and it is information

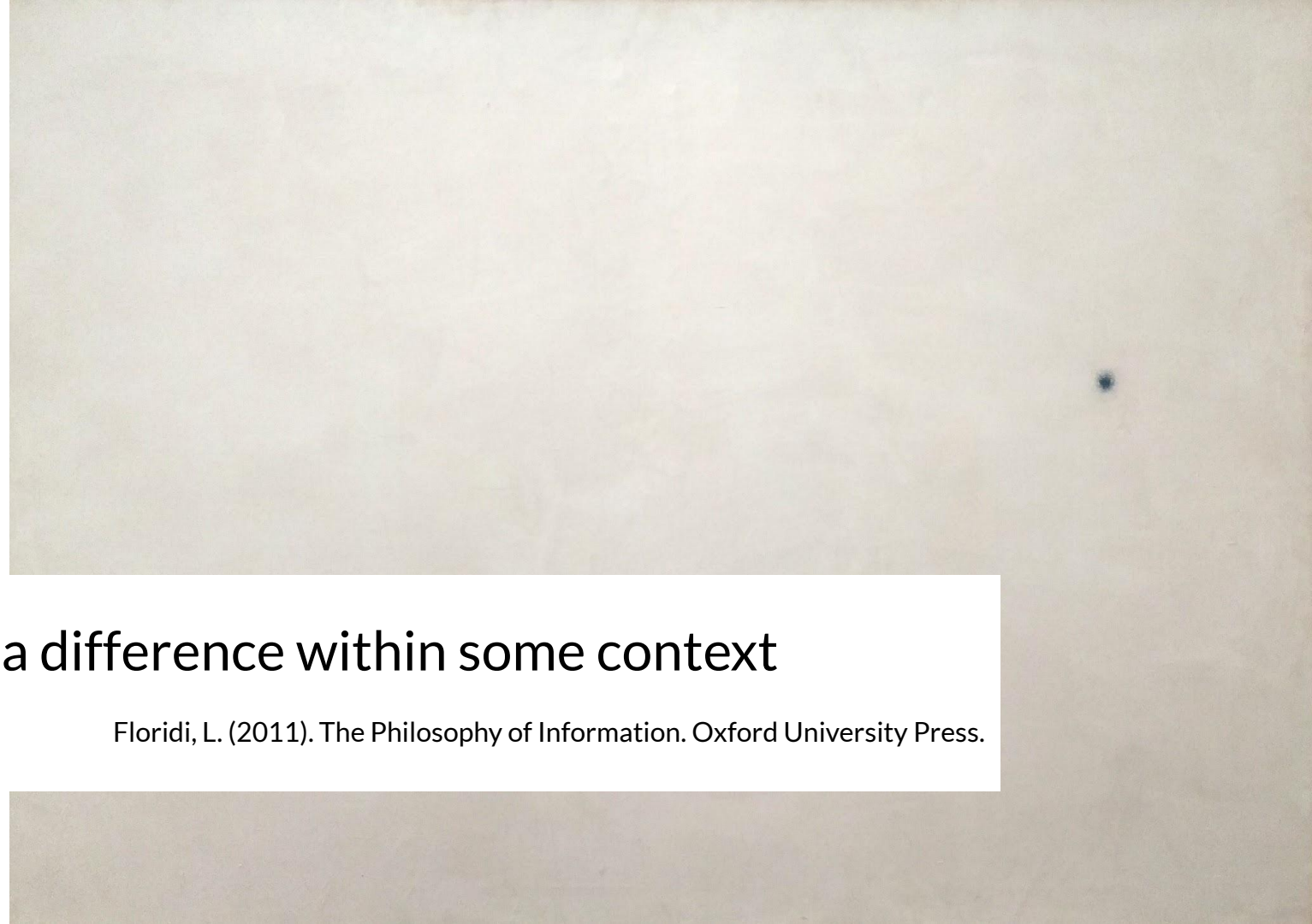
So, what are data, information, knowledge?

Datum



Joan Miró
Landscape
(1968)

Datum



A datum is a difference 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
  - Here, primary data used as indirect sources
  - About things other than those directly addressed by the primary data themselves

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

Information

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

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

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. https://doi.org/10.1016/0169-023X(95)00017-M

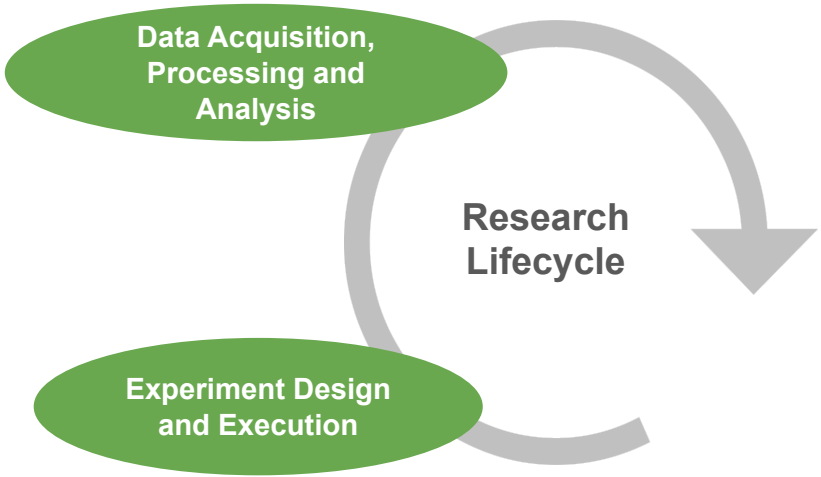
Knowledge

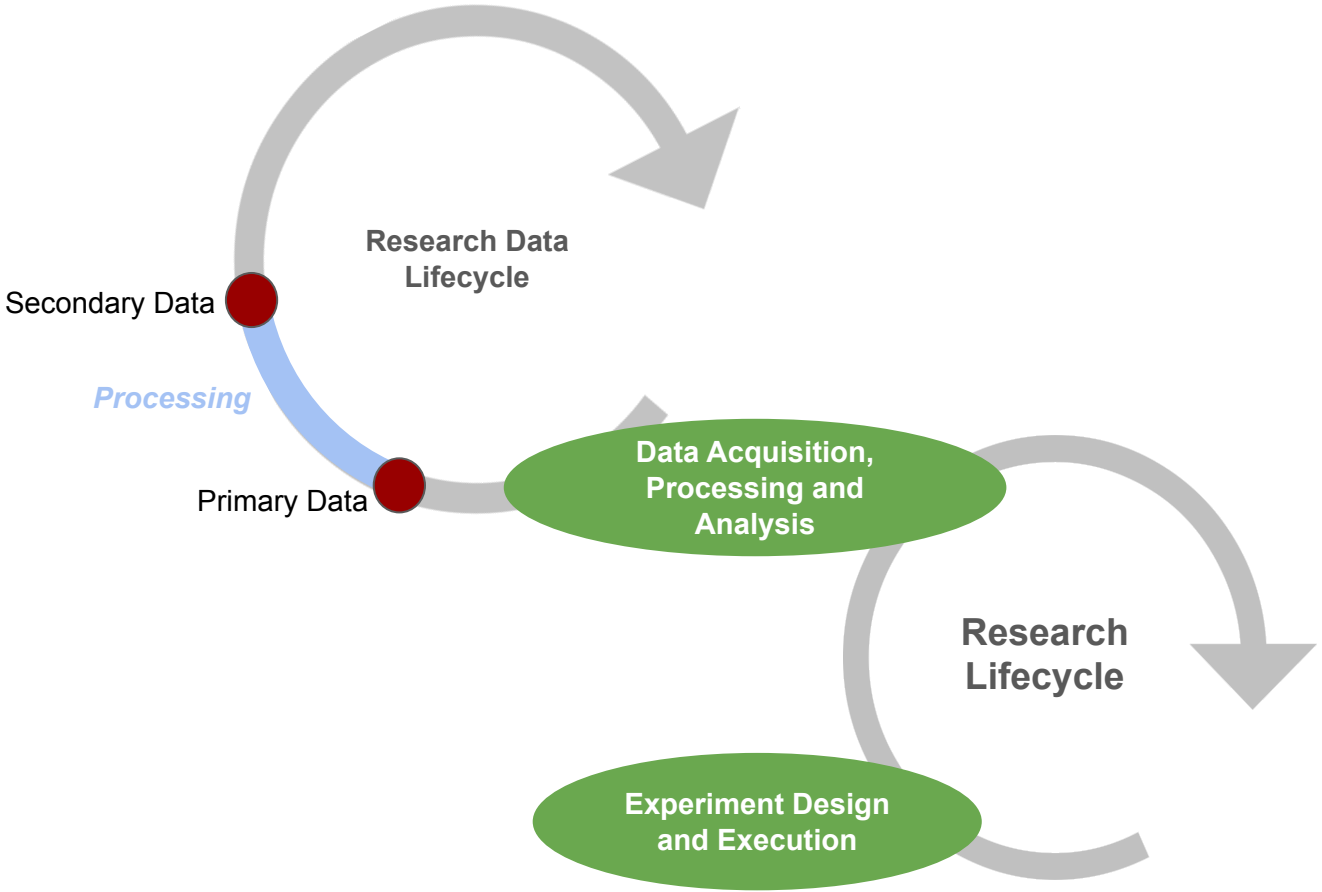
- Learned information
- Information incorporated in an agent’s reasoning resources
- Made ready for use within decision processes
- Output of learning processes
- Tacit knowledge is made explicit through externalization

