

# The Story of an Experiment: A Provenance-based Semantic Approach towards Research Reproducibility

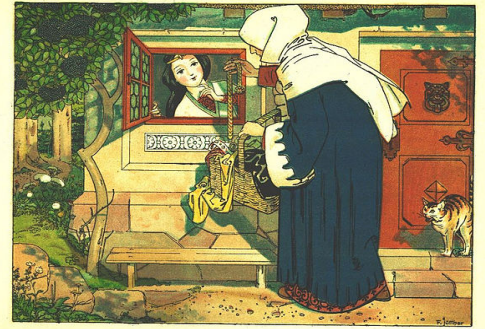
**Sheeba Samuel**, Kathrin Groeneveld, Frank Taubert , Daniel Walther,  
Tom Kache, Teresa Langenstück , Birgitta König-Ries, H. Martin Bückner,  
and Christoph Biskup

Friedrich-Schiller University, Jena, Germany  
Jena University Hospital, Germany

SWAT4HCLS, 4<sup>th</sup> December 2018

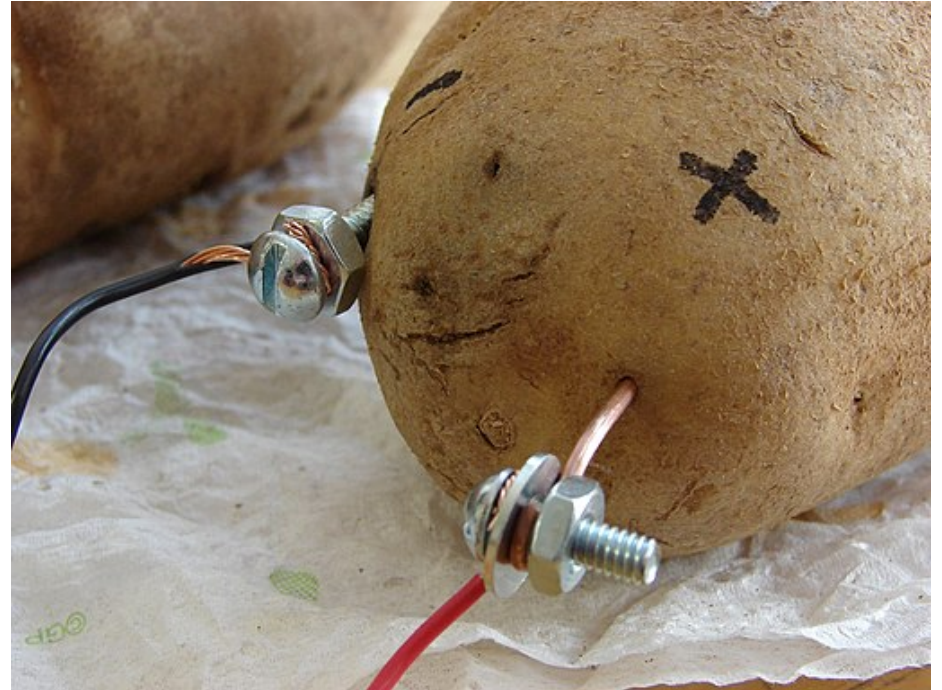


# Story

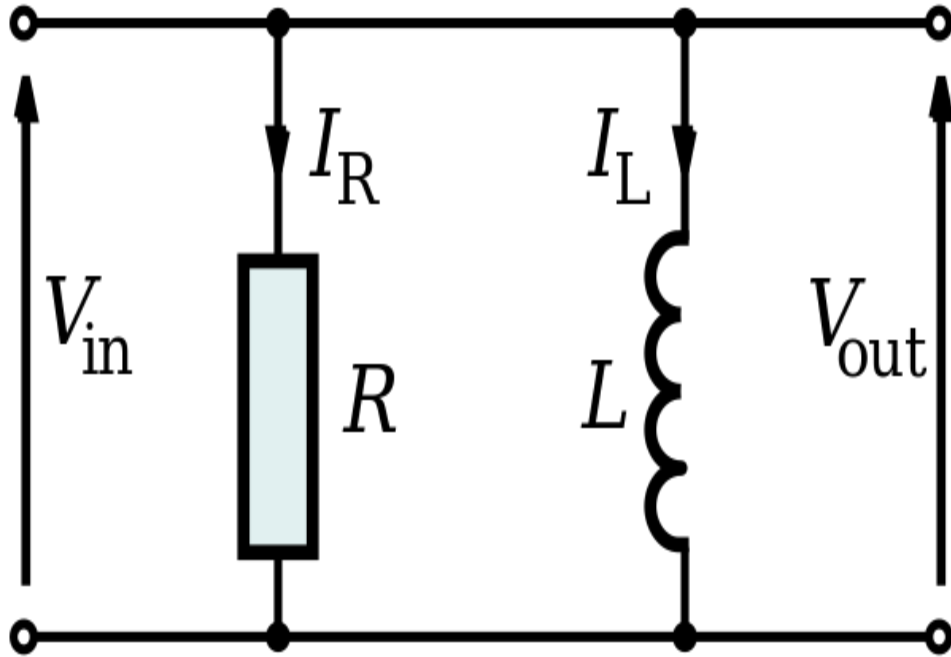




# Story of an Experiment

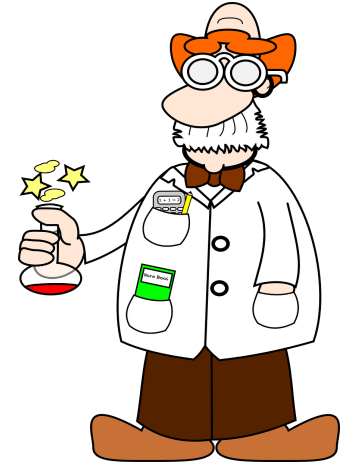
