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

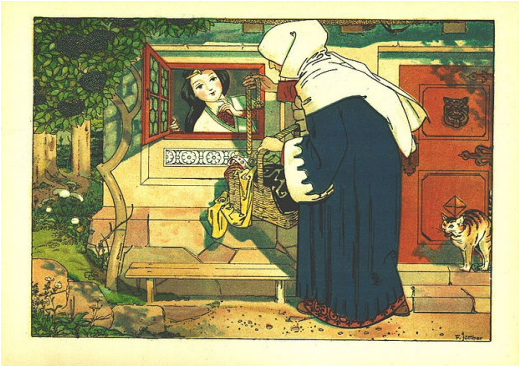
Sheeba Samuel, Kathrin Groeneveld, Frank Taubert , Daniel Walther, Tom Kache, Teresa Langenstück , Birgitta König-Ries, H. Martin Bücker, and Christoph Biskup

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

SWAT4HCLS, 4th December 2018

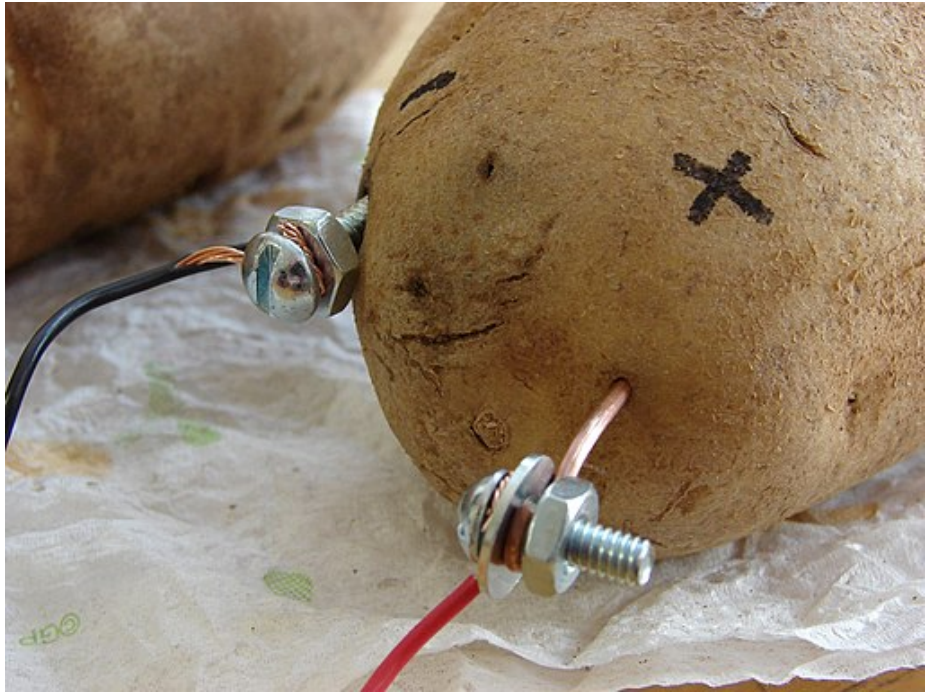
Deutsche Forschungsgemeinschaft

Story

