Science as a crowd sourced enterprise

Metadata Workshop, June 2019

Mark Greiner

Max-Planck Institute for Chemical Energy Conversion

Contents

1. Why I think Science is a crowd-sourced enterprise
2. Strategy to get the crowd providing FAIR data
3. Community research plan
4. Bonus: Proposed model on sample ontology

The original ideology of published science

- Why was science first published?

First scientific journal

- Ca. 1665 “Philosophical Transactions of the Royal Society”

http://www.mhs.ox.ac.uk/sis25/image.php?id=49

Motivation for First scientific journal

- produce a *collective* body of knowledge
- a public record of original contributions

- "a scientific news service"

- facilitate communication between individuals who had worked in isolation from one another

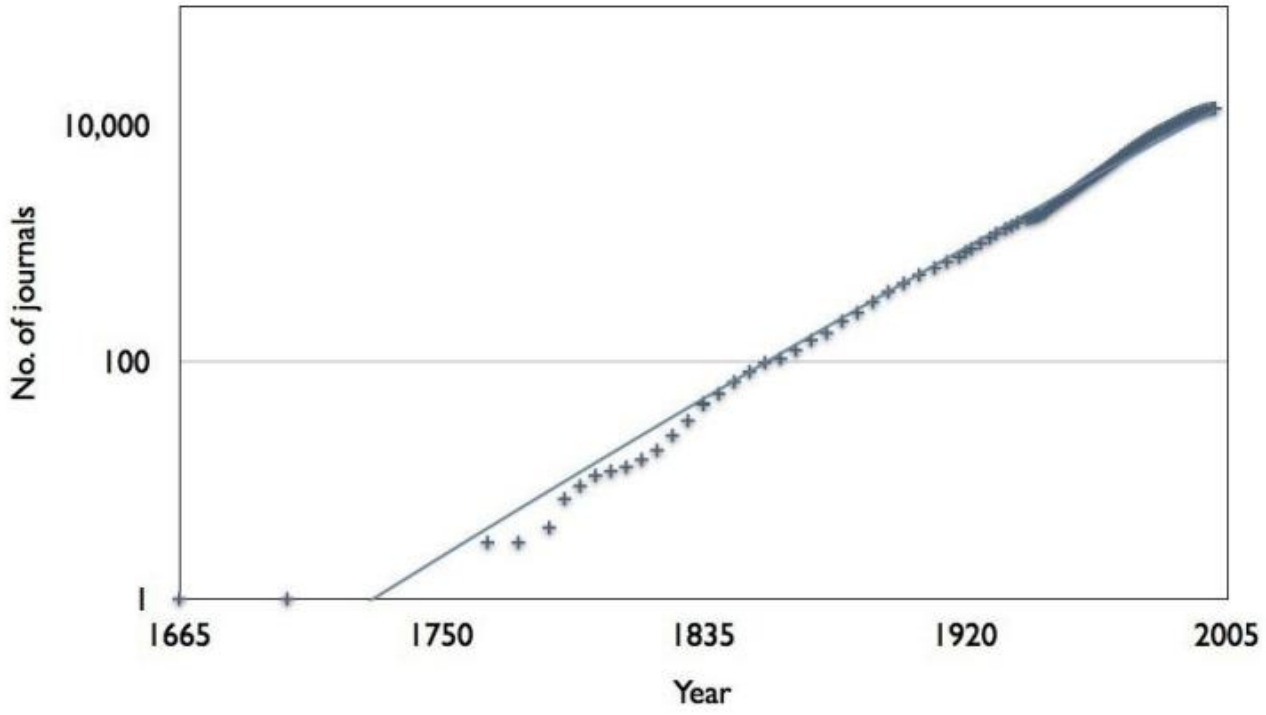
"The purposes and responsibilities for sharing" https://www.nap.edu/read/10613/chapter/4

First scientific journal

- Required author registration.
- Note time of new claim
- Peer reviewed
- Dissemination (via print) and archiving

Scientific publishing 300 years later

- 28 000 Scientific Journals
- 1.8 M articles per year



“The STM Report” Ware, Mabe (2012)

Think about how the world has changed in 300 years

- Globalization
- Internet
- Speed and reach of communication
- New ways of disseminating
  - Blogs, tutorials
- New ways of verifying claims
  - Eg. Forums, Reddit

Science as a crow-sourced enterprise

- *"science is fundamentally a cumulative enterprise. Each new discovery plays the role of one more brick in an edifice."*
  - Eric Lander

- *"For centuries scientists have relied on each other, on the self correcting mechanisms intrinsic to the nature of science, and on the traditions of their community to safeguard the integrity of the research process."*
  - (NRC, 1992)

"The purposes and responsibilities for sharing" https://www.nap.edu/read/10613/chapter/4

Open culture in Software and electronics

- Code sharing, Github, open source, Stack overflow…
  - You can learn anything. These skills are not protected and hidden from the community.
  - Big software firms have huge benefit from this.
- IoT
  - Raspberry Pi, Arduino
  - Lots of open projects, blogs,
  - Money is made by selling components
  - Industry has benefit of getting highly trained people, and also gets new ideas from community

Problems with current publishing culture

- Bias toward positive results
- “decorating” scientific results in embellishments about importance of the work
- These are problems could be solved

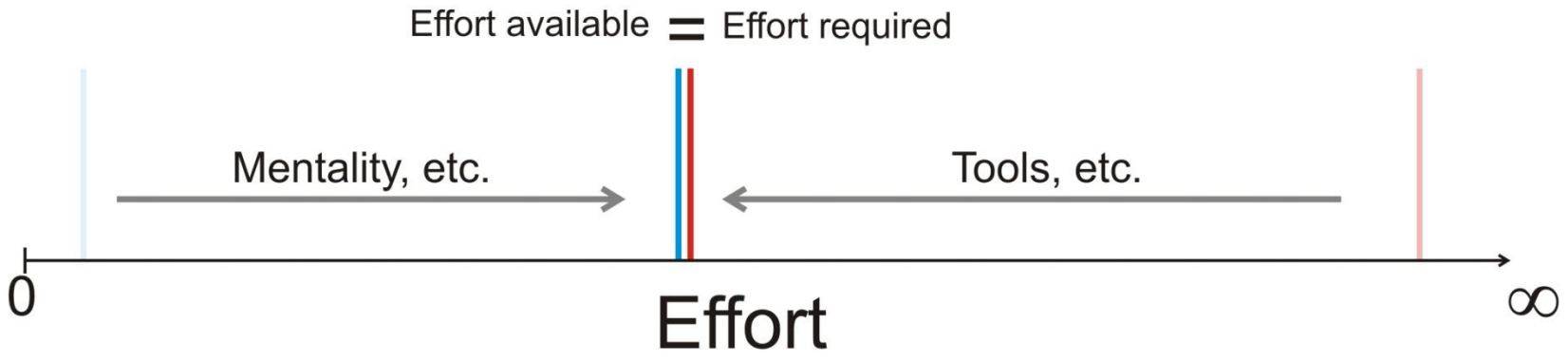
Research landscape

- $1.4 trillion globally (2012)
- Ca. 18% on basic research
  - Almost all basic research done in academia
- Majority of publications come from academia
  - What % of R&D results are open?
- Ca. 7 M researchers worldwide
  - 7.5 researchers per 1000 employed people