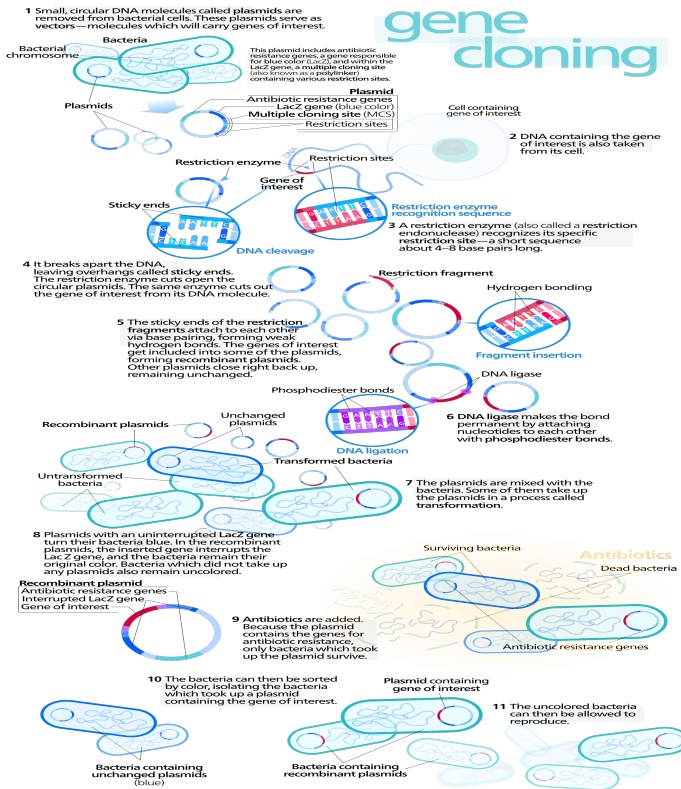


# Story of an Experiment

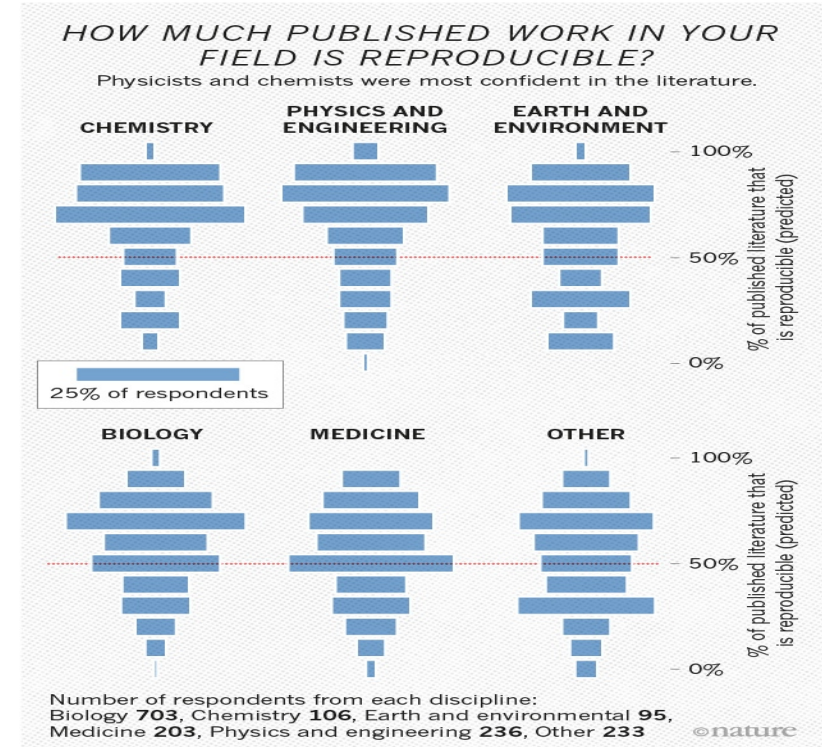
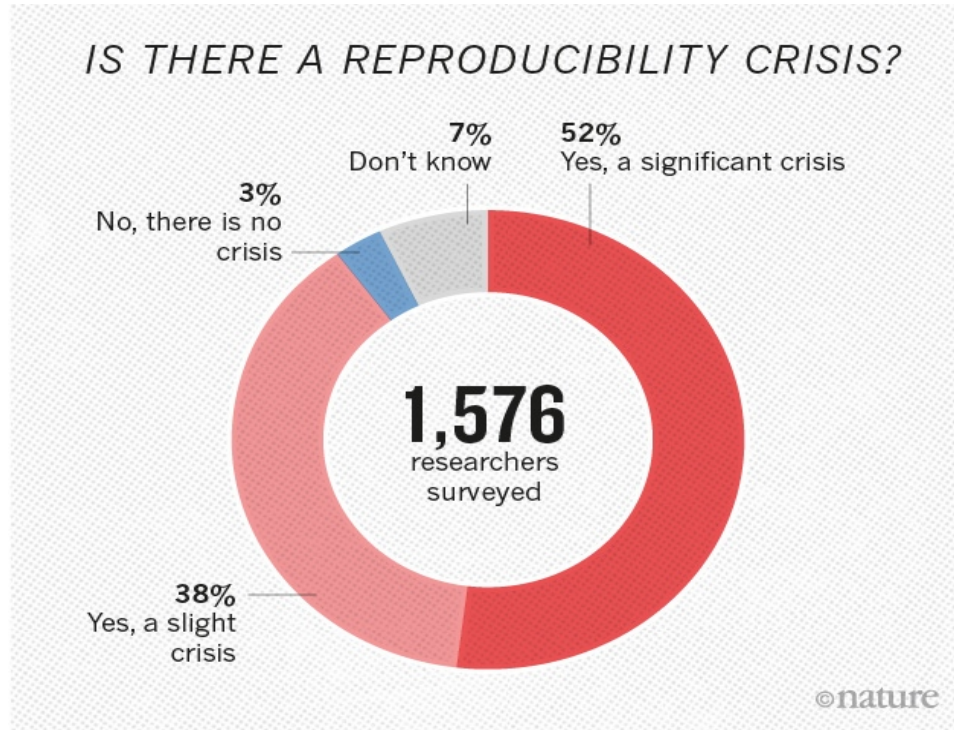




# Story of an Experiment



# Reproducibility



# Challenges that hinder research reproducibility

- Lack of documentation in digital media
- Non-availability of the datasets, code, workflow...
- Integration of data generated from different devices
- Incomplete and uncertain provenance information
- Lack of knowledge of the type of data and their formats and most importantly their semantics.