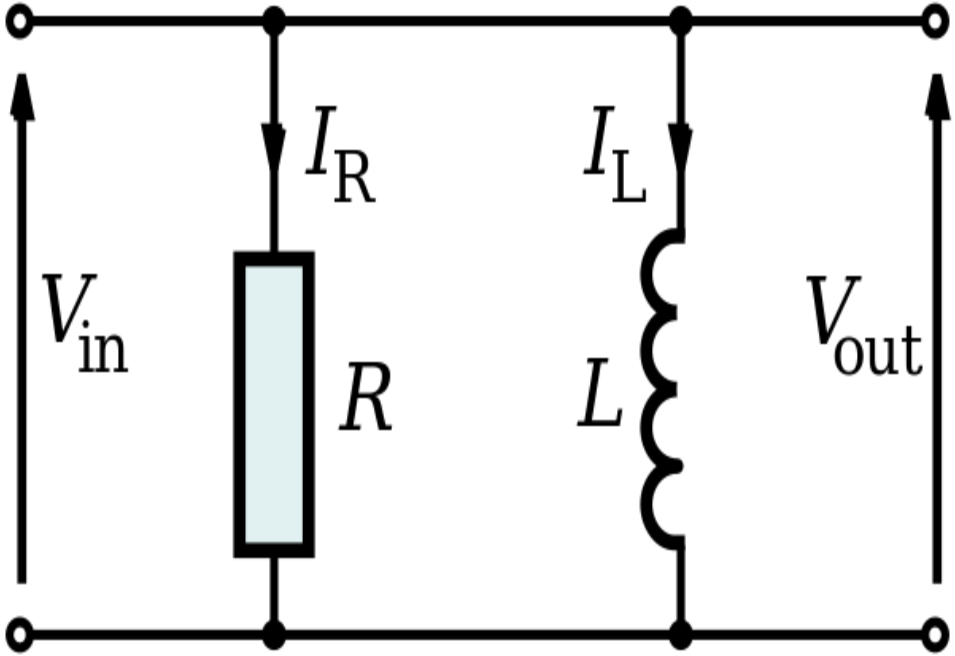


Snow White and the Seven Dwarfs https://commons.wikimedia.org/wiki/

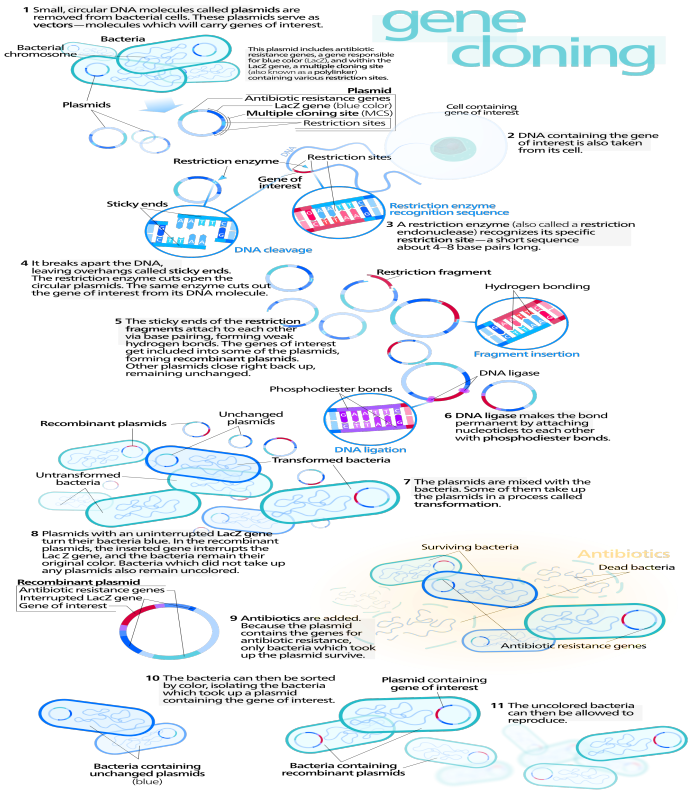
Story of an Experiment

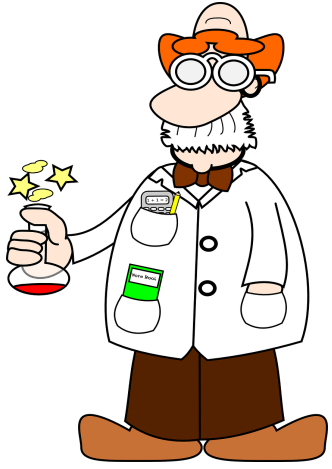


Story of an Experiment

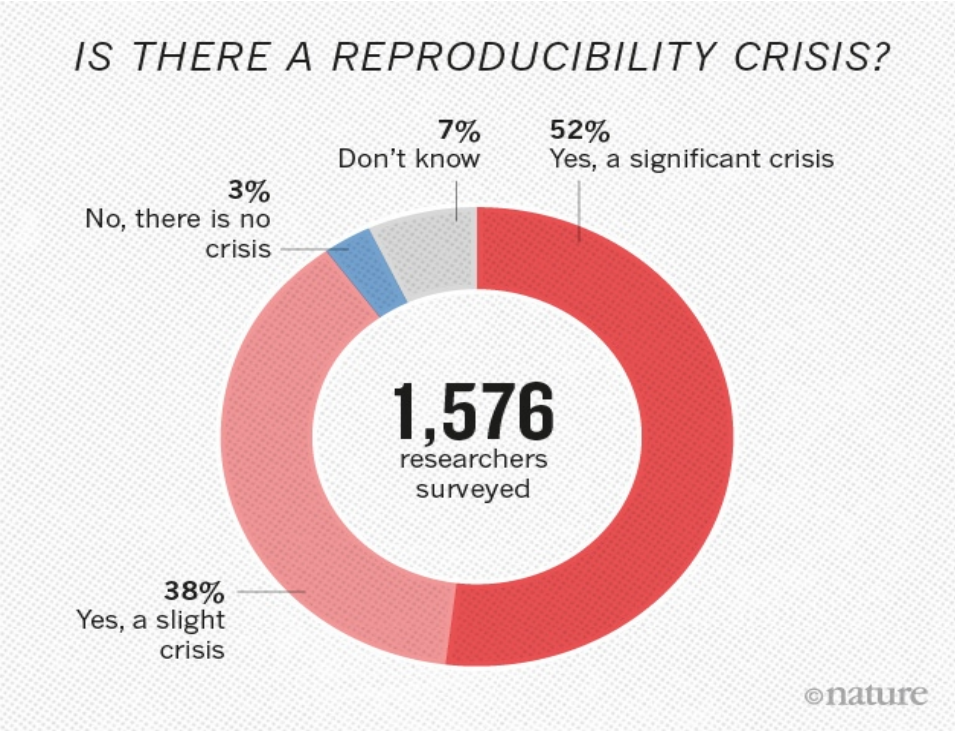


Story of an Experiment





Reproducibility



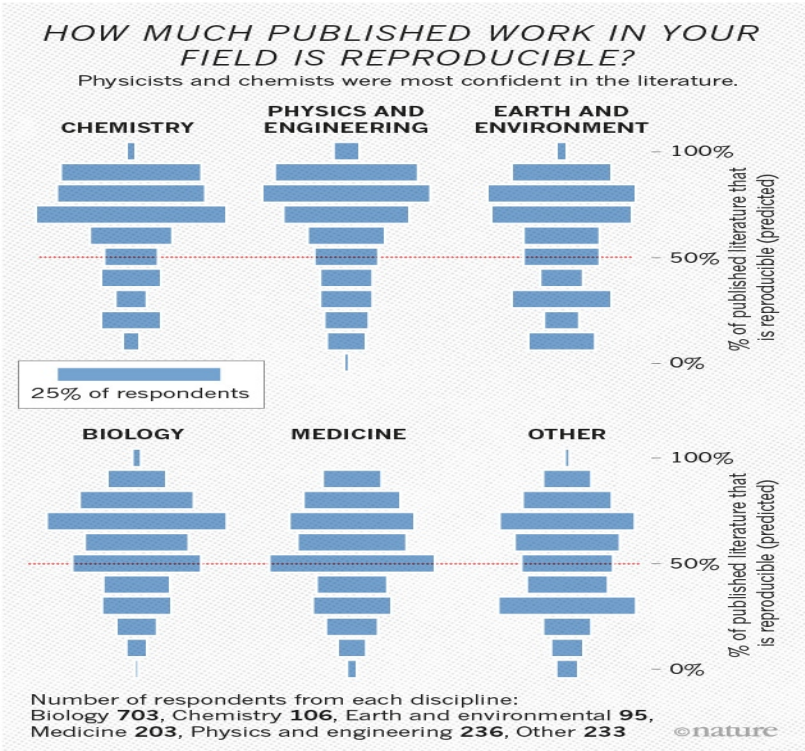


Image Courtesy: 1,500 scientists lift the lid on reproducibility, Nature