Premise

- Society will get more value from scientific research if it is FAIR.

d) Case 3: Approach from both sides



Changing user behavior

Mentality:
- Awareness of metadata
- Potential value of FAIR

Autocratic:
- Funding agencies
- Publishers

Incentives:
- Immediate value
- provide services

Providing Tools

Ontology:
- Build and test
- Make as domain aware as possible

e-labbook:
- Provide domain-specific plug-ins
- Forms

Parsing and auto-tagging:
- Develop NLP
- NN for auto-tagging

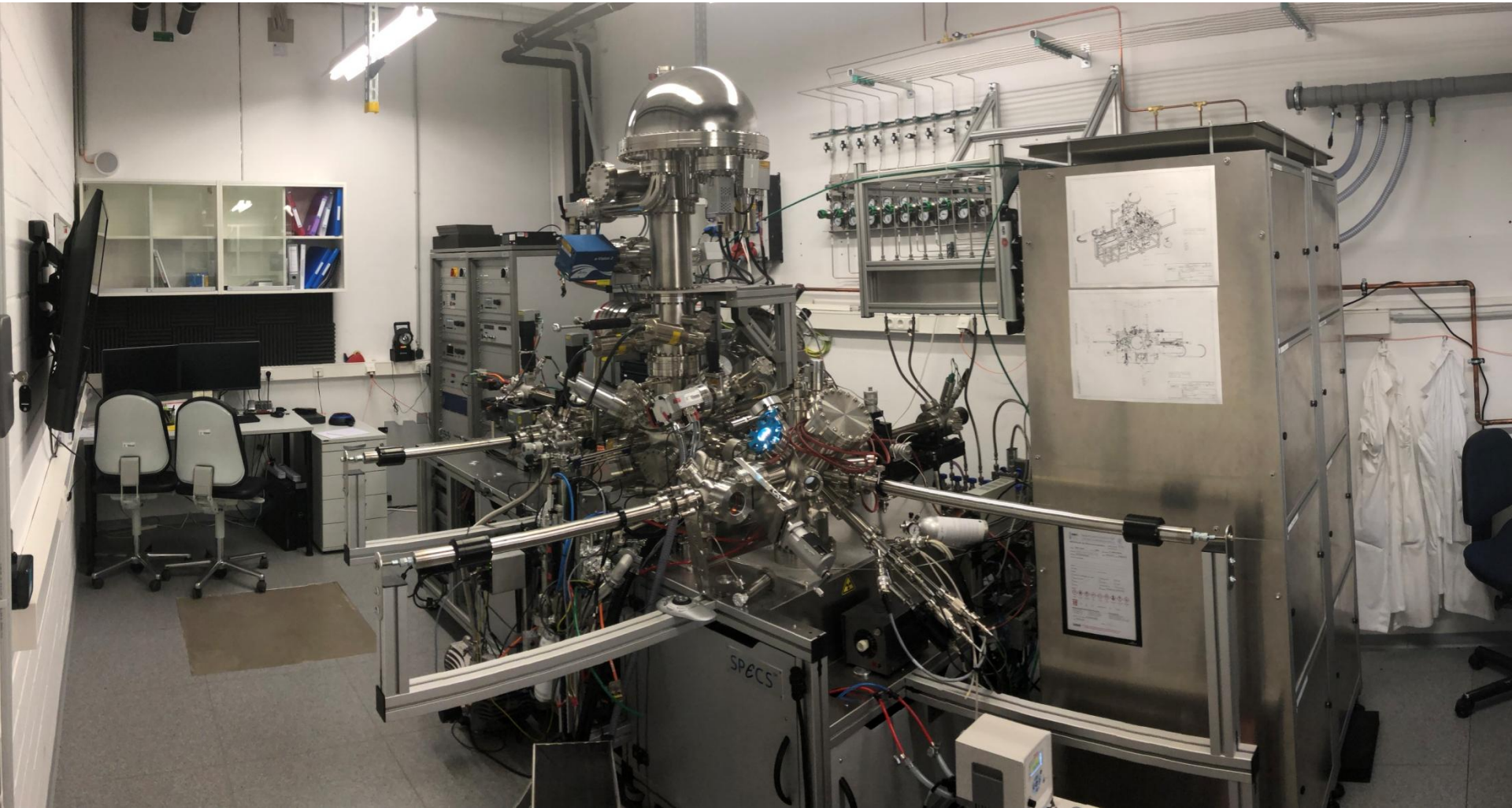
Vectorstock.com/14886332

Community research

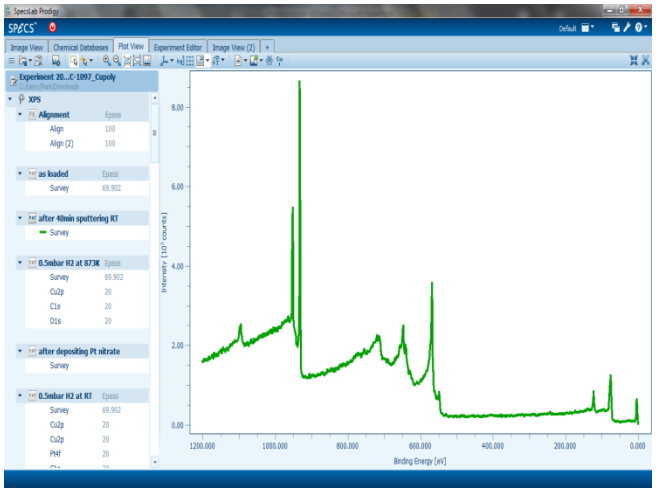
Community research

- To build any solution that I want the community to adopt, I need to make sure it has value to the community.