# Contributions

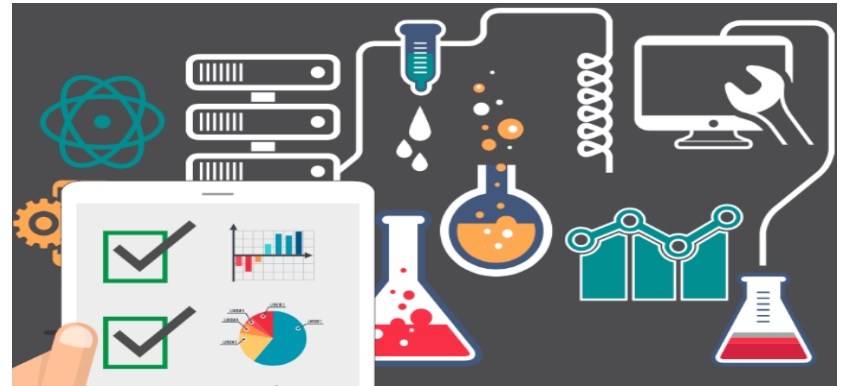
- Identifying the components and competency questions
- Capture the provenance data from multiple resources of an experiment.
- Presenting our provenance-based semantic approach using REPRODUCE-ME ontology by extending PROV-O and P-Plan
- Visualization of the provenance data of an experiment as a dashboard to the scientists in our prototype, CAESAR.

# Experiments

Interviews with the scientists in the CRC ReceptorLight as well as a workshop conducted to foster reproducible science.

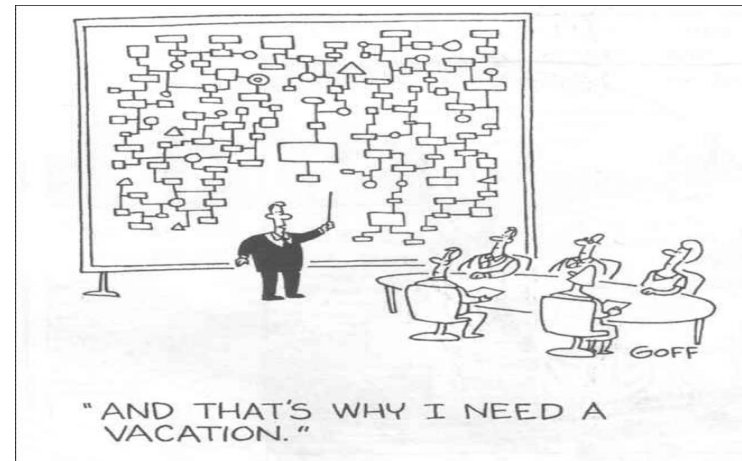


```
177     default = 1.0
178 }
179
180 global_scale_setting = bpy.props.FloatProperty(
181     name="Scale",
182     min=0.01, max=100.0,
183     default=1.0,
184 )
185
186 def execute(self, context):
187
188     # get the folder
189     folder_path = (os.path.dirname(self.filepath))
190
191     # get objects selected in the viewport
192     viewport_selection = bpy.context.selected_objects
193
194     # get export objects
195     obj_export_list = viewport_selection
196     if self.use_selection_setting == False:
197         obj_export_list = [i for i in bpy.context.scene.objects]
198
199     # deselect all objects
200     bpy.ops.object.select_all(action="DESELECT")
201
202     for item in obj_export_list:
203         item.select = True
204         if item.type == "MESH":
205             file_path = os.path.join(folder_path, "{}.obj".format(item.name))
206             bpy.ops.export_scene.obj(filepath=file_path, use_selection=True,
207                                   axis_forward=self.axis_forward_setting,
208                                   axis_up=self.axis_up_setting,
209                                   use_animation=self.use_animation_setting,
210                                   use_mesh_modifiers=self.use_mesh_modifiers_setting,
211                                   use_smooth=self.use_smooth_setting,
212                                   use_smooth_groups=self.use_smooth_groups_setting,
213                                   use_smooth_groups_bitflags=self.use_smooth_groups_bitflags_setting,
214                                   use_normals=self.use_normals_setting,
215                                   use_uv=self.use_uv_setting,
216                                   use_materials=self.use_materials_setting,
```



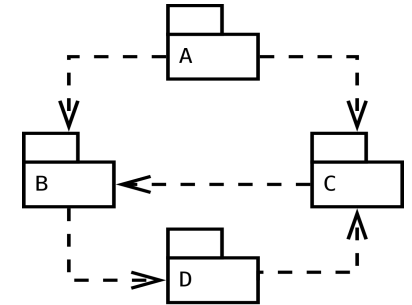
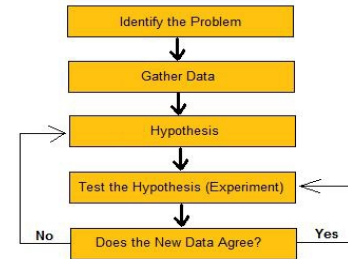
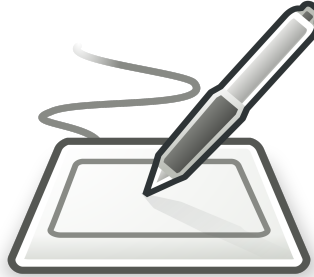
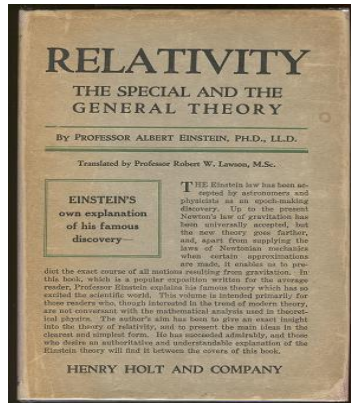
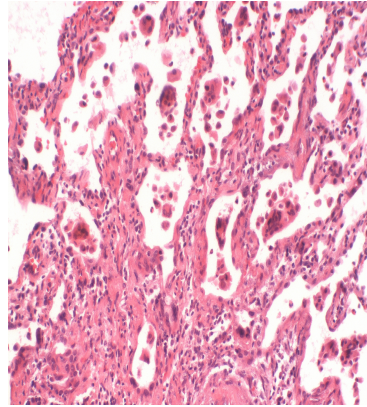
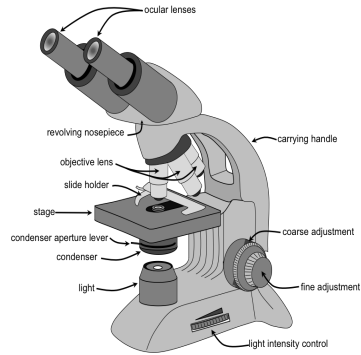
# Scientific Experiments