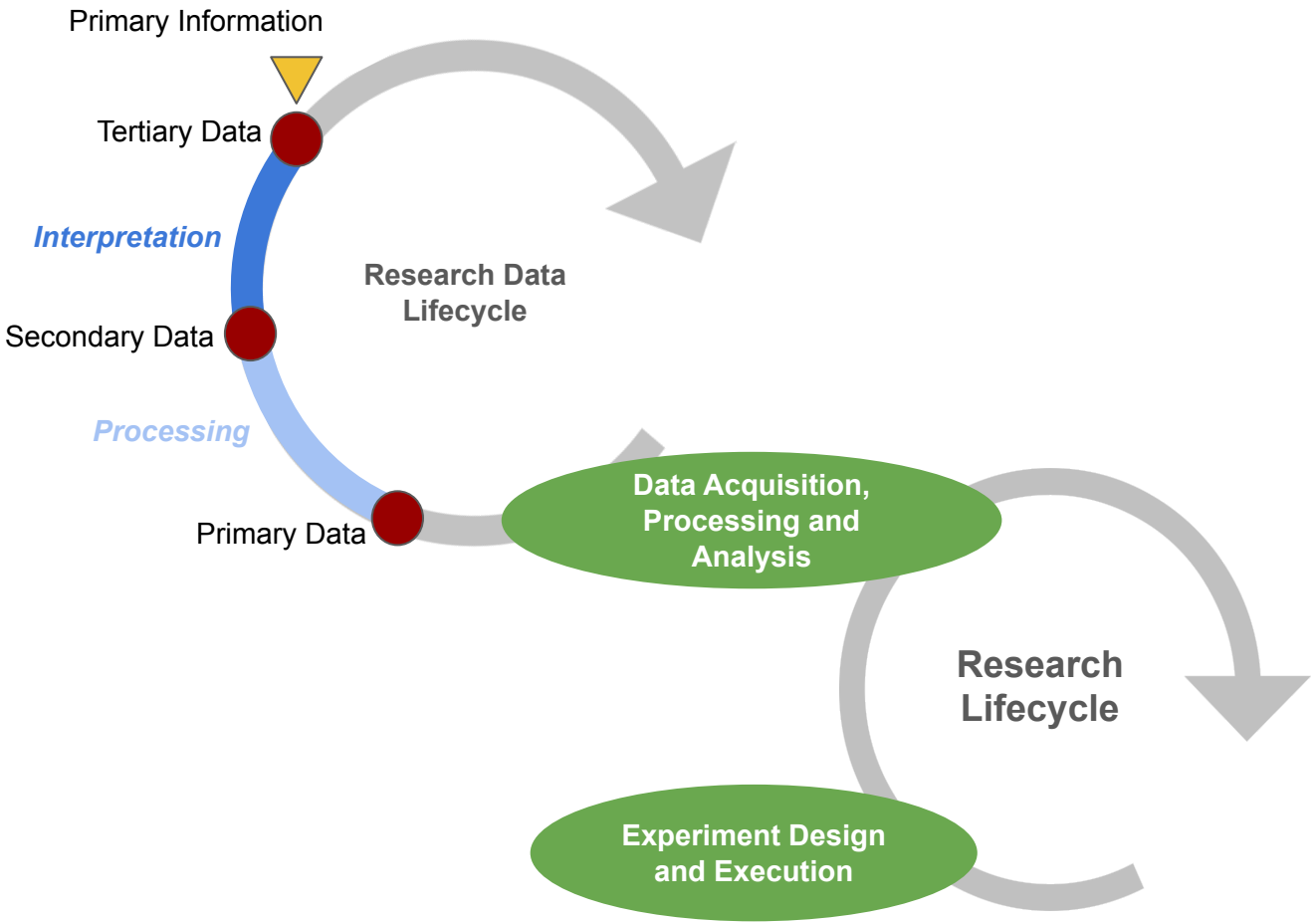
Aamodt, A. and Nygård, M. 1995. Different roles and mutual dependencies of data, information, and knowledge – An AI perspective on their integration. Data & Knowledge Engineering, 16(3): 191–222. https://doi.org/10.1016/0169-023X(95)00017-M

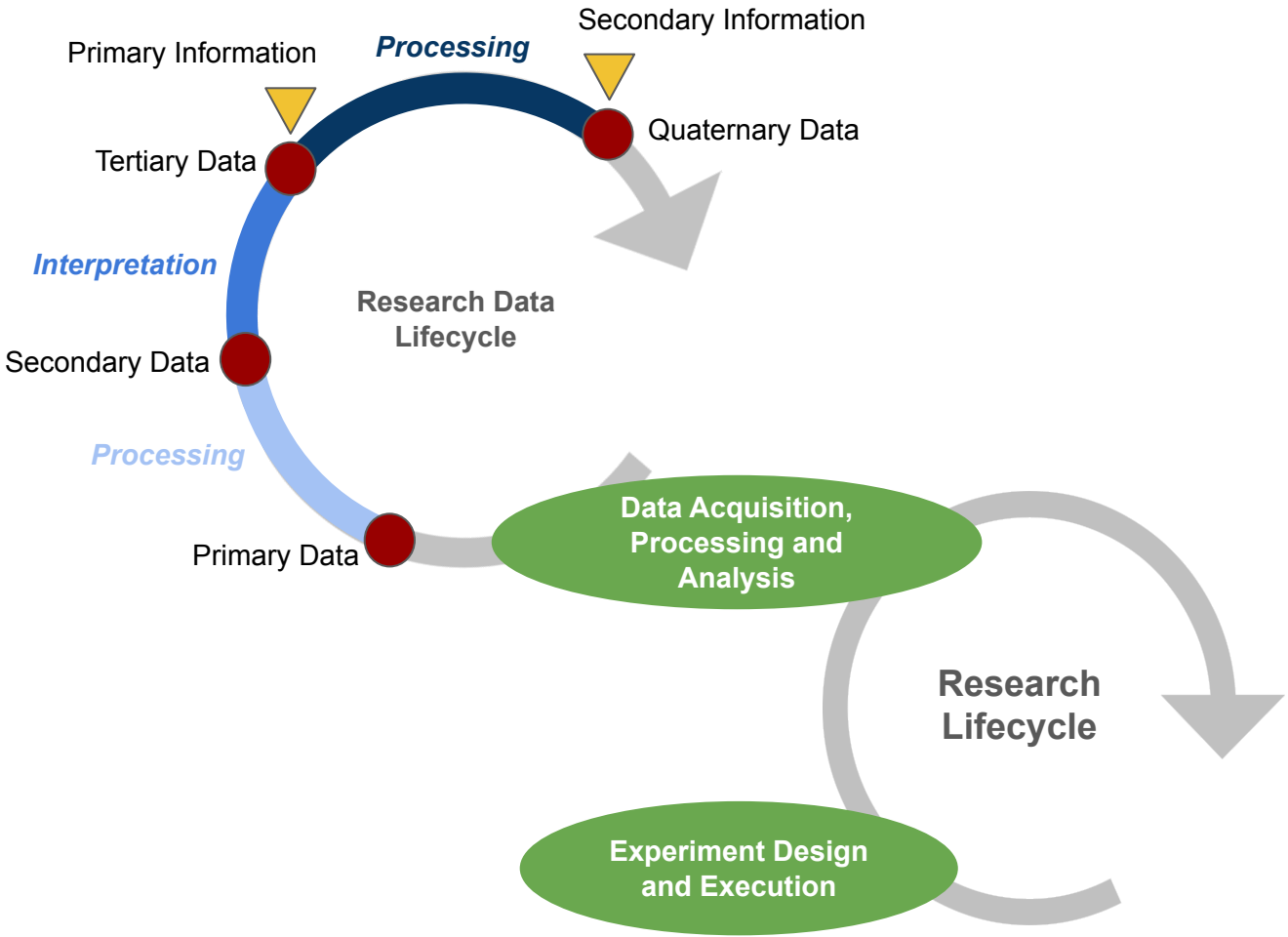
Chyi Lee, C. and Yang, J. 2000. Knowledge value chain. Journal of Management Development, 19(9), 783–794. https://doi.org/10.1108/02621710010378228