Challenges that hinder research reproducibility

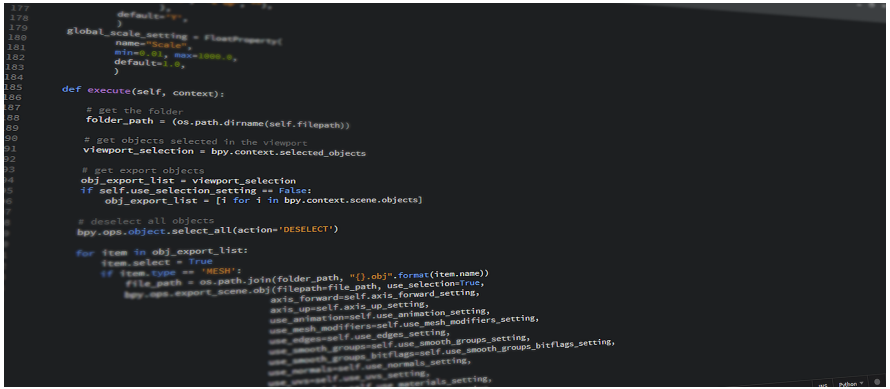
- Lack of documentation in digital media
- Non-availability of the datasets, code, workflow...
- Integration of data generated from different devices
- Incomplete and uncertain provenance information
- Lack of knowledge of the type of data and their formats and most importantly their semantics.