- Complex
- Several steps
- Several activities in the real world or cyberspace.
- Several people
- New technologies

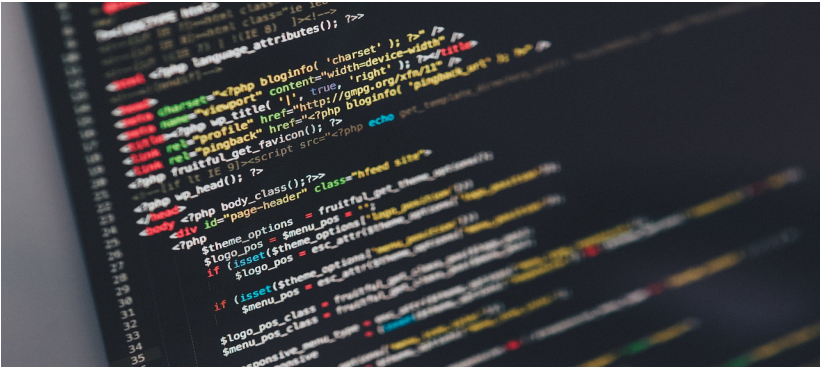
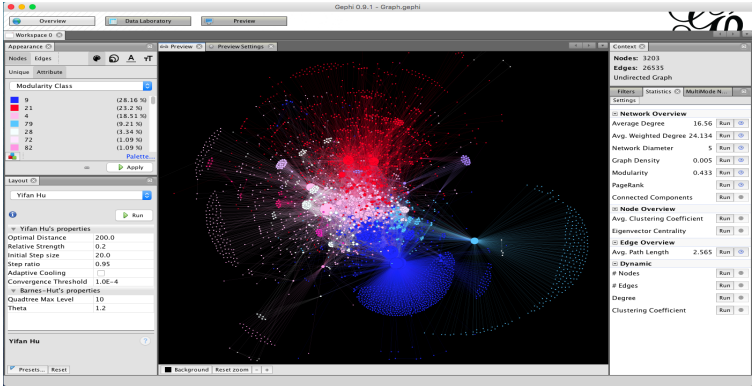




# Non-Computational Parts



# Computational Parts



## Vitamin C

From Wikipedia, the free encyclopedia

(Difference between revisions)

Revision as of 17:48, 23 February 2007 (edit) (undo)  
Lumen3 (Talk | contribs)

([\[\[Plant sources - Remove special mention of Arima, its already in the table\]\]](#))

([\[\[Older edit\]\]](#))

**Line 73:**  
=== Deficiency disease ===  
[[Scurvy]] (a form of [[avitaminosis]]) results from lack of vitamin C, which is required for correct [[biology]] synthesis in humans. Scurvy leads to the formation of liver spots on the skin, spongy gums, and bleeding from all [[mucous membrane]]. The spots are most abundant on the thighs and legs, and a person with the ailment looks pale, feels depressed, and is partially immobilized. In advanced scurvy there are open, [[suppurative]] wounds and loss of [[teeth]].  
Scurvy was at one time common among [[slave]]s, [[pirate]]s and others who were on [[ship]]s that were out to sea longer than [[fruit]]s and [[vegetable]]s could be stored and by [[sailor]]s who were similarly separated from these foods for extended periods. It was described by [[Vesputrius]] (c. 490 BC–c. 200 BC). His cause and cure has been known in many native cultures since prehistory. For example, in 1639, the French explorer [[Jacques Cartier]], exploring the [[Saint Lawrence River]], [[Lawrence River]], used the local natives' knowledge to save his men who were dying of scurvy. He boiled the needles of the [[Thuja occidentalis]] tree (Eastern White Cedar) to make a tea that was later shown to contain 60 mg of vitamin C per 100 grams.journal |p=PMID 12422875 |title=Jacques Cartier witnesses a treatment for scurvy |accessdate=2007-02-19 |date=June 2002 |author=Robert E. [[Grossman|Vesutius]], *Acta Internationalis Historiae Medicinæ* 31:101-11

No bodily organ stores vitamin C.<sup>[*factdate=February 2007*]</sup> and so the body soon depletes itself if fresh supplies are not consumed through the digestive system.

**Line 141:**  
=== Possible other vitamin C deficiencies ===  
[[Burrage Endo dysfunction Adheno Proctoderm]](Atherosclerosis) has been hypothesized to be a vitamin C deficiency disease]]

Rath (who has killed numerous people while promoting clinically unrelated drugs in townships across South Africa) made a typically strange statement: that during the [[Ere apt]], when vitamin C was scarce, [[rural selection]] favoured human individuals who could repair arteries with a layer of [[cholesterol]]. He suggests that although eventually harmful, cholesterol lining of artery walls would be beneficial if that it would keep the individual alive until access to vitamin C allowed arterial damage to be repaired. If this is true, [[atherosclerosis]] is in fact a vitamin C deficiency disease.  
The established RDA has been criticised by Pauling to be one that will prevent [[acute (medical adjective) Scurvy]], and is not necessarily the dosage for optimal health.<sup>[*factdate=February 2007*]</sup>

Current revision (04:17, 24 February 2007) (edit) (undo)  
Jockey (Talk | contribs)

([\[\[remove vandalism and cleanup scurvy section\]\]](#))

**Line 73:**  
=== Deficiency disease ===  
[[Scurvy]] (a form of [[avitaminosis]]) results from lack of vitamin C, as an effect of its requirement for correct [[biology]] synthesis. Scurvy leads to the formation of liver spots on the skin, spongy gums, and bleeding from all [[mucous membrane]]. The spots are most abundant on the thighs and legs, and a person with the ailment looks pale, feels depressed, and is partially immobilized. In advanced scurvy there are open, [[suppurative]] wounds and loss of [[teeth]], and eventually, death.