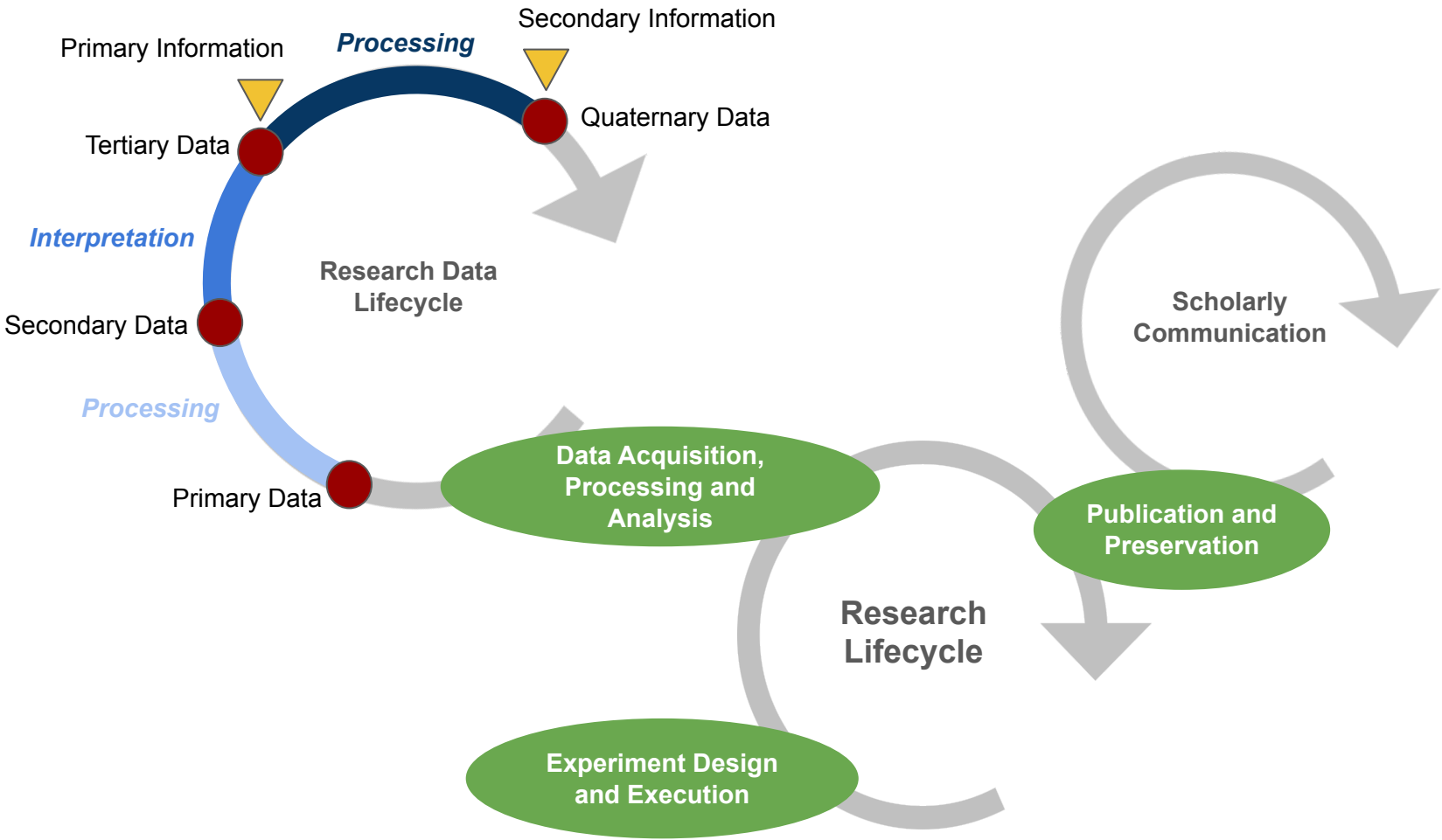
How does this apply to research?

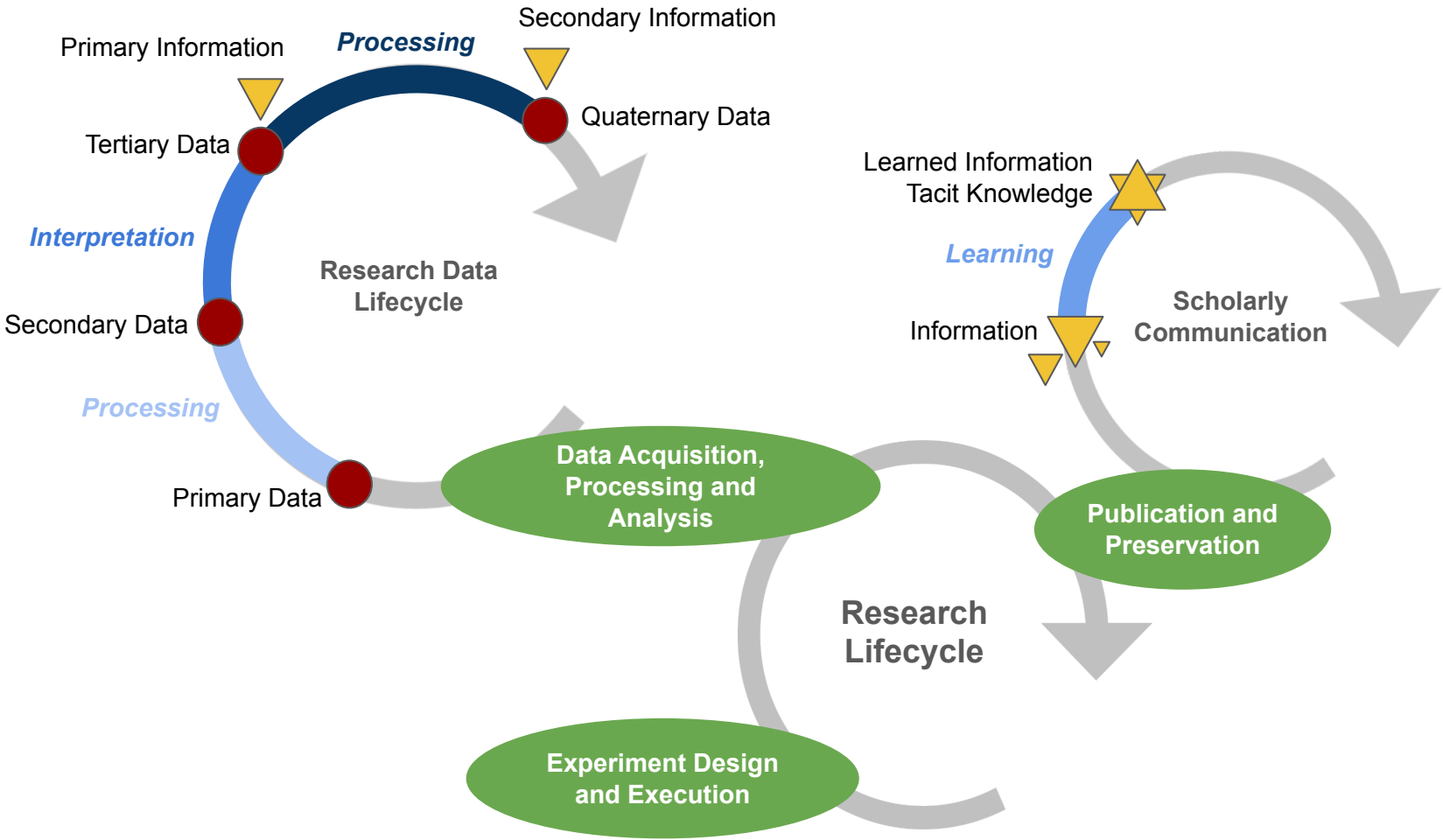

Data Acquisition, Processing and Analysis
Research Lifecycle
Experiment Design and Execution