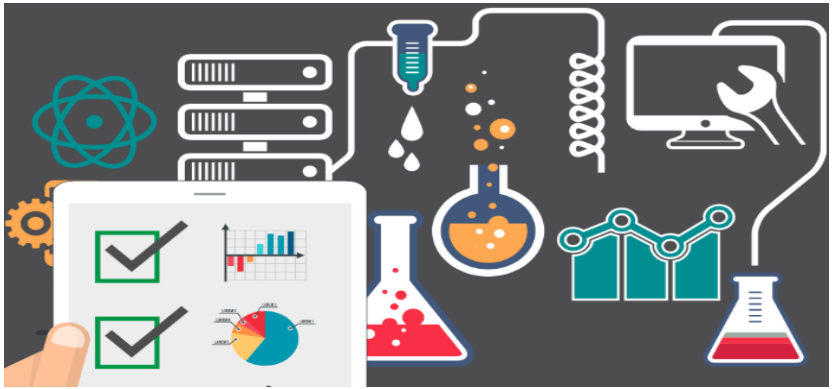
Contributions

- 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

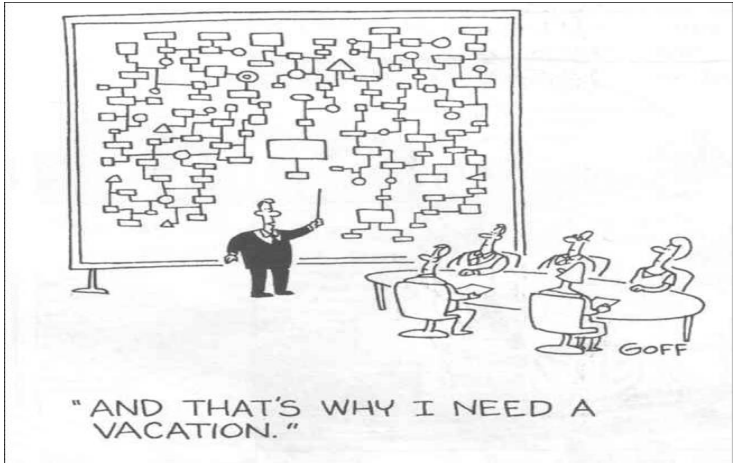
Interviews with the scientists in the CRC ReceptorLight as well as a workshop conducted to foster reproducible science.



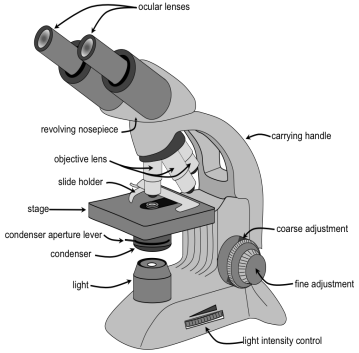


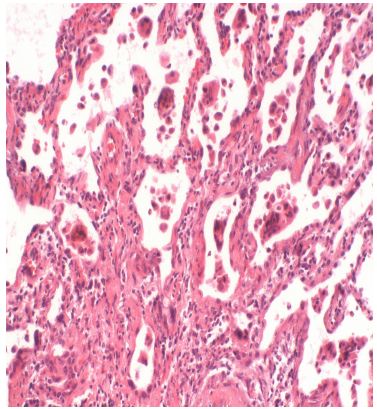
Scientific Experiments

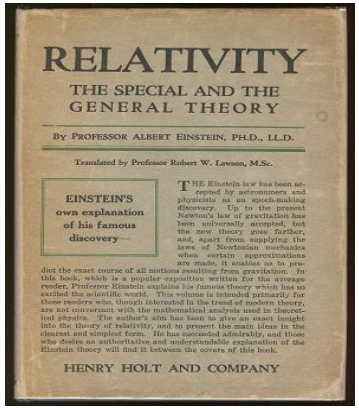
- Complex
- Several steps
- Several activities in the real world or cyberspace.
- Several people
- New technologies



Non-Computational Parts







RELATIVITY

THE SPECIAL AND THE GENERAL THEORY

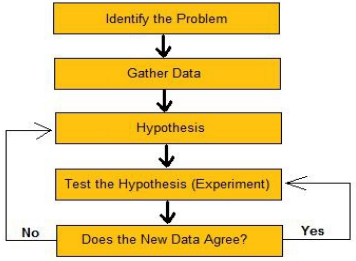
By Professor Albert Einstein, Ph.D., LL.D.

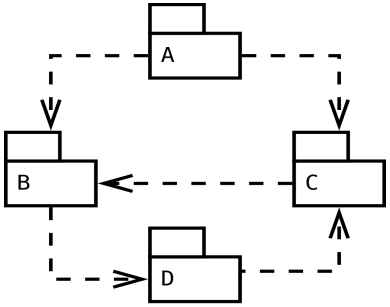
Translated by Professor Robert W. Lawson, M.Sc.

EINSTEIN'S own explanation of his famous discovery—

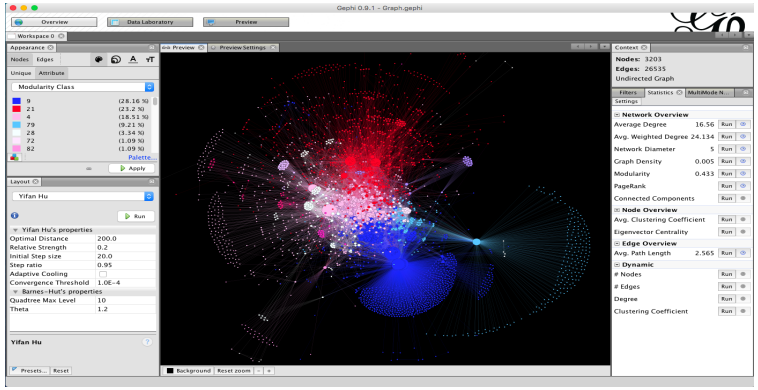
THE Einstein law has been accepted by astronomers and physicists as an epoch-making discovery. Up to the present Newton's law of gravitation has been universally accepted, but the new theory goes further, and, apart from supplying the laws of Newtonian mechanics when certain approximations are made, it enables us to predict the exact course of all motions resulting from gravitation. In this book, which is a popular exposition written for the average reader, Professor Einstein explains his famous theory which has so excited the scientific world. This volume is intended primarily for those readers who, though interested in the trend of modern theory, are not conversant with the mathematical analysis used in theoretical physics. The author's aim has been to give an exact insight into the theory of relativity, and to present the main ideas in the clearest and simplest form. He has succeeded admirably, and those who desire an authoritative and understandable explanation of the Einstein theory will find it between the covers of this book.