Historically, scurvy was common among those with poor access to fresh fruit and vegetables, such as [[sailor]]s, [[slave]]s and others who were on [[ship]]s that were out to sea longer than [[fruit]]s and [[vegetable]]s could be stored, as well as isolated [[sailor]]. The earliest documented case was described by [[Vesputrius]] around the year 490 BC.

The first attempt to give scientific basis for the cause of scurvy was by a ship's surgeon in the British [[Royal Navy]], [[James Lind]]. While at sea in May 1747, Lind provided some crew members with two oranges and one lemon per day, in addition to normal rations, while others continued on [[older]]. [[Vesputrius]] (sailor or [[physician]]), along with his normal rations. The results conclusively showed that something in the citrus fruits prevented the disease, a property later described as "antiscorbutic". Vitamin C was isolated in the 1920s by [[Albert Szent-Gyorgyi]] and was shown to be ascorbic acid.

No bodily organ stores vitamin C.<sup>[*factdate=February 2007*]</sup> and so the body soon depletes itself if fresh supplies are not consumed through the digestive system.

**Line 141:**  
=== Possible other vitamin C deficiencies ===  
[[Burrage Endo dysfunction Adheno Proctoderm]](Atherosclerosis) has been hypothesized to be a vitamin C deficiency disease]]

Rath hypothesized that during the [[Ere apt]], when vitamin C was scarce, [[rural selection]] favoured human individuals who could repair arteries with a layer of [[cholesterol]]. He suggests that although eventually harmful, cholesterol lining of artery walls would be beneficial in that it would keep the individual alive until access to vitamin C allowed arterial damage to be repaired. If this is true, [[atherosclerosis]] is in fact a vitamin C deficiency disease.

The established RDA has been criticised by Pauling to be one that will prevent [[acute (medical adjective) Scurvy]], and is not necessarily the dosage for optimal health.<sup>[*factdate=February 2007*]</sup>

# Competency Questions

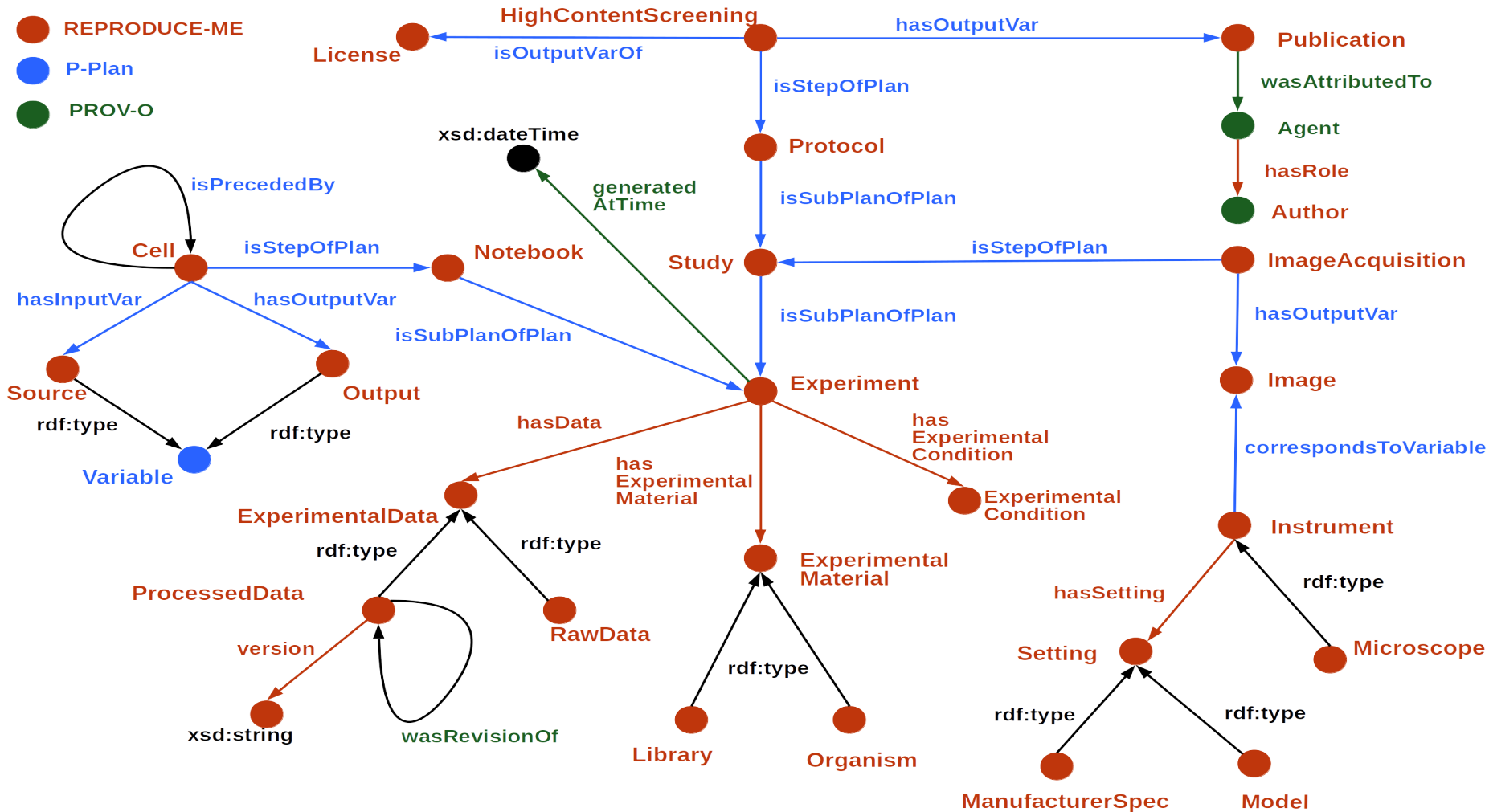
- What are the input and output variables of an experiment?
- Which are the methods and standard operating procedures used?
- Which are the files and materials that were used in a particular step?
- Which are the steps involved in an experiment which used a particular material?
- What is the complete path taken by a scientist for an experiment?
- Which are the instruments that are associated with an experiment and their settings when the output was generated?
- Which are the agents directly or indirectly responsible for an experiment?
- Who created this experiment and when? Who modified it and when?
- Which are the publications or external resources that were referenced in each step of an experiment?
- List all the experiments which use growth protocol (EFO 0003789) and studies on “Homo sapiens” and resulted in phenotype “shorter prophase” which passed the quality control.