Our Lab

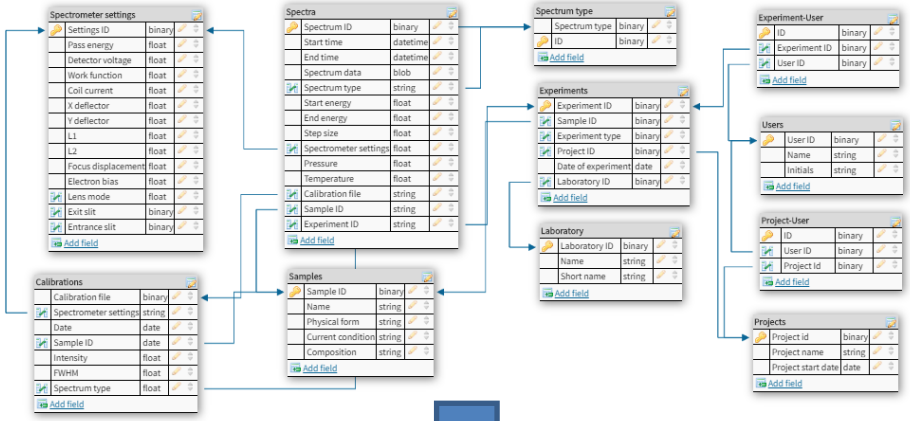


Raw spectra

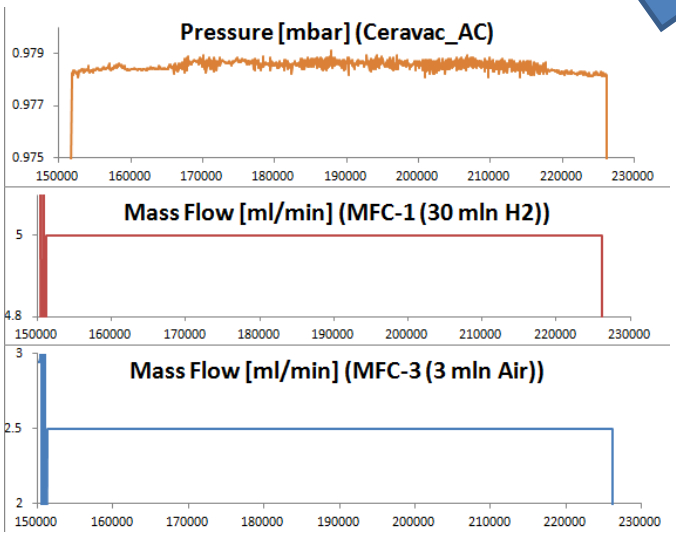


SQL database

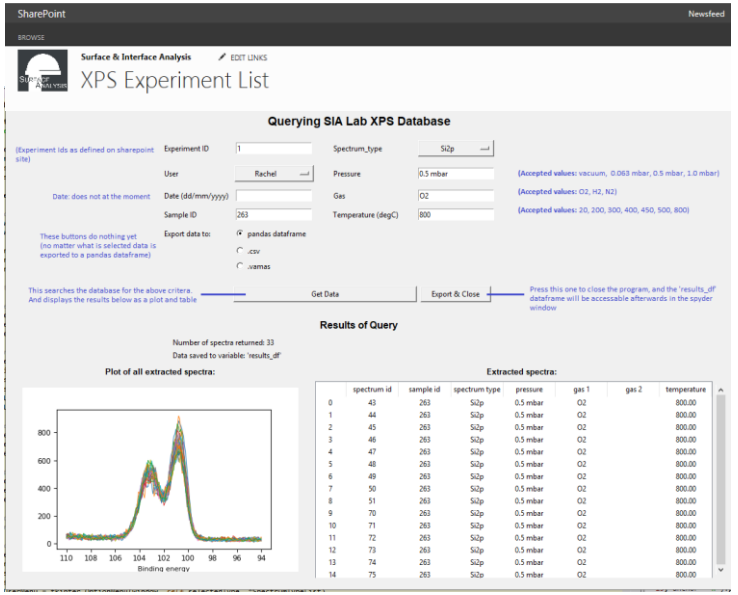
Link to samples, spectra, projects, users