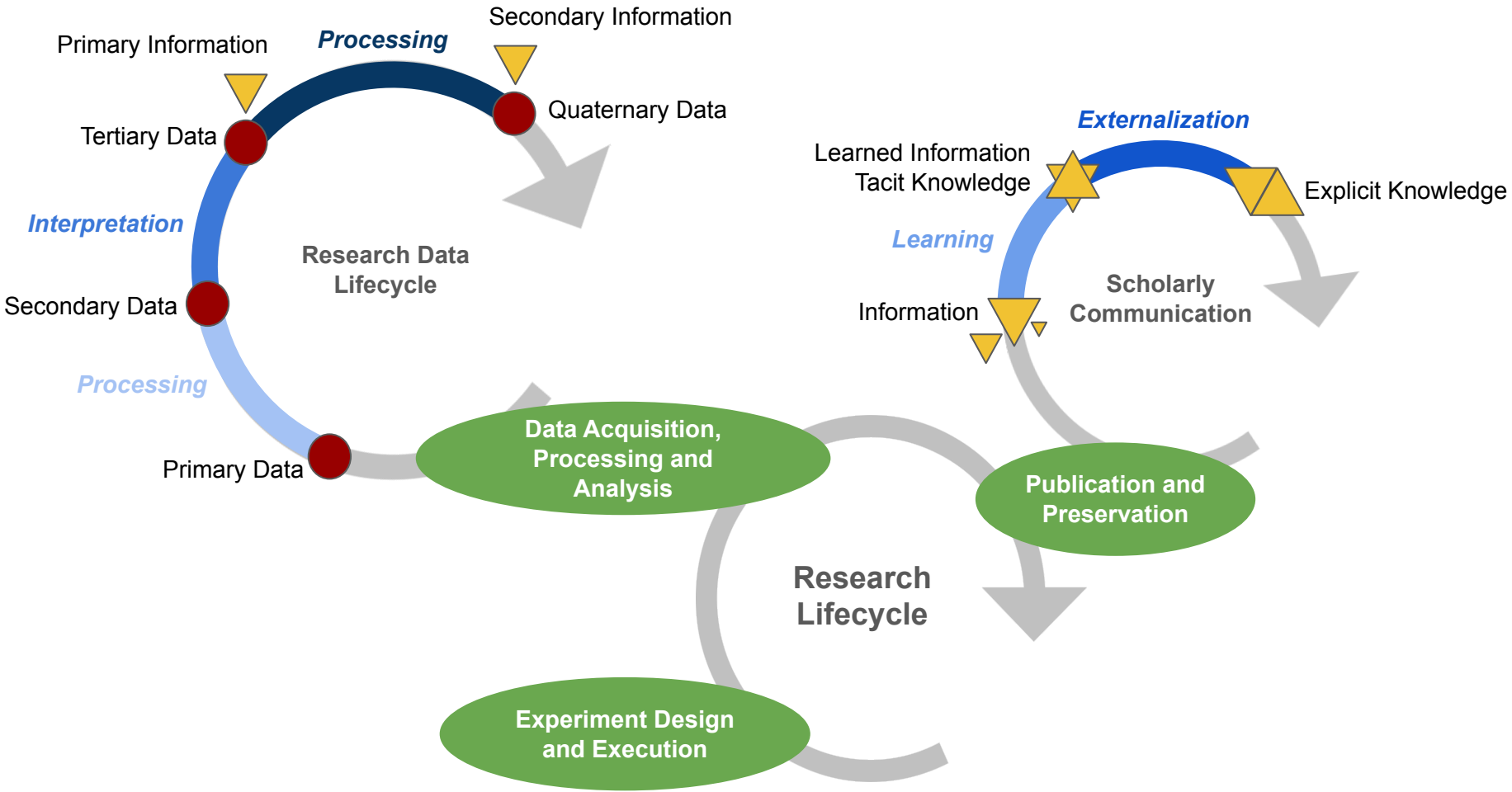

Research Data Lifecycle
Secondary Data
Processing
Primary Data
Data Acquisition, Processing and Analysis
Research Lifecycle
Experiment Design and Execution


Primary Information
Tertiary Data
Interpretation
Secondary Data
Processing
Primary Data
Research Data Lifecycle
Data Acquisition, Processing and Analysis
Research Lifecycle
Experiment Design and Execution


Primary Information
Processing
Secondary Information
Tertiary Data
Quaternary Data
Interpretation
Research Data Lifecycle
Secondary Data
Processing
Primary Data
Data Acquisition, Processing and Analysis
Research Lifecycle
Experiment Design and Execution


Primary Information
Processing
Secondary Information
Tertiary Data
Quaternary Data
Interpretation
Research Data Lifecycle
Secondary Data
Processing
Primary Data
Data Acquisition, Processing and Analysis
Publication and Preservation
Scholarly Communication
Research Lifecycle
Experiment Design and Execution


Primary Information
Processing
Secondary Information
Tertiary Data
Quaternary Data
Interpretation
Research Data Lifecycle
Secondary Data
Processing
Primary Data
Learned Information
Tacit Knowledge
Learning
Information
Scholarly Communication
Data Acquisition, Processing and Analysis
Publication and Preservation
Research Lifecycle
Experiment Design and Execution


Primary Information
Processing
Secondary Information
Quaternary Data
Tertiary Data
Interpretation
Research Data Lifecycle
Secondary Data
Processing
Primary Data
Data Acquisition, Processing and Analysis
Experiment Design and Execution
Research Lifecycle
Publication and Preservation
Externalization
Learned Information
Tacit Knowledge
Explicit Knowledge
Learning
Scholarly Communication
Information


PDF


CSV


JPEG

But for machines ...

… are **data**!

What the data mean is (sometimes more, sometimes less) clear to humans but for machines meaning is implicit (not expressly stated)

It is important for knowledge infrastructures to ensure meaning is explicit and formal

Information and knowledge-based systems

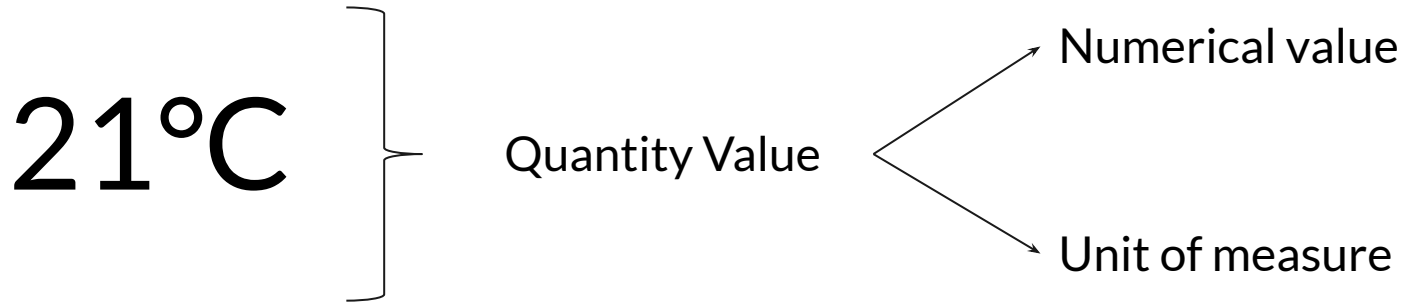
Not “just” databases

So, how do we represent meaning in information systems?

The technology framework

21°C

Temperature [°C]
21

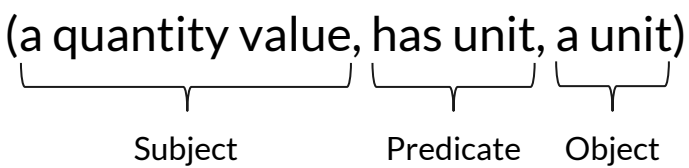

21°C
Quantity Value
Numerical value
Unit of measure

Numerical value: 21

Unit of measure: http://qudt.org/vocab/unit/DEG_C

Resource Description Framework (RDF)

- A W3C Recommendation and technology developed within the Semantic Web Activity
- A Statement-centric data model
- A Statement is a triple structure consisting of a subject, predicate and object



Resource Description Framework (RDF)

- Subject, predicate and object are Resources identified by Uniform Resource Identifier (URI)
- Object can be a literal value such as a number or a string
- URIs are typically prefixed, e.g. “qudt:” for http://qudt.org/schema/qudt/
- The resulting (RDF) data can be processed, e.g. querying using SPARQL

(http://qudt.org/vocab/unit/DEG_C,
http://www.w3.org/1999/02/22-rdf-syntax-ns#type,
http://qudt.org/schema/qudt/Unit)

Subject

Predicate

Object

(unit:DEG_C, rdf:type, qudt:Unit)

With prefixes

My thermometer observes that the temperature in my living room is 21°C

```
:obs1 rdf:type sosa:Observation .
:obs1 sosa:madeBySensor :thermometer1 .
:obs1 sosa:observedProperty :temperature .
:obs1 sosa:hasFeatureOfInterest :room1 .
:obs1 sosa:hasResult :qv1 .
:qv1 qudt:numericValue “21” .
:qv1 qudt:unit unit:DEG_C .
```

My thermometer observes that the temperature in my living room is 21°C

The meaning of data is now explicit and formal

Meaning is formalized using unambiguous and identified terms of shared vocabularies