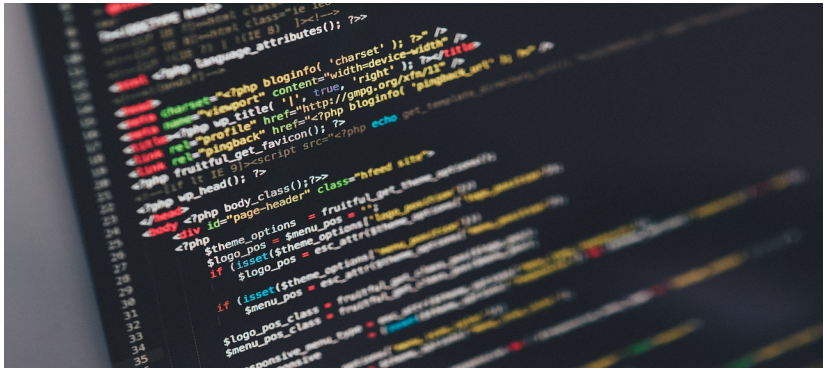
HENRY HOLT AND COMPANY

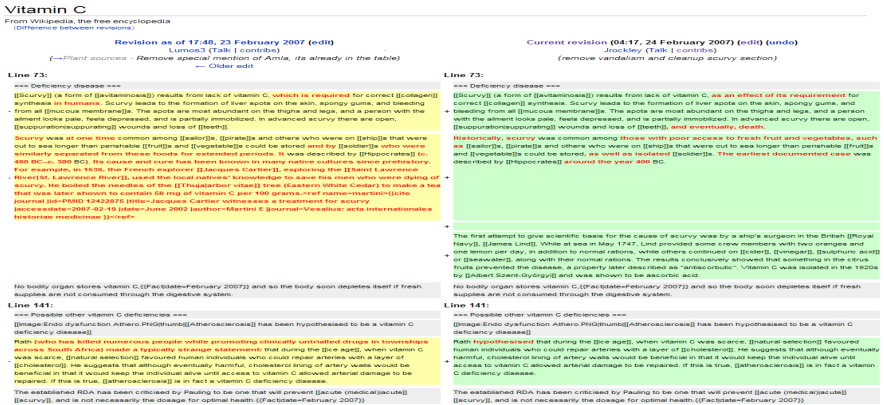




Computational Parts







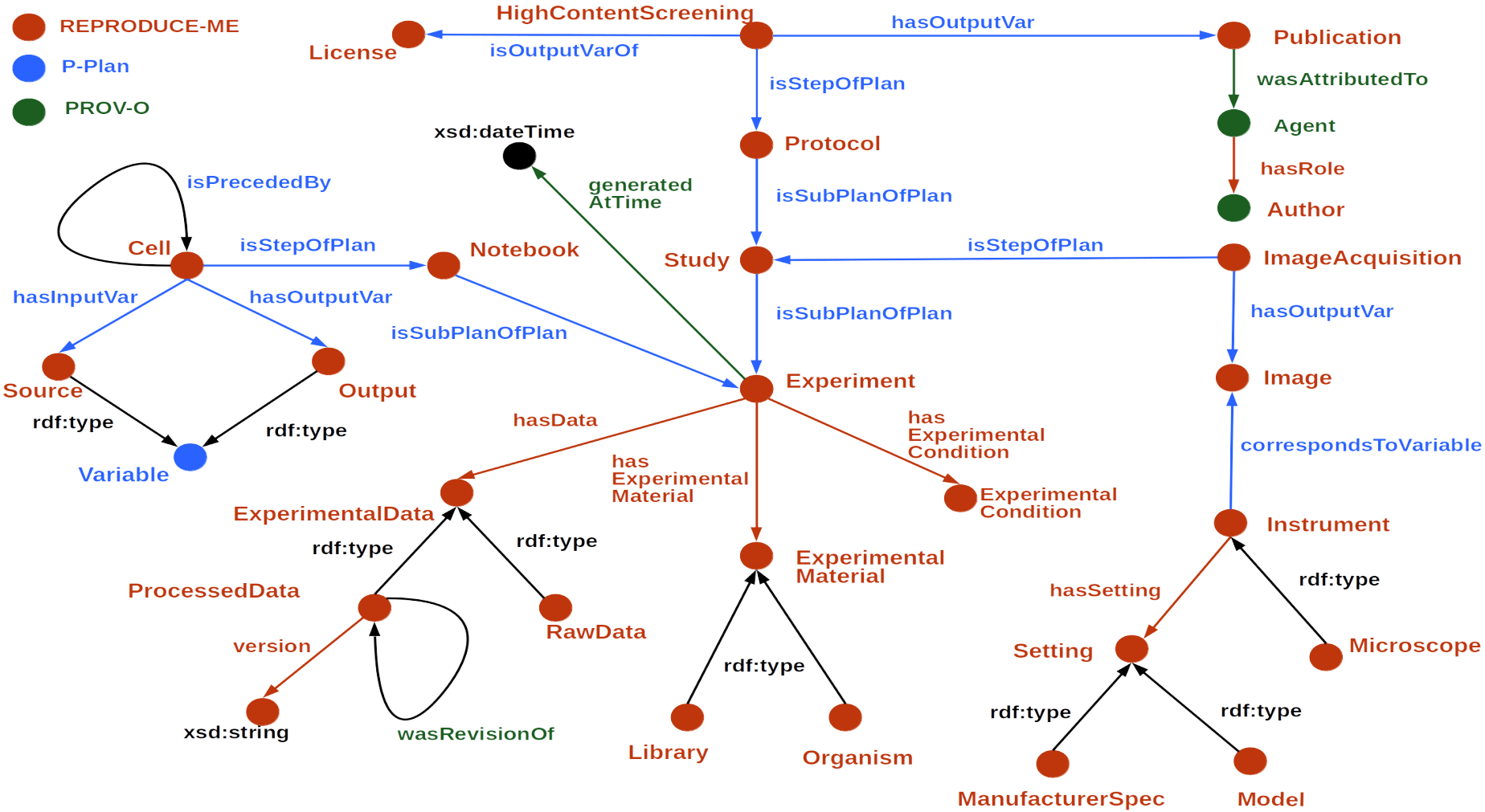
Competency Questions

- 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**

- **C**oll**A**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