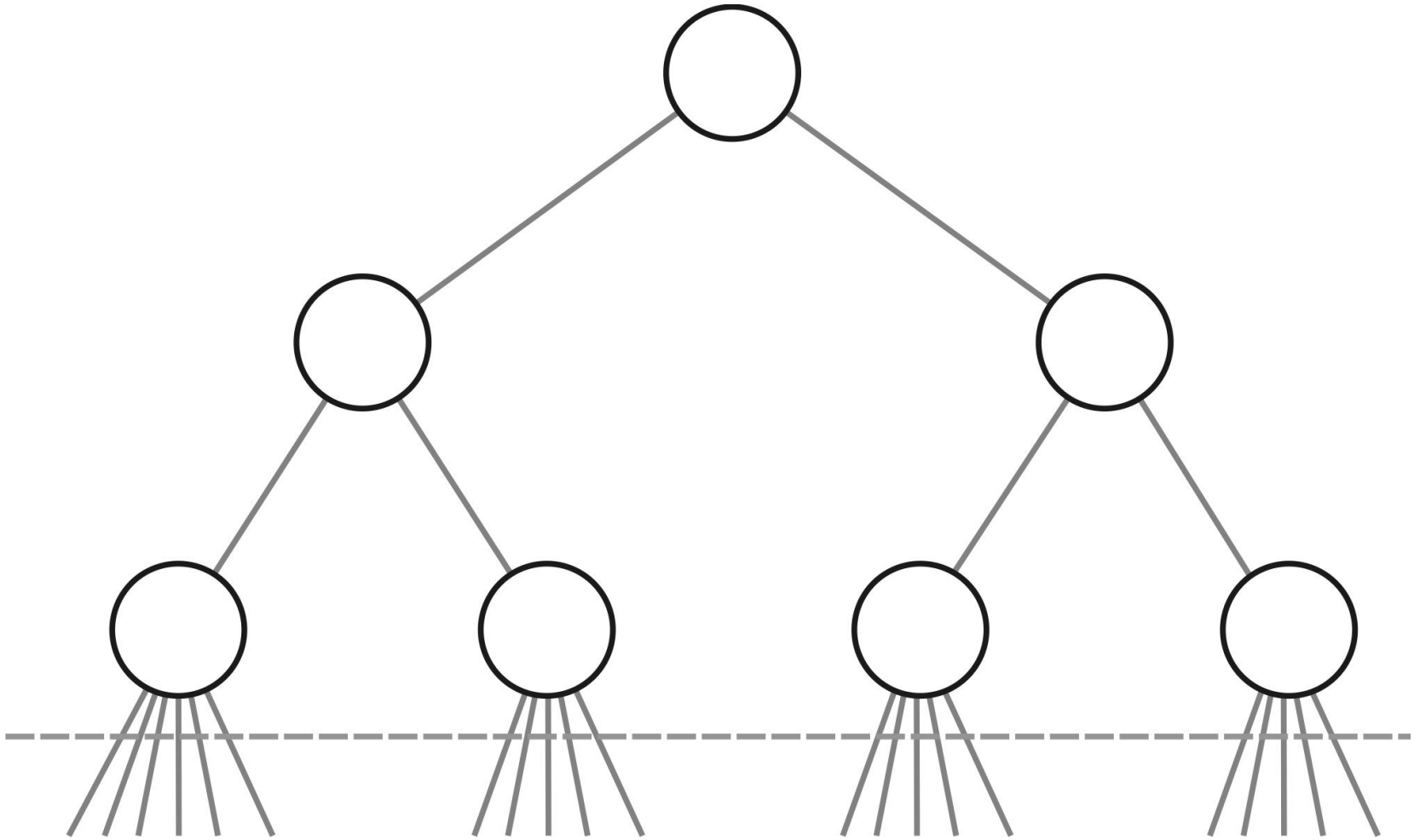
Experimental metadata



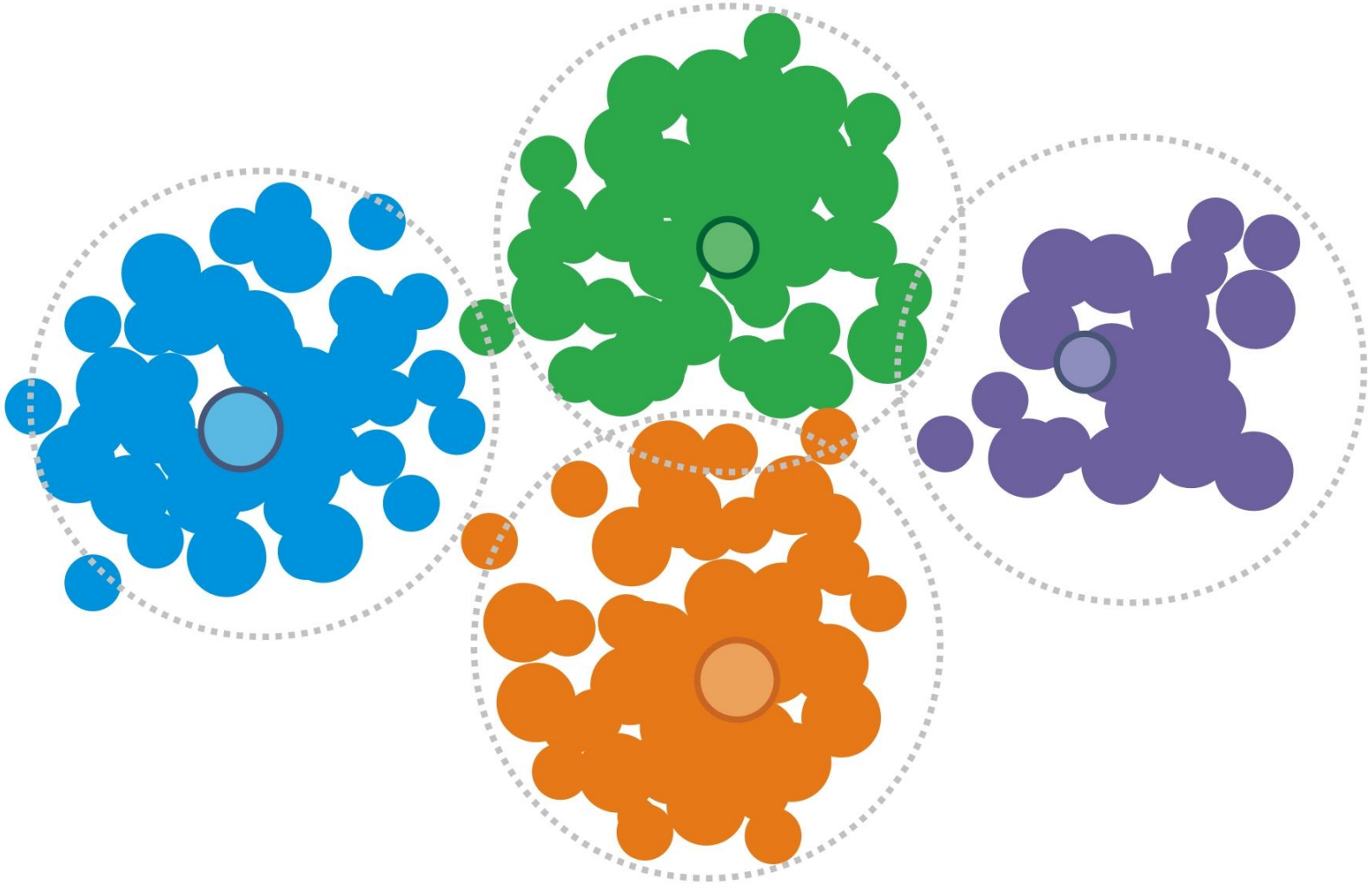
Sharepoint Web user interface



Abstract to domain specific ontologies



<1,1,2,0,0,0,0,0,5,0,1,…>



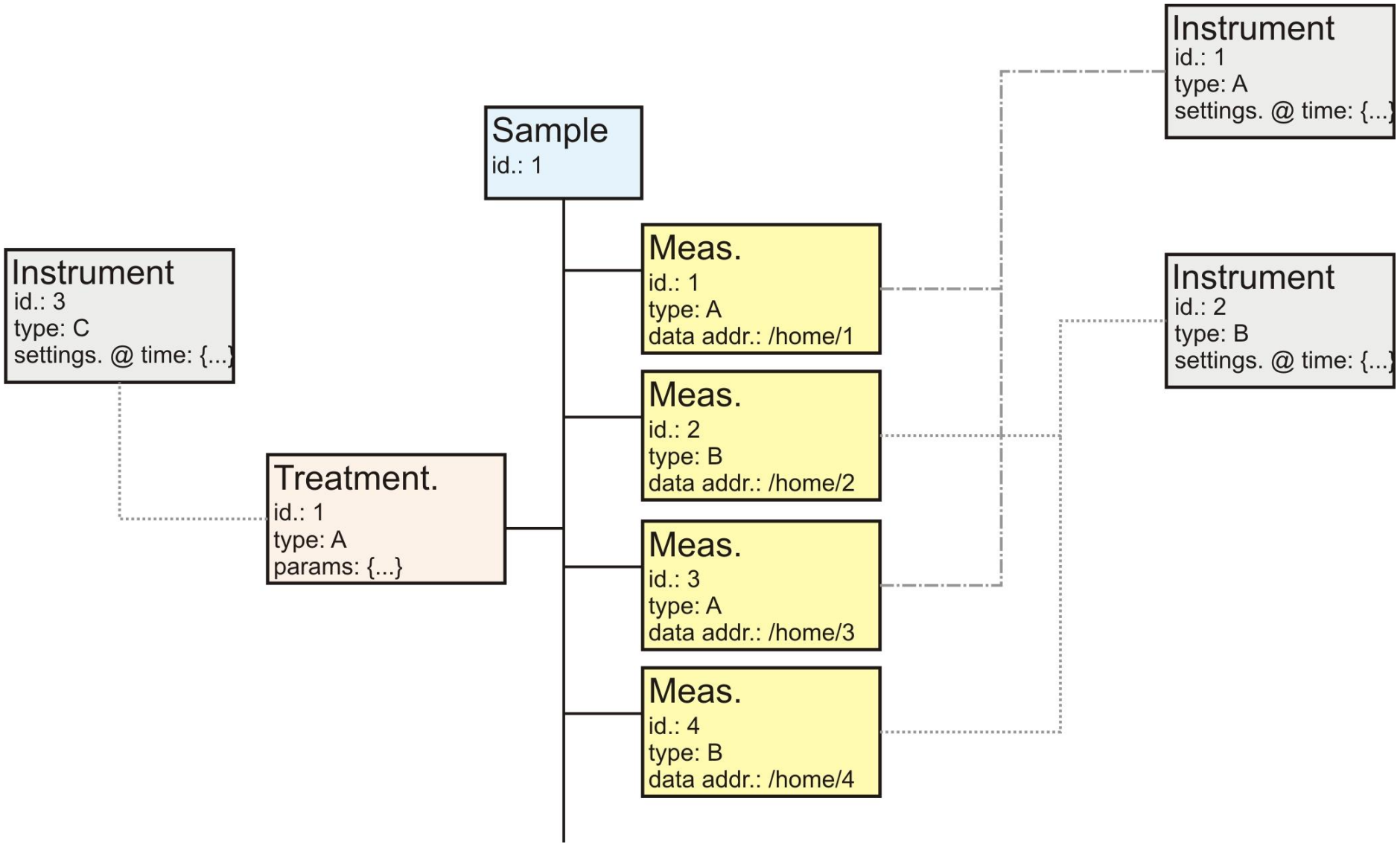
Build a hub to gather info

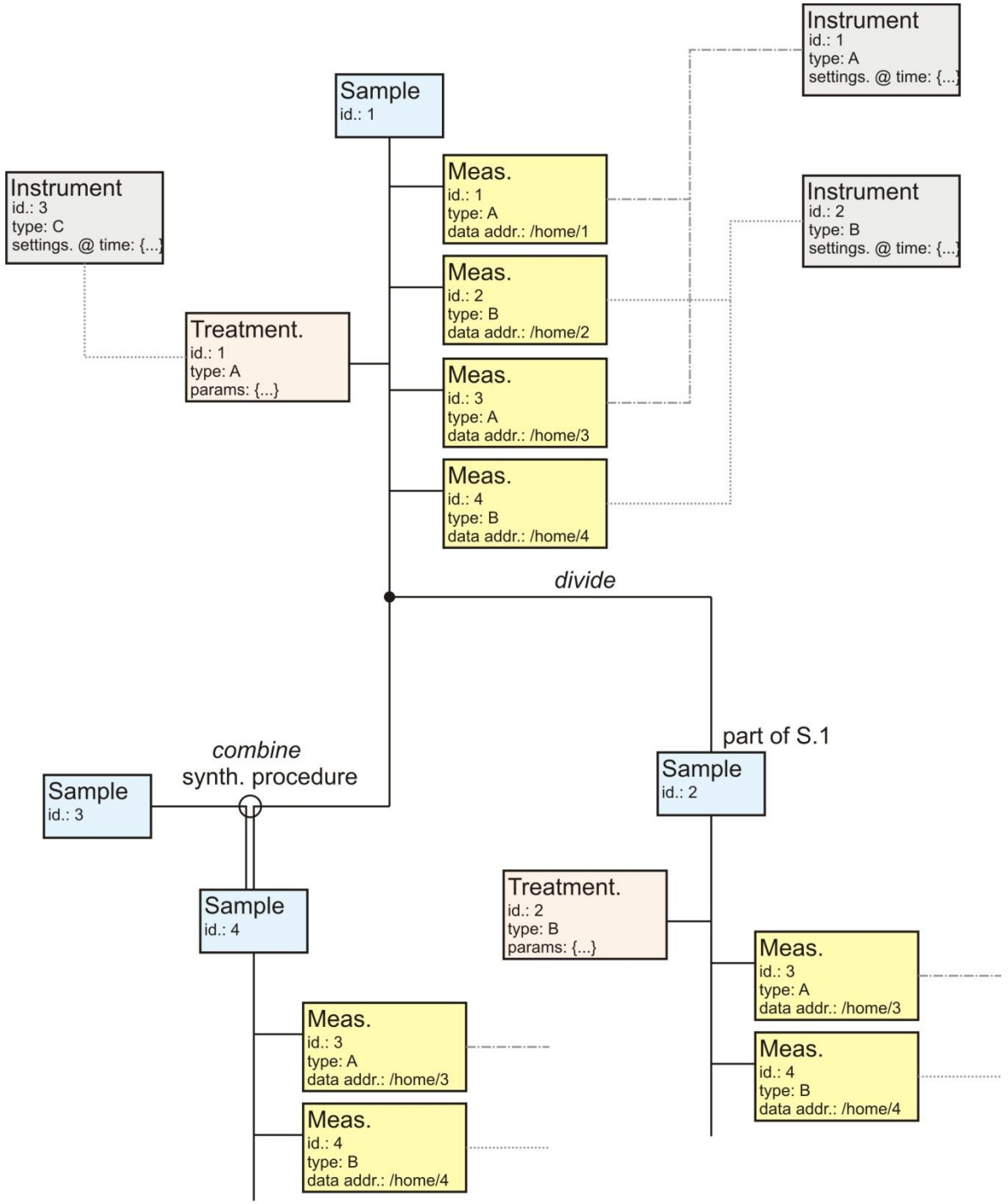
- Network with community
- Provide a place where they can enter their info