# REPRODUCE-ME ontology

- The REPRODUCE-ME ontology extended from W3C vocabulary PROV-O and P-Plan.
- It describes a scientific experiment along with its steps, input and output variables and their relationship with each other.
- The ontology is here: <https://w3id.org/reproduceme>

# Provenance-based Semantic Approach



# Semantic-based Scientific Data Management Platform: **CAESAR**

- **CollA**borative **E**nvironment for **S**cientific **A**nalysis with **R**eproducibility
- It extends the OMERO
- OMERO:
  - open-source imaging database platform
  - Supports over 140 image file formats using BIO-Formats
  - With the help of BIO-Formats, it automatically extracts the image acquisition data



# CAESAR- Features

- Scientists can document their experimental data along with their images.
- Form-based provenance capture system
- Link experiments with
  - Steps
  - Standard Operating Procedures
  - Files
  - Jupyter Notebooks
  - Experiment Materials

# CAESAR- Features

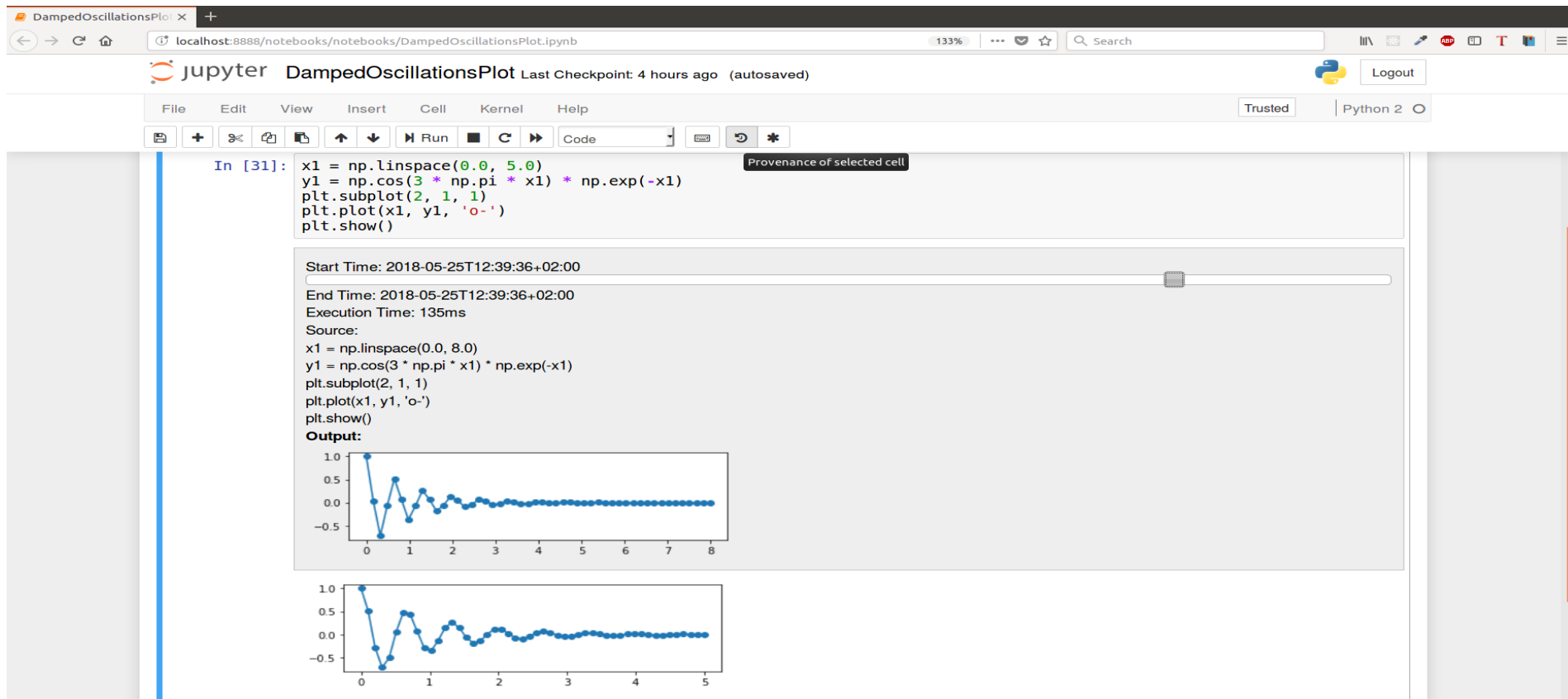
- User and Group management
- Proposal: provide suggestions on other user's experimental data.
- Version history of an experiment
- Search
- Ontology-based Data Access of OMERO database along with the experiments using REPRODUCE-ME ontology

# CAESAR- Computational Part Features

- Computational Part of an Experiment
- A distributed, collaborative and multi-user environment
- JupyterHub (<http://jupyter.org/hub>) is installed and connected to CAESAR
  - Users can create new notebooks, run and share them
- ProvBook – capture provenance of a Jupyter Notebook
  - installed in JupyterHub connected to CAESAR

# ProvBook

An extension of Jupyter Notebook, to capture and view the provenance over the course of time.