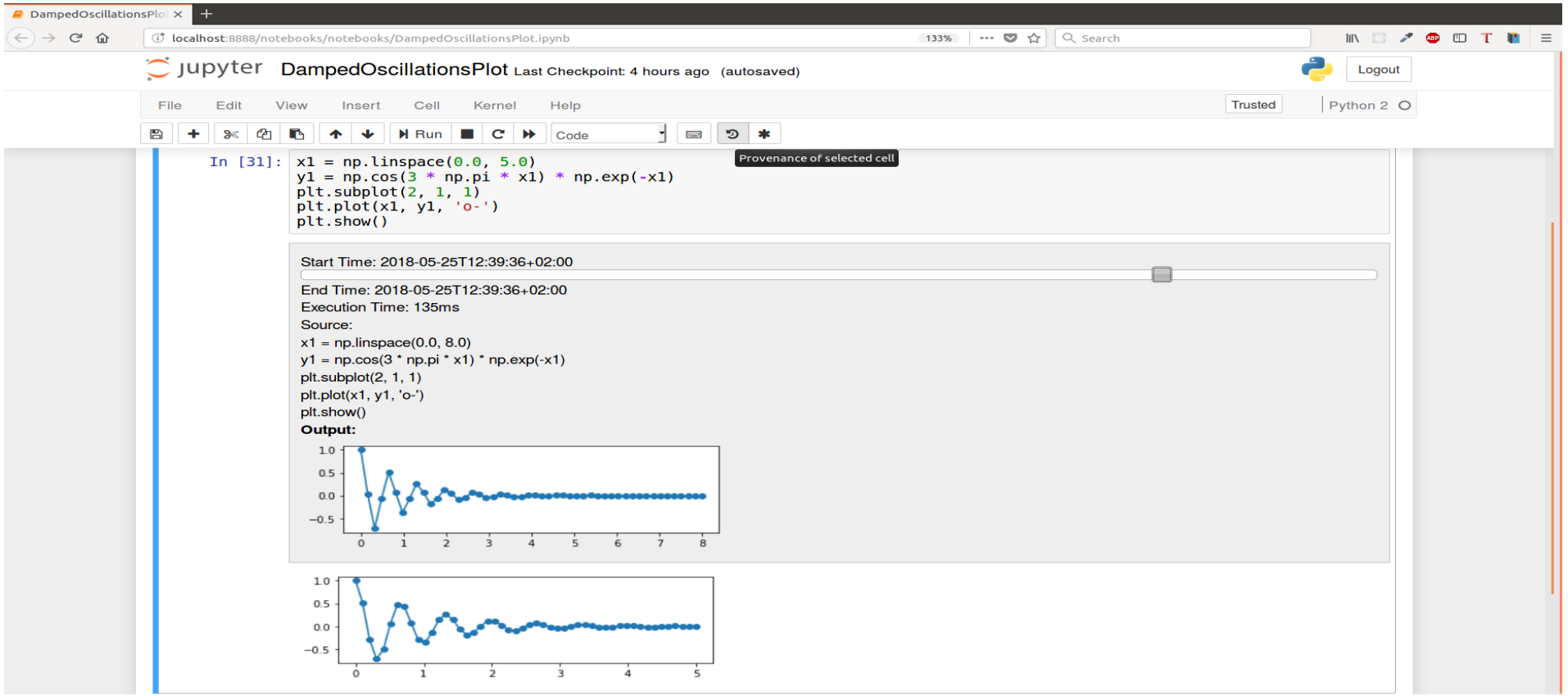
CAESAR- Computational Part Features

- 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

Combining P-Plan and the REPRODUCE-ME Ontology to Achieve Semantic Enrichment of Scientific Experiments using Interactive Notebooks, ESWC 2018 Satellite Events

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.