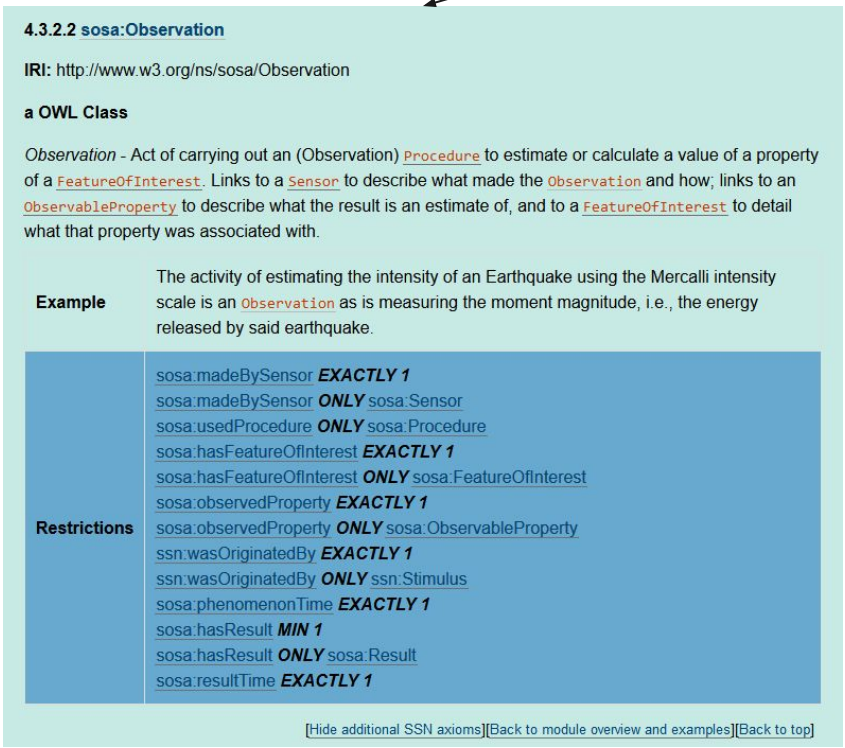
Vocabularies

- Concepts and relations (qudt:Unit, sosa:Observation, rdf:type, etc.) are constituents of vocabularies
- Also known as terminologies, controlled vocabularies, thesauri, taxonomies, ontologies
- They are most useful if sociotechnical systems widely share them
- This is key to attain *semantic* data interoperability
- There exist numerous languages to specify vocabularies (e.g. RDFS, OWL)
- These enable stating that a Resource is a concept or a relation, a sub concept of another concept, ...

http://www.w3.org/ns/sosa/Observation

Lookup by human

Lookup by machine



4.3.2.2 sosa:Observation

IRI: http://www.w3.org/ns/sosa/Observation

a OWL Class

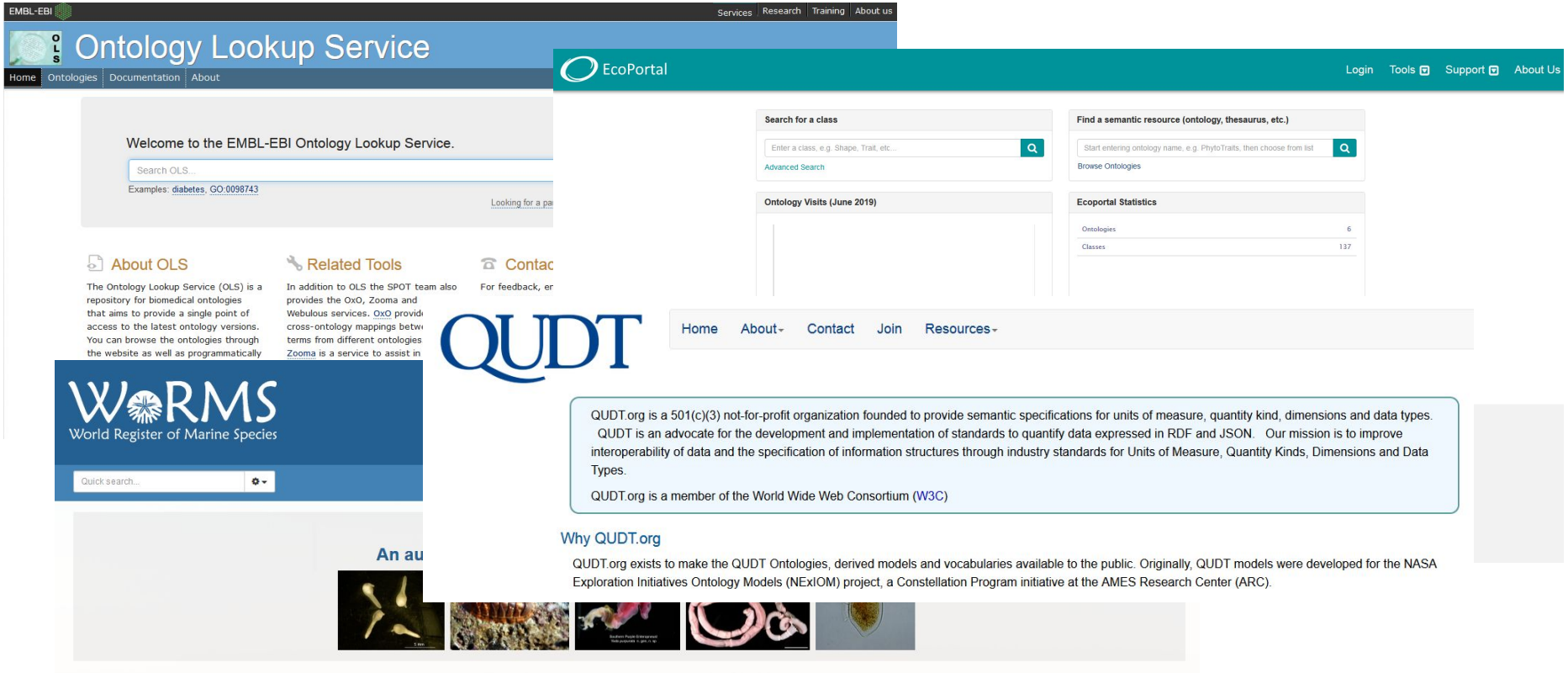
Observation - Act of carrying out an (Observation) `Procedure` to estimate or calculate a value of a property of a `FeatureOfInterest`. Links to a `Sensor` to describe what made the `Observation` and how; links to an `ObservableProperty` to describe what the result is an estimate of, and to a `FeatureOfInterest` to detail what that property was associated with.

Example	The activity of estimating the intensity of an Earthquake using the Mercalli intensity scale is an `Observation` as is measuring the moment magnitude, i.e., the energy released by said earthquake.
Restrictions	sosa:madeBySensor ***EXACTLY 1*** sosa:madeBySensor ***ONLY*** sosa:Sensor sosa:usedProcedure ***ONLY*** sosa:Procedure sosa:hasFeatureOfInterest ***EXACTLY 1*** sosa:hasFeatureOfInterest ***ONLY*** sosa:FeatureOfInterest sosa:observedProperty ***EXACTLY 1*** sosa:observedProperty ***ONLY*** sosa:ObservableProperty ssn:wasOriginatedBy ***EXACTLY 1*** ssn:wasOriginatedBy ***ONLY*** ssn:Stimulus sosa:phenomenonTime ***EXACTLY 1*** sosa:hasResult ***MIN 1*** sosa:hasResult ***ONLY*** sosa:Result sosa:resultTime ***EXACTLY 1***

[Hide additional SSN axioms][Back to module overview and examples][Back to top]

sosa:Observation rdf:type owl:Class .
sosa:Observation rdfs:label “Observation” .
…

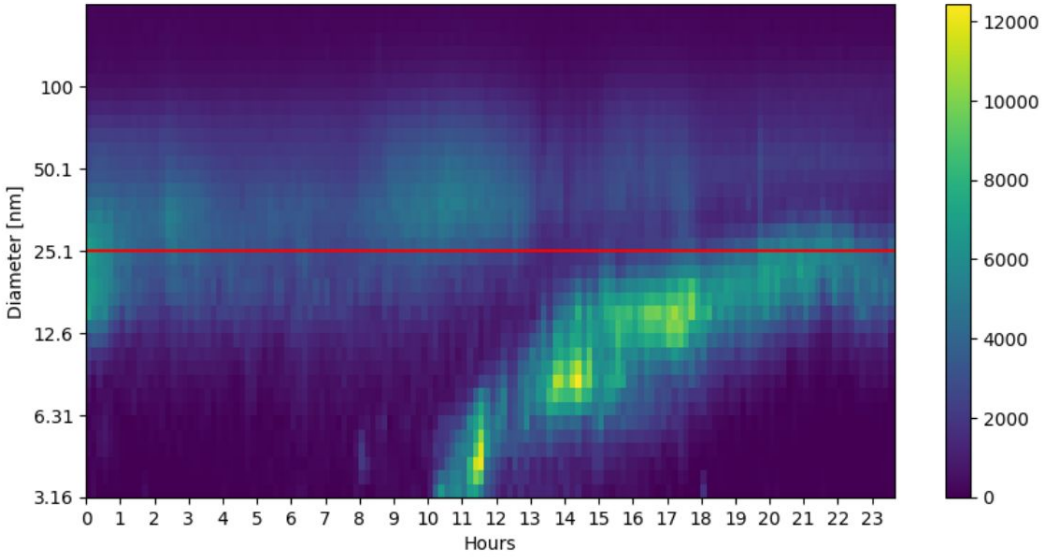
How do I find the vocabulary I need?

