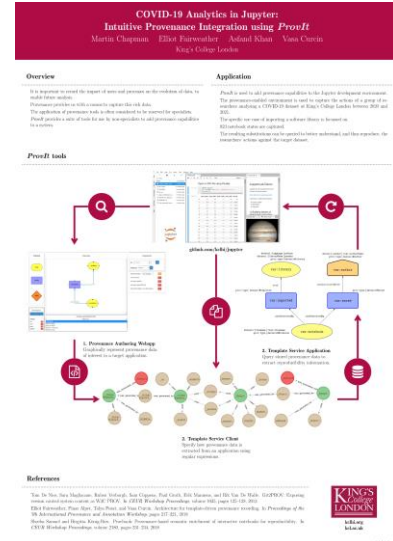


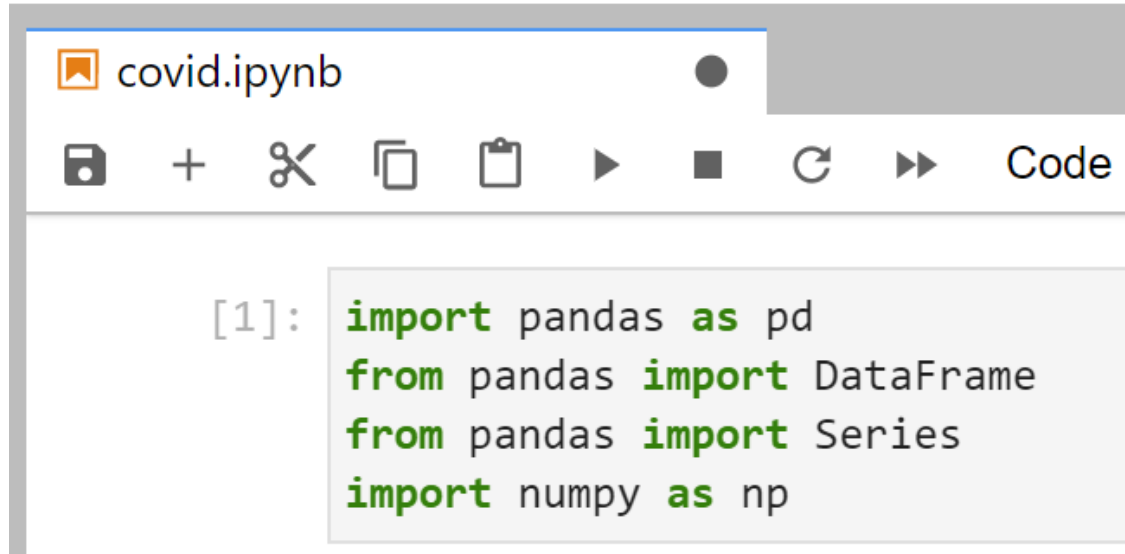
COVID-19 Analytics in Jupyter: Intuitive Provenance Integration using *ProvlIt*

Martin Chapman, Elliot Fairweather, Asfand Khan and Vasa Curcin
King's College London



Provenance-enabled Jupyter environment: <https://github.com/kclhi/jupyter>

How do I properly specify the data I want to capture?



The image shows a Jupyter Notebook window titled 'covid.ipynb'. The toolbar includes icons for saving, adding, deleting, copying, pasting, running, and a 'Code' button. The code cell contains the following Python code:

```
[1]: import pandas as pd
      from pandas import DataFrame
      from pandas import Series
      import numpy as np
```

How do I properly specify the data I want to capture?