- Provide a hub where they can submit data and have immediate benefit from a service
- Auto-ID of XPS spectra

Pilot project XPS

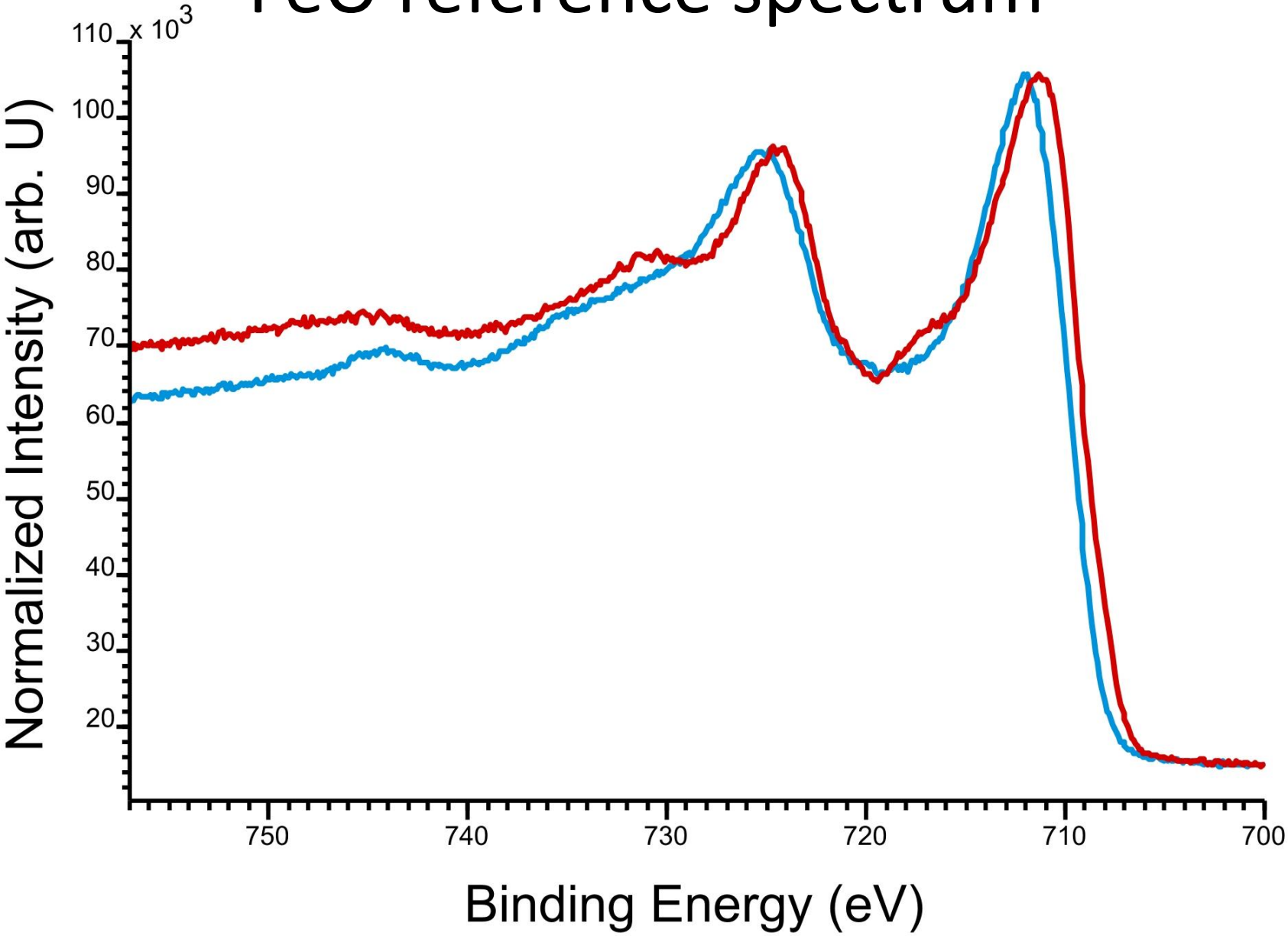
- XPS is a common method
- Train neural network on XPS data for auto-peak ID
- Step 1: Get the data
- Step 2: Label the data
- Step 3: Curate
- Step 4: train models
- Step 5: Auto ID