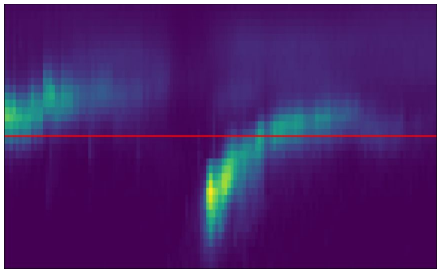


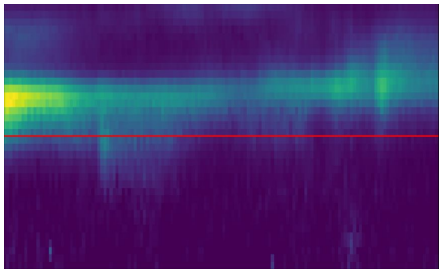
Part II

A knowledge infrastructure in atmospheric physics as a case in point

SMEAR


Diameter [nm]
100
50.1
25.1
12.6
6.31
3.16
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hours
12000
10000
8000
6000
4000
2000
0





Measured is particle size distribution, the primary observational data.

Described are new particle formation events, the derived data (actually, information)

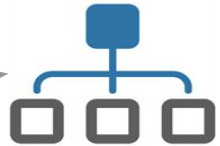
Let’s look at how two research groups implement data analysis.

Is the meaning of data explicit, unambiguous, formal?

Research Group A

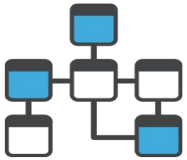
Access **observational data** published by research infrastructure

SMEAR CSV



S String
N Number
T/F Boolean

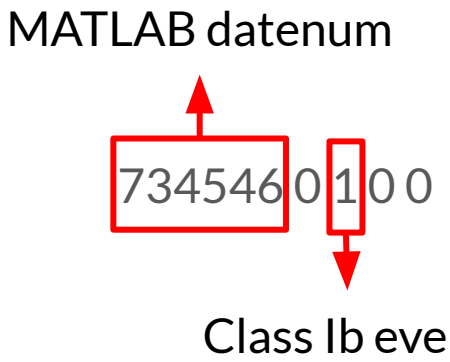
MATLAB



A Array
O Object
D Date

Research Group B

734545 0 0 0 1
734546 0 1 0 0
734547 0 0 0 1
734548 0 0 0 1
734549 0 0 0 1
734550 0 0 0 1
734551 0 0 0 1
734552 0 0 1 0



Encoding of data describing NPFE
by Research Group A

2011-07-04,2
2011-07-04,2
2011-07-05,4
2011-07-06,4
2011-07-07,4
2011-07-08,4
2011-07-09,NE
2011-07-10,BD

Encoding of data describing NPFE
by Research Group B

Classifications

Research Group A

Class I: parameters derived with good confidence
Class Ia: very clear and strong NPF
Class Ib: other Class I events
Class II: parameter derivation not possible
Non Event
Undefined
Bad Data
Partly Bad Data

dal Maso et al. (2005) BOREAL ENV. RES. Vol. 10:323-336

Research Group B

Class 0: Undefined
Class 1: Strong NPF
Class 2: Intermediate NPF
Class 3: Weak NPF
Class 4: ?
NE: Non Event
BD: Bad Data

Hamed et al. (2007) Atmos. Chem. Phys. 7:355-376

Problem: Heterogeneous Derived Data

734545 0 0 0 1	2011-07-04,2
734546 0 1 0 0	2011-07-04,2
734547 0 0 0 1	2011-07-05,4
734548 0 0 0 1	2011-07-06,4
734549 0 0 0 1	2011-07-07,4
734550 0 0 0 1	2011-07-08,4
734551 0 0 0 1	2011-07-09,NE
734552 0 0 1 0	2011-07-10,BD

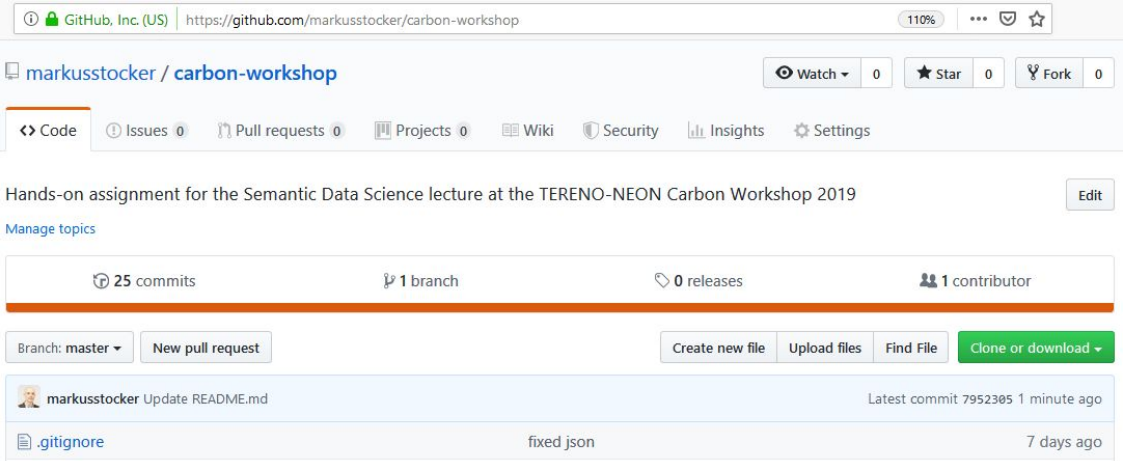
Problem: Implicit Semantics

734545 0 0 0 1
734546 0 1 0 0
734547 0 0 0 1
734548 0 0 0 1
734549 0 0 0 1
734550 0 0 0 1
734551 0 0 0 1
734552 0 0 1 0

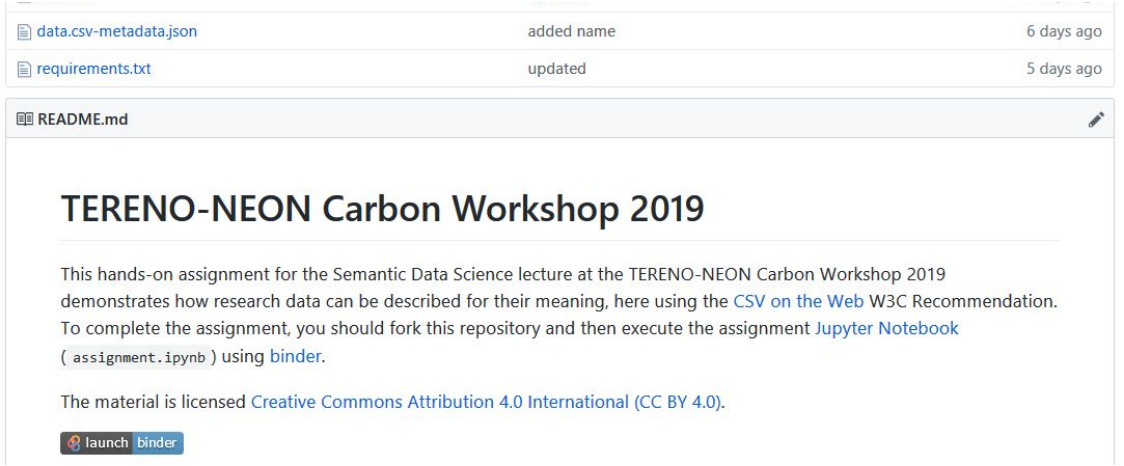
Can we improve this?