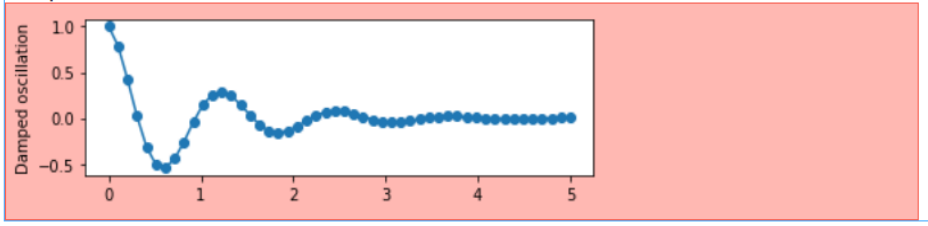
ProvBook Diff

☐ Hide unchanged cells Export diff

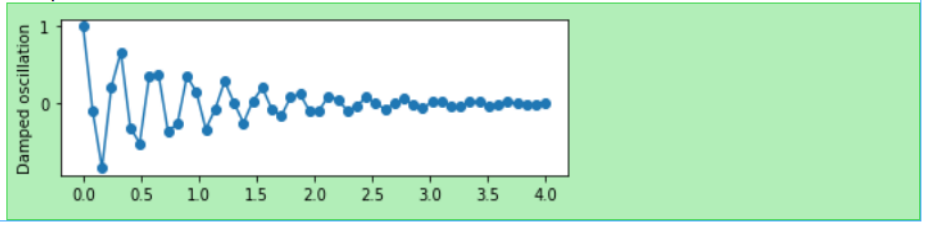
Base	Remote
In [16]:	In [16]:
`x1 = np.linspace(0.0, 5.0)`	`x1 = np.linspace(0.0, 4.0)`
`y1 = np.cos(18 * np.pi * x1) * np.exp(-x1)`	`y1 = np.cos(18 * np.pi * x1) * np.exp(-x1)`
`plt.subplot(2, 1, 1)`	`plt.subplot(2, 1, 1)`
`(...)`	`(...)`