Sample metadata


Instrument
id.: 1
type: A
settings. @ time: {...}
Sample
id.: 1
Instrument
id.: 3
type: C
settings. @ time: {...}
Meas.
id.: 1
type: A
data addr.: /home/1
Instrument
id.: 2
type: B
settings. @ time: {...}
Meas.
id.: 2
type: B
data addr.: /home/2
Treatment.
id.: 1
type: A
params: {...}
Meas.
id.: 3
type: A
data addr.: /home/3
Meas.
id.: 4
type: B
data addr.: /home/4


Instrument
id.: 1
type: A
settings. @ time: {...}
Sample
id.: 1
Meas.
id.: 1
type: A
data addr.: /home/1
Instrument
id.: 3
type: C
settings. @ time: {...}
Instrument
id.: 2
type: B
settings. @ time: {...}
Meas.
id.: 2
type: B
data addr.: /home/2
Treatment.
id.: 1
type: A
params: {...}
Meas.
id.: 3
type: A
data addr.: /home/3
Meas.
id.: 4
type: B
data addr.: /home/4
divide
part of S.1
combine
synth. procedure
Sample
id.: 3
Sample
id.: 2
Sample
id.: 4
Treatment.
id.: 2
type: B
params: {...}
Meas.
id.: 3
type: A
data addr.: /home/3
Meas.
id.: 4
type: B
data addr.: /home/4
Meas.
id.: 3
type: A
data addr.: /home/3
Meas.
id.: 4
type: B
data addr.: /home/4

FeO reference spectrum



How can I support my claim?

- XRD said FeO, ICP said FeO, Sigma Aldrich said 99.99% FeO
- But … XPS said Fe_2O_3
- XRD measures volume of ca. 50-200 µm x 100 $µm^2$
- ICP measures whole sample (actually sibling of XPS sample)
- XPS measures volume of 5 nm x 300 $µm^2$
- Data represent different portions of sample

Open question…

- How can we encode this uncertainty?
- Do we embed it into the schema as assertions based on measured data, with probabilities associated with them?

Thank You!

- ISO 19156
- Cox

- RDA Oct 23 conference

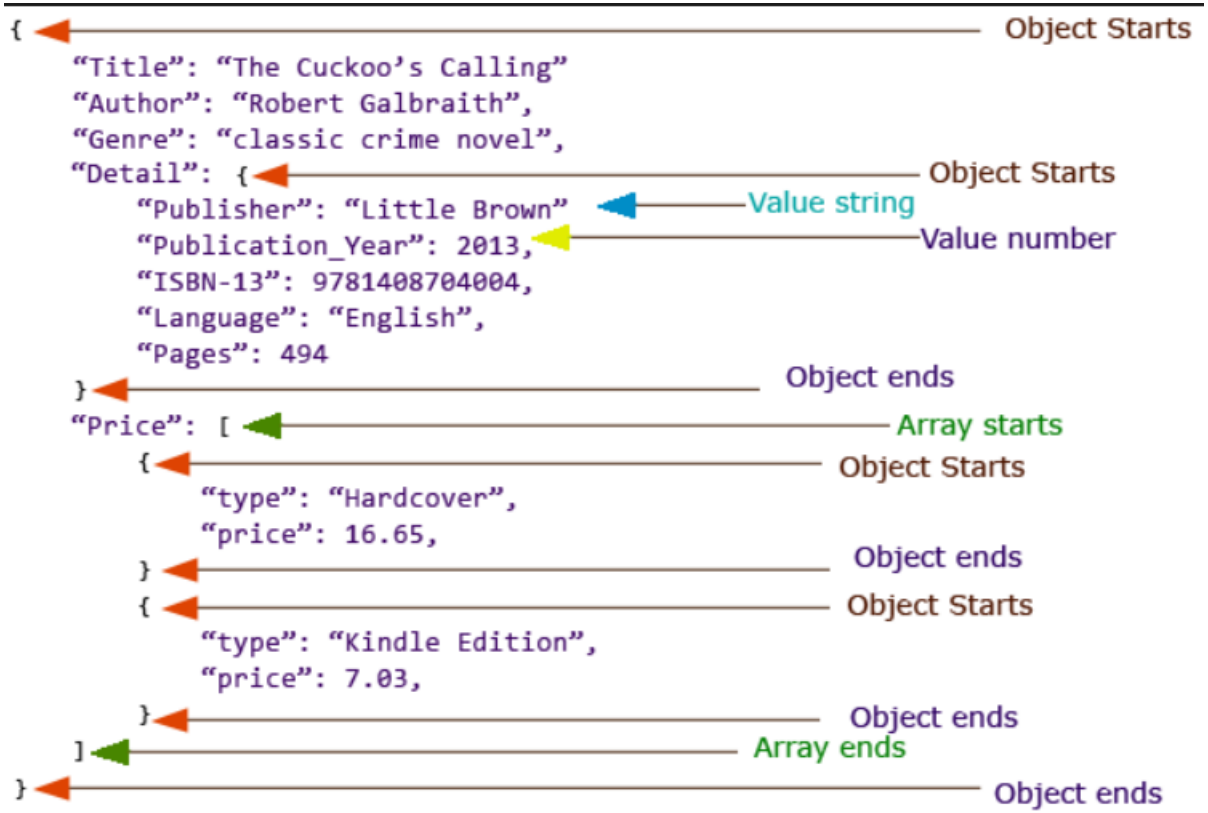
Pilot projects

- Pilot project to use data science methods on community provided, large and diverse data sets