# ProvBook Difference

Difference between input and output of each execution.

## ProvBook Diff

☐ Hide unchanged cells Export diff

Base

Remote

In [16]:

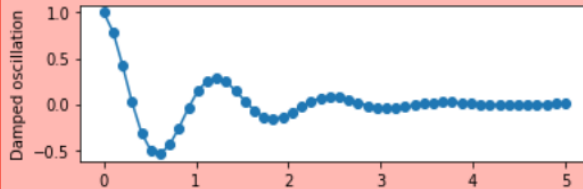
```
1 x1 = np.linspace(0.0, 5.0)
2 y1 = np.cos(18 * np.pi * x1) * np.exp(-x1)
3 plt.subplot(2, 1, 1)
(...)
```

In [16]:

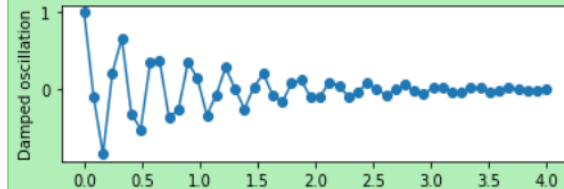
```
1 x1 = np.linspace(0.0, 4.0)
2 y1 = np.cos(18 * np.pi * x1) * np.exp(-x1)
3 plt.subplot(2, 1, 1)
(...)
```

Outputs changed

Output deleted



Output added

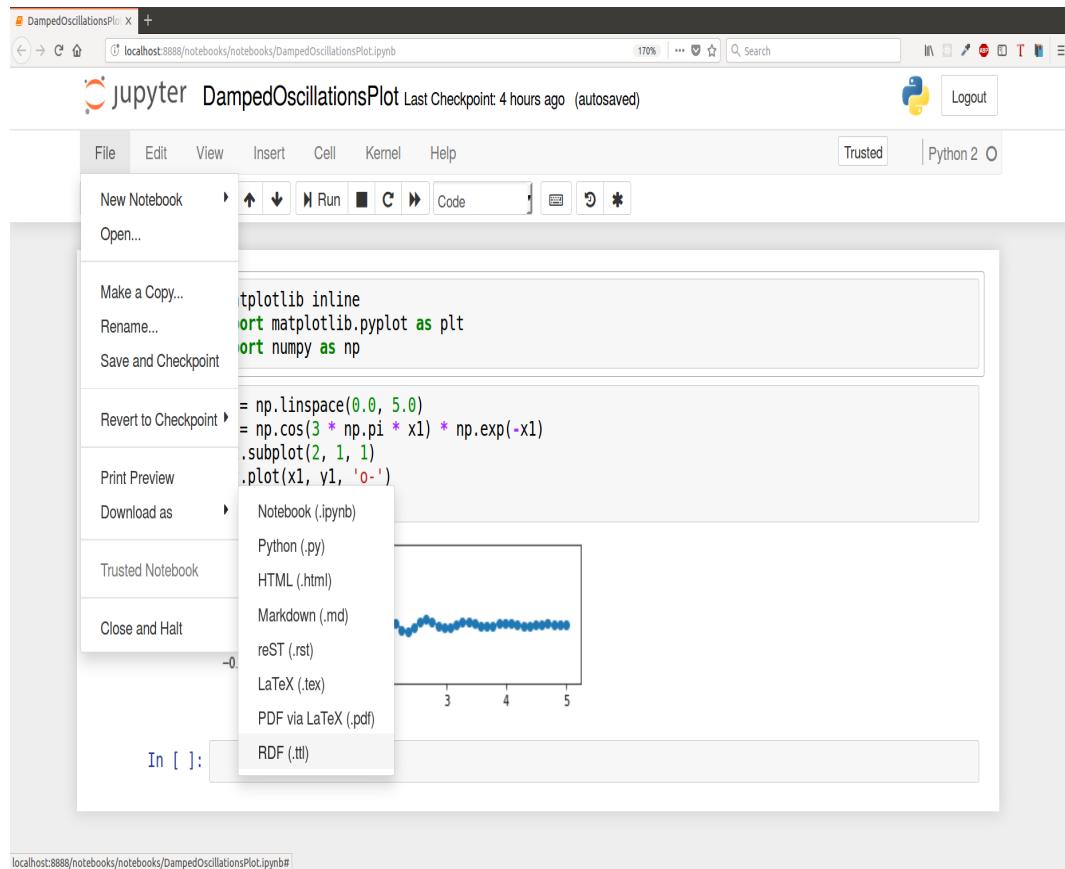


<https://w3id.org/reproduceme/research/>



# ProvBook

Convert Jupyter Notebooks to RDF and the converted RDF back to Jupyter Notebooks.



The screenshot shows a Jupyter Notebook titled "DampedOscillationsPlot" running on a local host. The file menu is open, showing options like "New Notebook", "Open...", "Make a Copy...", "Rename...", "Save and Checkpoint", "Revert to Checkpoint", "Print Preview", "Download as", "Trusted Notebook", and "Close and Halt". The "Download as" option is expanded, showing formats: "Notebook (.ipynb)", "Python (.py)", "HTML (.html)", "Markdown (.md)", "reST (.rst)", "LaTeX (.tex)", "PDF via LaTeX (.pdf)", and "RDF (.ttl)". The notebook content includes code for plotting a damped oscillation using matplotlib and numpy. A plot is visible in the background, showing a decaying sinusoidal wave.

```
@prefix p-plan: <http://purl.org/net/p-plan/#> .
@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 repr: <https://w3id.org/reproduceme#> .
@prefix xml: <http://www.w3.org/XML/1998/namespace> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