Outputs changed

Output deleted



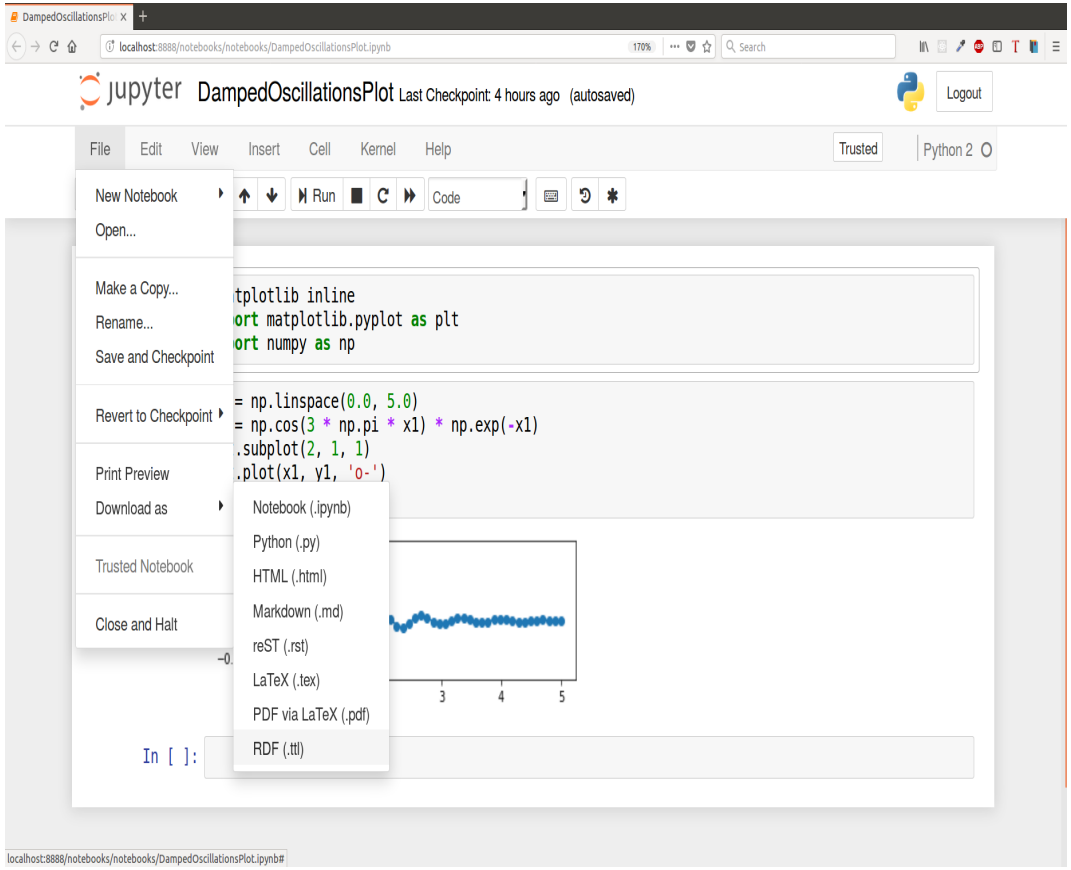
Output added



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

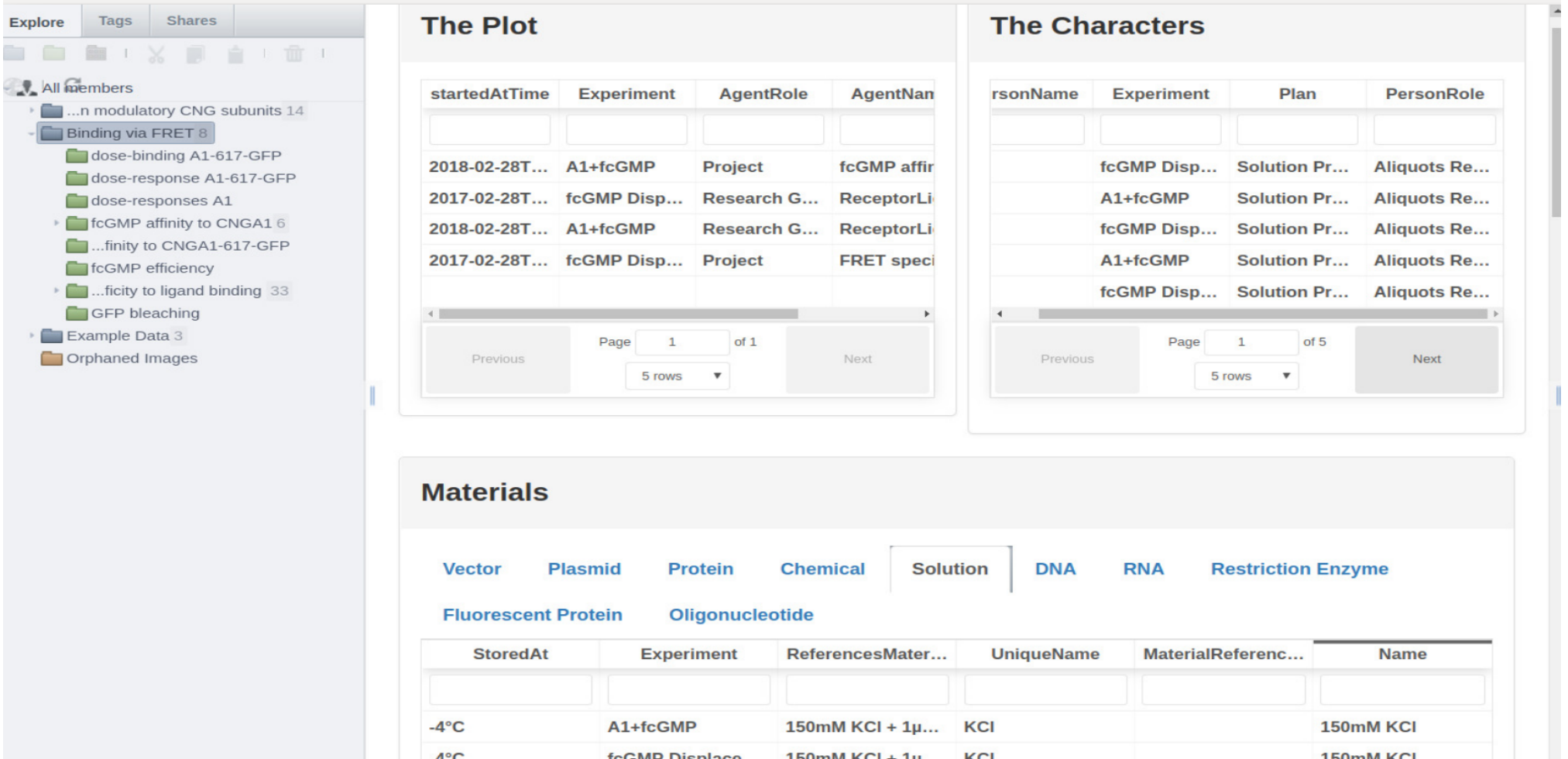
ProvBook

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




```
@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



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 Bücker, 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