Isolating one method is not enough

- The knowledge comes from the correlations between linked methods
- Correlations between multiple properties
- Correlations between structure and function
- This requires that data from diverse methods needs to be connected in a machine interpretable way

JSON


{
Object Starts
"Title": "The Cuckoo's Calling"
"Author": "Robert Galbraith",
"Genre": "classic crime novel",
"Detail": {
Object Starts
"Publisher": "Little Brown"
Value string
"Publication_Year": 2013,
Value number
"ISBN-13": 9781408704004,
"Language": "English",
"Pages": 494
}
Object ends
"Price": [
Array starts
{
Object Starts
"type": "Hardcover",
"price": 16.65,
}
Object ends
{
Object Starts
"type": "Kindle Edition",
"price": 7.03,
}
Object ends
]
Array ends
}
Object ends

XML

```
<Books>
    <Book ISBN="0553212419">
        <title>Sherlock Holmes: Complete Novels...
        <author>Sir Arthur Conan Doyle</author>
    </Book>
    <Book ISBN="0743273567">
        <title>The Great Gatsby</title>
        <author>F. Scott Fitzgerald</author>
    </Book>
    <Book ISBN="0684826976">
        <title>Undaunted Courage</title>
        <author>Stephen E. Ambrose</author>
    </Book>
    <Book ISBN="0743203178">
        <title>Nothing Like It In the World</title>
        <author>Stephen E. Ambrose</author>
    </Book>
</Books>
```

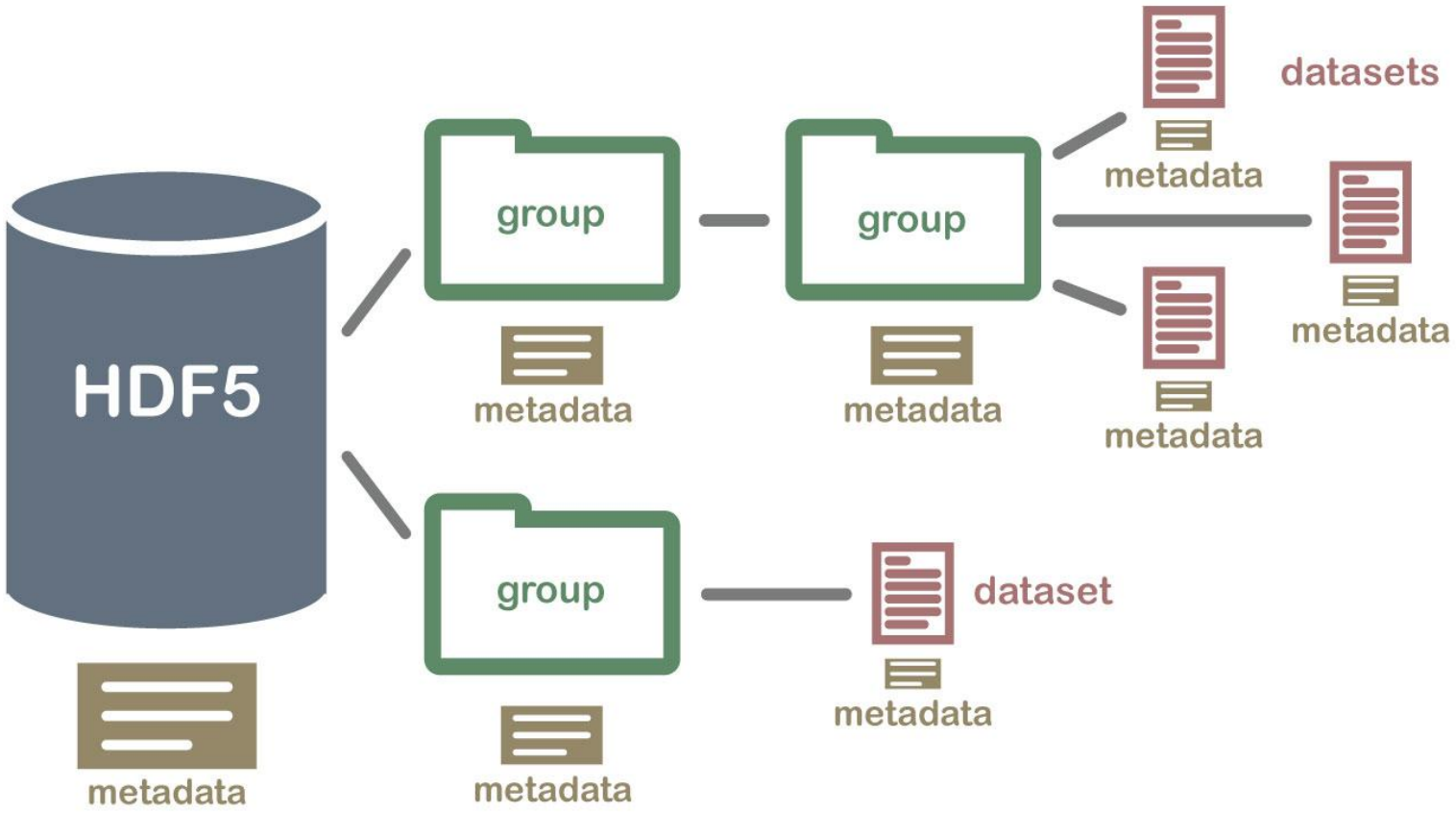
VAMAS

```
VAMAS Surface Chemical Analysis Standard Data Transfer Format 1988 May 4
Specs
Phoibos
SPECS
?
42
Casa Info Follows CasaXPS Version 2.3.19rev1.2v
0
Created by SpecsLab Prodigy, Version 4.47.1-r76131
SourceAnalyserAngle: Not Specified
CasaRowLabel:survey i.Vac
CasaRowLabel:after 65 min sputter
CasaRowLabel:vacuum

CasaRowLabel:5 mbar He
CasaRowLabel:5 mbar He
CasaRowLabel:5 mbar He

CasaRowLabel:vacuum
CasaRowLabel:5 mbar He
NORM
REGULAR
8
1
Loop
d
0
0
0
0
277
Survey
survey i.Vac
2019
3
15
16
55
24
0
12
Casa Info Follows
0
0
0
0
File = E:/OwnCloud/Experiment 2019-03-15_Cu in He.sle
Group = survey i.Vac
First of group = 1
Analyzer lens = AngleResolvedMode22:1.5kV
Analyzer slit = 4:7x20c\C:mesh
Scan mode = FixedAnalyzerTransmission
Kinetic energy start = 284.61
XPS
0
Al
1486.61
0
0
```

HDF5


HDF5
metadata
group
metadata
group
metadata
datasets
metadata
metadata
metadata
group
metadata
dataset
metadata

Publishing is old-school

- Disseminating information via embedded graphics in prose does not lend itself well to large scale data analysis
- The true facts as disguised in a lot of fluff, and the data is not in a consistent format.
- We need a semantic web for scientific data to improve the value of scientific efforts

Show picture of cowboy holding up nature journal

Incentives to publish

- "By providing intellectual credit publicly for innovative claims in natural philosophy, the journal encouraged scientists to disclose knowledge that they might otherwise have kept secret."