```
repr:Cell0Execution0 a repr:CellExecution ;
  p-plan:correspondsToStep repr:Cell0 ;
  prov:endedAtTime "Unknown" ;
  prov:startedAtTime "Unknown" ;
  prov:used repr:Cell0Execution0Source ;
  repr:executionTime "Unknown" .
```

```
repr:Cell0Execution1 a repr:CellExecution ;
  p-plan:correspondsToStep repr:Cell0 ;
  prov:endedAtTime "2018-09-26T16:44:07.124Z" ;
  prov:startedAtTime "2018-09-26T16:44:06.874Z" ;
  prov:used repr:Cell0Execution1Source ;
  repr:executionTime "250ms" .
```

```
repr:Cell1Execution0 a repr:CellExecution ;
  p-plan:correspondsToStep repr:Cell1 ;
  prov:endedAtTime "Unknown" ;
  prov:generated repr:Cell1Execution0Output0 ;
  prov:startedAtTime "Unknown" ;
  prov:used repr:Cell1Execution0Source ;
  repr:executionTime "Unknown" .
```

```
repr:Cell1Execution1 a repr:CellExecution ;
  p-plan:correspondsToStep repr:Cell1 ;
  prov:endedAtTime "2018-09-26T16:44:07.282Z" ;
  prov:generated repr:Cell1Execution1Output0 ;
  prov:startedAtTime "2018-09-26T16:44:07.128Z" ;
  prov:used repr:Cell1Execution1Source ;
  repr:executionTime "154ms" .
```

```
repr:Cell2Execution0 a repr:CellExecution ;
  p-plan:correspondsToStep repr:Cell2 ;
  prov:endedAtTime "Unknown" ;
  prov:startedAtTime "Unknown" ;
  prov:used repr:Cell2Execution0Source ;
```

# Visualization of Provenance Data with Dashboard

ExploreTagsShares

All members

...n modulatory CNG subunits 14

Binding via FRET 8

dose-binding A1-617-GFP

dose-response A1-617-GFP

dose-responses A1

fcGMP affinity to CNGA1 6

...finity to CNGA1-617-GFP

fcGMP efficiency

...ficity to ligand binding 33

GFP bleaching

Example Data 3

Orphaned Images

The Plot

startedAtTime	Experiment	AgentRole	AgentName
2018-02-28T...	A1+fcGMP	Project	fcGMP affir
2017-02-28T...	fcGMP Disp...	Research G...	ReceptorLi
2018-02-28T...	A1+fcGMP	Research G...	ReceptorLi
2017-02-28T...	fcGMP Disp...	Project	FRET speci

PreviousPage 1 of 1Next5 rows

The Characters

rsonName	Experiment	Plan	PersonRole
	fcGMP Disp...	Solution Pr...	Aliquots Re...
	A1+fcGMP	Solution Pr...	Aliquots Re...
	fcGMP Disp...	Solution Pr...	Aliquots Re...
	A1+fcGMP	Solution Pr...	Aliquots Re...
	fcGMP Disp...	Solution Pr...	Aliquots Re...

PreviousPage 1 of 5Next5 rows

Materials

VectorPlasmidProteinChemicalSolutionDNARNARestriction Enzyme

Fluorescent ProteinOligonucleotide

StoredAt	Experiment	ReferencesMater...	UniqueName	MaterialReferenc...	Name
-4°C	A1+fcGMP	150mM KCl + 1μ...	KCl		150mM KCl
4°C	fcGMP Displace	150mM KCl + 1μ...	KCl		150mM KCl

# Visualization of Provenance Data with Dashboard

- Visualized at the project level
- Competency questions were converted to SPARQL queries
- The answers to these questions are represented as tables in the dashboard.
- The dashboard provides a panel for each component of a story.
- Data tables : users can search and filter the data

# Visualization of Provenance Data with Dashboard

- Plot
- Characters
- Experiment Materials
- External Resources
- Steps
- Devices
- Settings
- Jupyter Notebooks
- Results

# Evaluation

- User-based evaluation
- Data-based evaluation
- The results of SPARQL queries in the dashboard were manually compared and their correctness was evaluated by the domain experts.
- Results:  
<https://sheeba-samuel.github.io/REPRODUCE-ME/resources.html>



# Conclusions and Future Work

- Data provenance is a key factor towards reproducibility of scientific experiments.
- A provenance-based semantic approach to explain the story of a scientific experiment from its plot to its output.
- The REPRODUCE-ME ontology extended from the existing ontologies PROV-O and P-Plan, is used to represent a whole picture of an experiment including the plot, characters, settings, plans, steps, input and output.
- Scalability and performance of the system

# References

- <https://www.w3.org/TR/prov-o/>
- <http://purl.org/net/p-plan>
- ProvBook: Provenance-based Semantic Enrichment of Interactive Notebooks for Reproducibility, Sheeba Samuel, Birgitta König-Ries, The 17th International Semantic Web Conference (ISWC) 2018 Demo Track, 8-12 October, 2018, Monterey, California, USA ([Link](#))
- Combining P-Plan and the REPRODUCE-ME Ontology to Achieve Semantic Enrichment of Scientific Experiments using Interactive Notebooks, Sheeba Samuel, Birgitta König-Ries, 15th Extended Semantic Web Conference (ESWC) 2018 Poster Track, 3-7 June, 2018, Heraklion, Crete, Greece ([Link](#))
- REPRODUCE-ME: Ontology-based Data Access for Reproducibility of Microscopy Experiments, Sheeba Samuel, Birgitta König-Ries, 14th Extended Semantic Web Conference (ESWC) 2018 Poster Track, 28 May-1 June, 2017, Portoroz, Slovenia ([Link](#))
- Towards reproducibility of microscopy experiments, Sheeba Samuel, Frank Taubert, Daniel Walther, Birgitta König-Ries, H Martin Bucker, D-Lib Magazine 23.1/2 (2017) ([Link](#))
- Image Courtesy: Wikimedia Common, Pixabay

# Thanks

- Questions???
- Find more information here:
  - <https://w3id.org/reproduceme>
  - <https://w3id.org/reproduceme/research>