Part III

Hands-on exercise



https://github.com/markusstocker/carbon-workshop

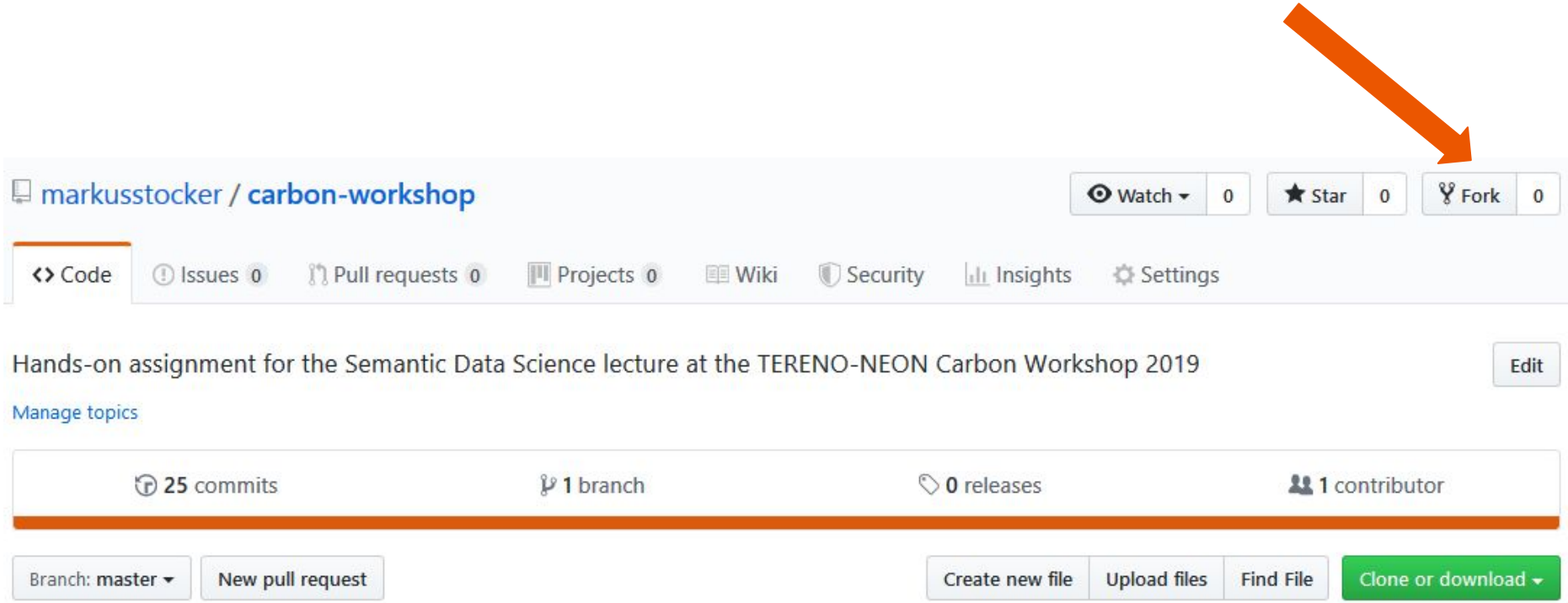


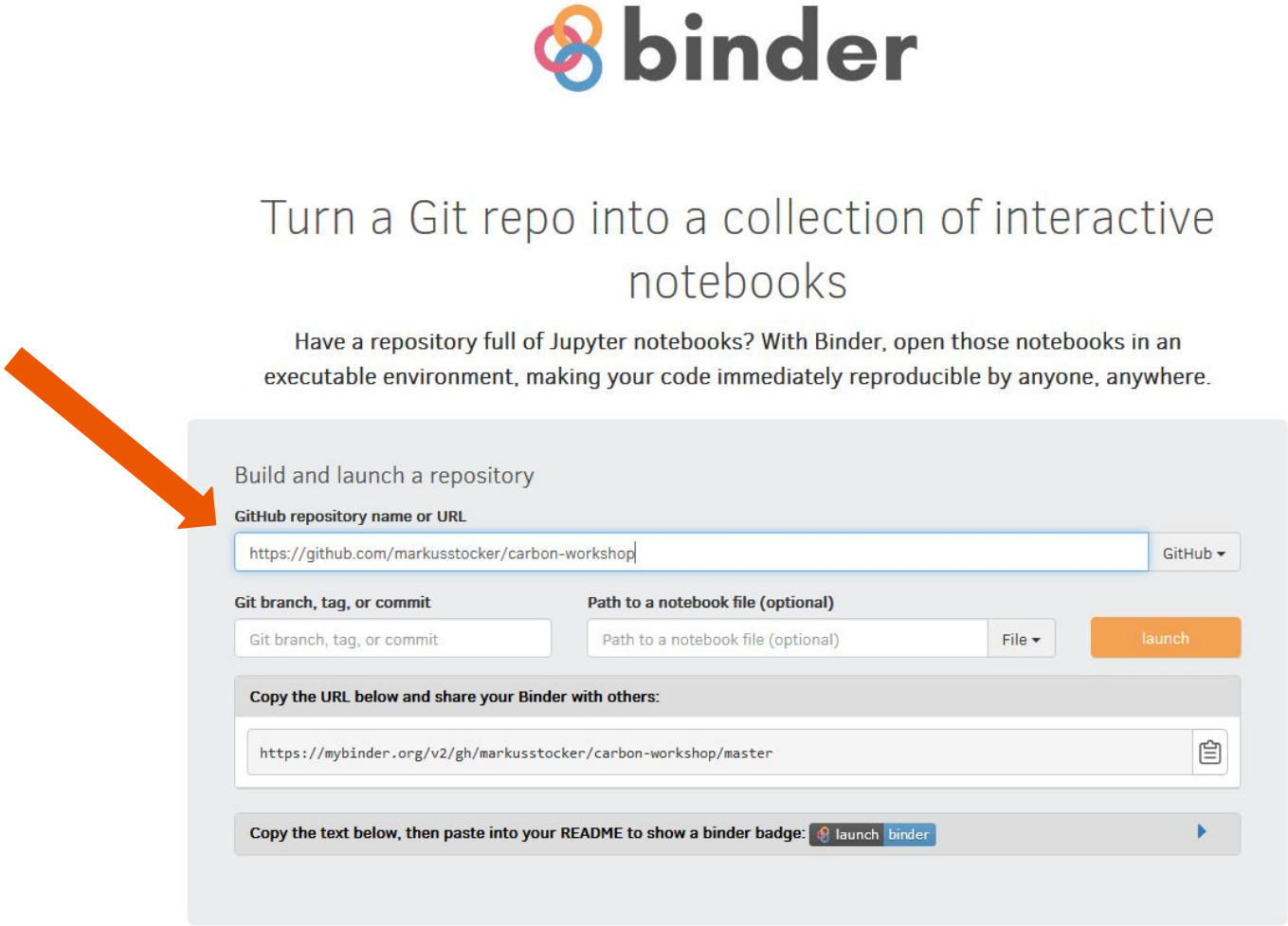
Aims

- Building on the presented case in atmospheric physics
- Interpret observational data and generate derivative data about events, in CSV
- Inspect the corresponding “metadata” that specifies data semantics
- Convert the CSV data into RDF as a format with explicit and formal data semantics
- Understand the resulting RDF data
- Formulate SPARQL queries on RDF to demonstrate a kind of RDF data processing

Steps

- Fork the repository https://github.com/markusstocker/carbon-workshop
- Use http://mybinder.org to launch the forked repository
- Wait until Jupyter Notebook (Lab) is launched
- Open assignment.ipynb and follow the instructions


markusstocker / carbon-workshop
Watch 0
Star 0
Fork 0
Code
Issues 0
Pull requests 0
Projects 0
Wiki
Security
Insights
Settings
Hands-on assignment for the Semantic Data Science lecture at the TERENO-NEON Carbon Workshop 2019
Edit
Manage topics
25 commits
1 branch
0 releases
1 contributor
Branch: master
New pull request
Create new file
Upload files
Find File
Clone or download


binder
Turn a Git repo into a collection of interactive notebooks
Have a repository full of Jupyter notebooks? With Binder, open those notebooks in an executable environment, making your code immediately reproducible by anyone, anywhere.
Build and launch a repository
GitHub repository name or URL
https://github.com/markusstocker/carbon-workshop
GitHub
Git branch, tag, or commit
Path to a notebook file (optional)
Git branch, tag, or commit
Path to a notebook file (optional)
File
launch
Copy the URL below and share your Binder with others:
https://mybinder.org/v2/gh/markusstocker/carbon-workshop/master
Copy the text below, then paste into your README to show a binder badge:
launch
binder

https://hub.gke.mybinder.org/user/markusstocker-carbon-workshop-6yhy7k50/notebooks/assignment.ipynb

jupyter assignment (autosaved)

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3

Run Markdown

NPFE Data with CSV on the Web

This notebooks exemplifies how data about new particle formation events can be described using CSV on the Web.