- "The *Philosophical Transactions of the Royal Society* created a sense of competition among scientists to be the first to publish a new scientific finding"

https://www.nap.edu/read/10613/chapter/4

"The purposes and responsibilities for sharing"

The user-entered metadata will follow a regimented logical decision tree to determine the information that must be entered. When the experimenter initiates a measurement, she/he will be prompted to enter the needed information by means of a chatbot that follows the decision tree. The resulting metadata will be automatically entered into the relational database and associated with the measured data until the experiment is completed.
The measured data will also be directly committed to the database and linked to the metadata. An experiment (i.e. all experimental instructions) will be programmed into the instrument and the measurements will be initiated. The measurements will be automated and capable of running without constant human intervention. All instrumental parameters will be logged during the measurements and linked to the associated raw data.

1. Format of data logging for one example experiment

Indexing data	Raw Data		Logged data								User inputted data				
Data entry ID	XPS Spectrum	QMS Spectrum	Date/Time	Pressure	Sample temperature	Mass flow1	Gas ID 1	Mass flow2	Gas ID 2	Spectrometer Settings	Sample ID	Project ID	XPS line	Experiment type	Pretreatments
Integer	*2 x n matrix*	*2 x m matrix*	*String*	*Float*	*Float*	*Float*	*String*	*Float*	*String*	*Dictionary*	*String*	*String*	*String*	*String*	*String List*
5123	1022, 1001, ... 491.0, 491.1, ...	2.3E-12, 2.21E-12, ... 0.0, 0.1, ...	2018.11.17 13:20:55	0.515	670.5	5.0	Oxygen	2.0	Methanol	{Pass Energy: 20, Detector: 2100,...}	CEC-1021	P112	Co2p	Temp. prog. oxidation	Polish, sputter, anneal

2. Selection from newly measured data

	XPS Spectrum	XPS Line
Integer	*2 x n matrix*	*String*
5123	1022, 1001, ... 491.0, 491.1, ...	Co2p
5124	1150, 1125, ... 491.0, 491.1, ...	Co2p
5125	1095, 1109, ... 491.0, 491.1, ...	Co2p

Compare with reference data

3. XPS Reference Spectrum Database

Data ID	XPS Spectrum	XPS Line	Spectrometer Settings	Compound ID	Compound ID No.	Prep. Conditions	Experiment no.
Integer	*2 x n matrix*	*String*	*Dictionary*	*String*	*Integer*	*Dictionary*	*Integer*
1152	1022, 1001, ... 491.0, 491.1, ...	Co2p	{Pass Energy: 20, Detector: 2100,...}	Co3O4	11524		
1153	1022, 1001, ... 491.0, 491.1, ...	Co2p	{Pass Energy: 20, Detector: 2100,...}	CoO	11524		
1154	1022, 1001, ... 491.0, 491.1, ...	Co2p	{Pass Energy: 20, Detector: 2100,...}	Co	11524		

Data Processing
e.g. subtraction, averaging, etc.

Name or data processing data
e.g. name of Casa file

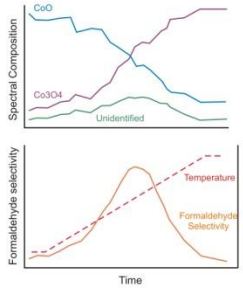
PCA, k-means clustering, neural network, regression

Identification of compounds present

4. Processed Data from one experiment

Compound 1	Compound 2	Compound 3	Goodness of fit Using identified Compounds
String	*String*	*String*	*Float*
CoO	Co3O4	Unidentified	0.85

5. Trend plotting



The report processing will quickly show trends from an experiment and will be used to identify unexplainable anomalies, such as spectral components that do not belong to the database of known compounds

Look-up indirect materials properties

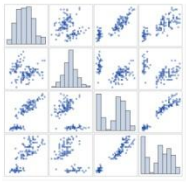
The XPS spectra will act as a material fingerprint, and will be used to link measured data with materials properties that were not directly measured in the experiment.
This analysis will enable heterogeneous data sets to be compared. For instance, using oxidation state trends, one could compare activity trends across diverse elements and diverse compounds.

6. Materials Property Database

Data ID	Compound ID No.	Compound ID	Material Type	Cation ox. state	Elec. Bandgap	Eg ref.	Group electroneg.	E-neg. ref.	Lewis acidity	E-neg. ref.	Crystal Struct.
Integer	*Integer*	*String*	*String*	*Integer List*	*Float*	*String*	*Float*	*String*	*Float*	*String*	*String*
1152	11524	Co3O4	Oxide	2,3	1.6	DOI:10.1021...	8.4	DOI:11.5023...	0.35, 0.5	DOI:05.9001...	cubic, spinel, Fd-3m, 227
1153	11524	CoO	Oxide	2	2.4	DOI:12.2521...	7.5	DOI:22.9525...	0.35	DOI:42.0014...	cubic, Fm3m
1154	11524	Co	Metal	0	0.0	DOI:11.1121...	4.8	DOI:09.5274...	N/A	DOI:14.7040...	Co

Big cross-material trends

Scatter plot matrix



Descriptor search

List of potential descriptors

Descriptor	Amt. of variance described
D1	0.73
D2	0.42
D3	0.22
D4	0.18
.	.
.	.
.	.

Select sensical descriptors

Cation hardness/Group electronegativity predicts formaldehyde selectivity with success rate of 0.73

Verification of descriptors using DFT simulations

Understanding of descriptors

Hypothesis for design of better material

Synthesize and test material. Is it better or not?

Researchers have been finding descriptors since the dawn of chemistry research. E.g. electronegativity, Lewis acitidy, hardness, et.c. These have been very useful for material property predictions and have also been verified and justified by ab initio calculations (when possible).
We will continue to do the same, but faster and using much larger and more complex sets of data.
In this way, we will be able to see trends that were not previously visible.

Science is crowd sourced

- Little bit here, little bit there
- Parameter space is enormous and large trends can generally not be achieved via a single lab
- To find patterns we need to utilize the data of the crowd. We need to connect labs doing related work. They need to leverage each others work to make progress faster
- They need to be connected

What do I do when I think I have a great idea?

- I Google it to see if anyone else has tried it.
- If so, did they do it the same way I would have?
- Can I make it better?
- Can I take their starting pint and go further?
- Is it a ‘destined to fail’ idea?
- These are good things to know before getting started implementing an idea.

Now we know it can be done better

- Now we know that the self-regulating process can be done much faster and better

- Now we know algorithms + computers + lots of data can be used to find big patterns’