We interpret particle size distribution data (primary data) as measured by an observation system of the SMEAR research infrastructure in order to detect the occurrence of new particle formation events on selected days in Hyytiälä, Finland. Detected events are then described, whereby we generate secondary (derivative) data about events. These data are stored to disk in CSV format.

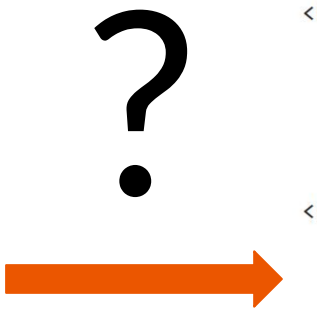
In addition, we use CSV on the Web to describe the secondary CSV data using common, shared terminology that is unambiguously identified and described on the web. This makes the secondary CSV data more interoperable, reusable and understandable to machines. For instance, we can use the description to transform the CSV data into RDF and then leverage SPARQL to query this data.

Before we start, we need to load required Python modules as well a few user-defined functions.

In [ ]:
```
import requests, io, os, pandas as pd, numpy as np
from urllib.parse import urlencode
from pytz import timezone
from datetime import datetime, timedelta
from csvwlib import CSVWConverter
from matplotlib import pyplot as plt
from rdflib.plugins.sparql.results.csvresults import CSVResultSerializer

def fetch(date):
    time_from = timezone('Europe/Helsinki').localize(datetime.strptime(date, '%Y-%m-%d'))
    time_to = time_from + timedelta(days=1)

    query = {
        'table': 'HYY_DMPS', 'quality': 'ANY', 'averaging': 'NONE', 'type': 'NONE',
        'from': str(time_from), 'to': str(time_to), 'variables': 'd316e1,d355e1,d398e1,'\
        'd447e1,d501e1,d562e1,d631e1,d708e1,d794e1,d891e1,d100e2,d112e2,d126e2,d141e2,d158e2,'\
        'd178e2,d200e2,d224e2,d251e2,d282e2,d316e2,d355e2,d398e2,d447e2,d501e2,d562e2,d631e2,'\
        'd708e2,d794e2,d891e2,d100e3,d112e3,d126e3,d141e3,d158e3,d178e3,d200e3'
    }

    url = 'https://avaa.tdata.fi/smear-services/smeardata.jsp?' + urlencode(query)
    response = requests.post(url)

    return pd.read_csv(io.StringIO(response.text))


def plot(data):
    d = data.copy(deep=True)
    d = d.iloc[:, 6:].values
    m = len(d)
    n = len(d[0])
    x = range(0, m)
    y = range(0, n)
    x, y = np.meshgrid(x, y)
    z = np.transpose(np.array([row[1:] for row in d]).astype(np.float))
    plt.figure(figsize=(10, 5), dpi=100)
    plt.pcolormesh(x, y, z)
```

```
{
    "@context": "http://www.w3.org/ns/csvw",
    "url": "data.csv",
    "tableSchema": {
        "aboutUrl": "http://avaa.tdata.fi/{date}",
        "columns": [{
          "name": "date",
          "propertyUrl": "http://purl.obolibrary.org/obo/STATO_0000093",
          "dc:description": "The ISO date at which the event occurred",
          "datatype": {
              "base": "date",
              "format": "yyyy-MM-dd"
          }
        },{
          "name": "start",
          "propertyUrl": "http://purl.obolibrary.org/obo/RO_0002537",
          "dc:description": "The ISO event start time",
          "datatype": {
              "base": "time",
              "format": "HH:mm"
          }
        },{
          "name": "end",
          "propertyUrl": "http://purl.obolibrary.org/obo/RO_0002538",
          "dc:description": "The ISO event end time",
          "datatype": {
              "base": "time",
              "format": "HH:mm"
          }
        },{
          "name": "class",
          "propertyUrl": "http://purl.obolibrary.org/obo/OBI_0000999",
          "dc:description": "The event classification according to the dal Maso et al. scheme",
          "datatype": {
              "base": "string",
              "format": "Ia|Ib|II"
          },
          "valueUrl": "http://avaa.tdata.fi/class/{class}"
        },{
          "name": "type",
          "virtual": true,
          "propertyUrl": "rdf:type",
          "valueUrl": "http://purl.obolibrary.org/obo/ENVO_01001372"
        }]
    }
}
```

date	start	end	class
2007-05-05	12:00	14:00	Ib
2007-04-15	11:00	12:00	Ia
2013-04-04	10:00	11:30	Ia



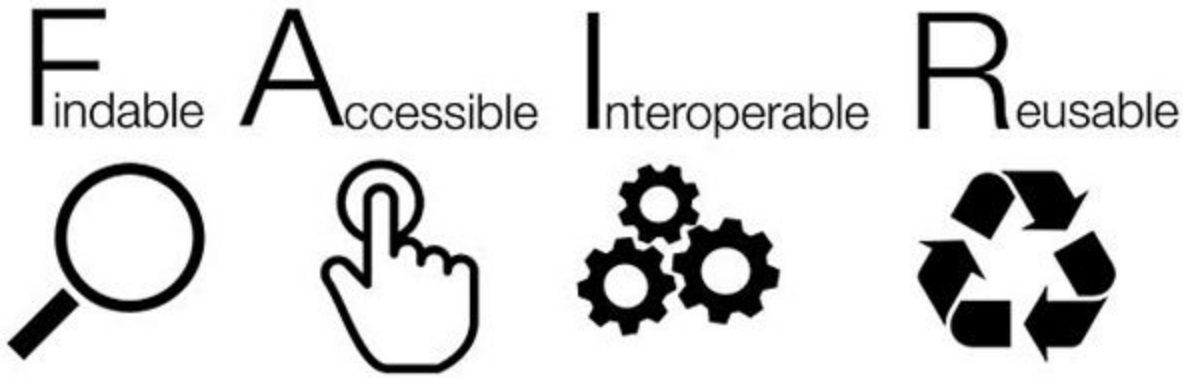
```
<http://avaa.tdata.fi/2007-04-15> a ns1:ENVO_01001372 ;
    ns1:OBI_0000999 <http://avaa.tdata.fi/class/Ia> ;
    ns1:RO_0002537 "11:00:00"^^xsd:time ;
    ns1:RO_0002538 "12:00:00"^^xsd:time ;
    ns1:STATO_0000093 "2007-04-15"^^xsd:date .

<http://avaa.tdata.fi/2007-05-05> a ns1:ENVO_01001372 ;
    ns1:OBI_0000999 <http://avaa.tdata.fi/class/Ib> ;
    ns1:RO_0002537 "12:00:00"^^xsd:time ;
    ns1:RO_0002538 "14:00:00"^^xsd:time ;
    ns1:STATO_0000093 "2007-05-05"^^xsd:date .

<http://avaa.tdata.fi/2013-04-04> a ns1:ENVO_01001372 ;
    ns1:OBI_0000999 <http://avaa.tdata.fi/class/Ia> ;
    ns1:RO_0002537 "10:00:00"^^xsd:time ;
    ns1:RO_0002538 "11:30:00"^^xsd:time ;
    ns1:STATO_0000093 "2013-04-04"^^xsd:date .
```

Part IV

Discussion

How are data semantics explicit and formal here?

```
<http://avaa.tdata.fi/2007-04-15> a ns1:ENVO_01001372 ;
    ns1:OBI_0000999 <http://avaa.tdata.fi/class/Ia> ;
    ns1:RO_0002537 "11:00:00"^^xsd:time ;
    ns1:RO_0002538 "12:00:00"^^xsd:time ;
    ns1:STATO_0000093 "2007-04-15"^^xsd:date .

<http://avaa.tdata.fi/2007-05-05> a ns1:ENVO_01001372 ;
    ns1:OBI_0000999 <http://avaa.tdata.fi/class/Ib> ;
    ns1:RO_0002537 "12:00:00"^^xsd:time ;
    ns1:RO_0002538 "14:00:00"^^xsd:time ;
    ns1:STATO_0000093 "2007-05-05"^^xsd:date .

<http://avaa.tdata.fi/2013-04-04> a ns1:ENVO_01001372 ;
    ns1:OBI_0000999 <http://avaa.tdata.fi/class/Ia> ;
    ns1:RO_0002537 "10:00:00"^^xsd:time ;
    ns1:RO_0002538 "11:30:00"^^xsd:time ;
    ns1:STATO_0000093 "2013-04-04"^^xsd:date .
```



Wilkinson, M. D. *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 https://doi.org/10.1038/sdata.2016.18

Do you publish your (CSV) research data?

Are you going to use these technologies?