covid.ipynb



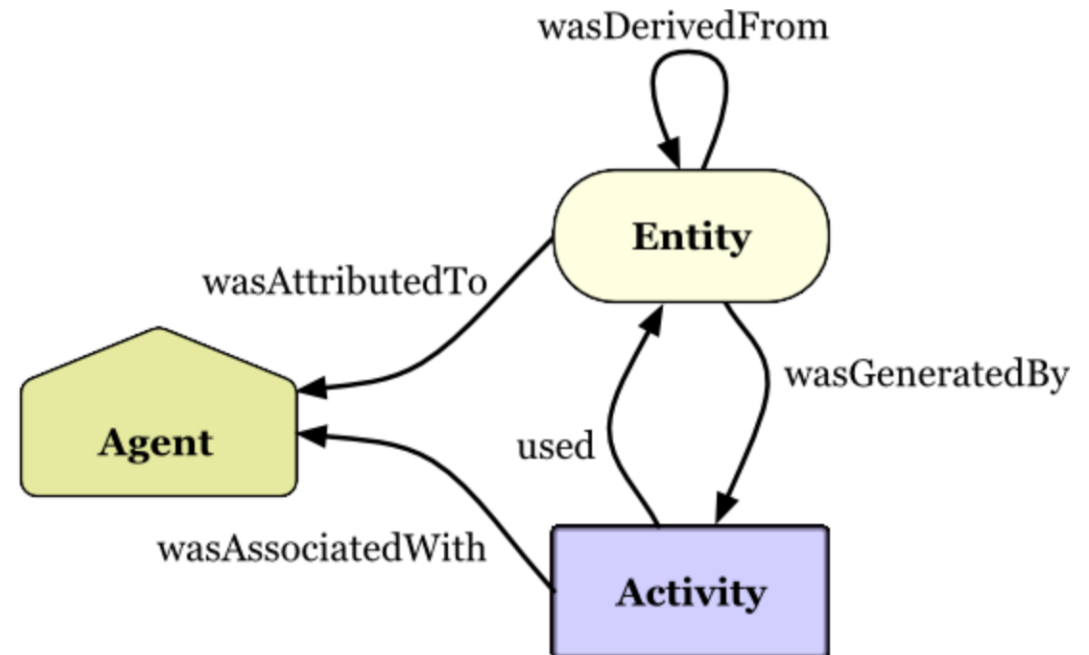
```
[1]: import pandas as pd
      from pandas import DataFrame
      from pandas import Series
      import numpy as np
```

W3C Working Group Note

2. Intuitive overview of PROV

This section provides an explanation of the main concepts in PROV. As with the rest of this document, it should be treated as a starting point for understanding the model. The PROV data model document [[PROV-DM](#)] provides precise definitions and constraints [[PROV-CONSTRAINTS](#)] to be followed.

The following diagram provides a high level overview of the structure of PROV records, limited to some key PROV concepts discussed in this document. Note that because PROV is meant to describe how things were created or delivered, PROV relations are named so they can be used in assertions about the past.



How do I properly specify the data I want to capture?

covid.ipynb

Code

```
[1]: import pandas as pd
from pandas import Series, DataFrame
import numpy as np
```

entity

activity

agent

Namespace

+

default

var

vvar

zone

datasci

Canvas

```
graph BT
    notebook((var:notebook)) -- wasGeneratedBy --> imported[var:imported]
    imported -- used --> library((var:library))
```

Choose conversion format:

PROV-N

PROV-N

PROV-XML

PROV-RDF

PROV-TURTLE

PROV-TriG

Inspector

ID: var:library var

Attribute: datasci:

datasci:language vvar:language

zone:id: import

zone:type: parallel

prov:type: datasci#Library

datasci:libraryName: vvar:libraryName

How do I properly specify the data I want to capture?

covid.ipynb

Code

```
[1]: import pandas as pd
from pandas import Series
from pandas import DataFrame
import numpy as np
```

Provenance
Authoring Webapp

Palette

- entity
- activity
- agent

Namespace

default

- var
- vvar
- zone
- datasci

Canvas

var:library

var:imported

var:notebook

used

wasGeneratedBy

Inspector

ID: var:library var ✓

Attribute: datasci: +

datasci:language vvar:language ✓

zone:id: import ✗

zone:type: parallel ✗

prov:type: datasci#Library ✗

datasci:libraryName: vvar:libraryName ✗

Choose conversion format:

PROV-N

PROV-XML










PROV-RDF

PROV-TURTLE

PROV-TriG

How do I extract this data from a running system?

covid.ipynb

         Code

```
[1]: import pandas as pd
      from pandas import DataFrame
      from pandas import Series
      import numpy as np
```

datasci:language	vvar:language
datasci:libraryName	vvar:libraryName
prov:type	datasci#Library
zone:id	import
zone:type	parallel

prov:type	datasci#Imported
zone:id	import
zone:type	parallel

prov:type	datasci#Notebook
-----------	------------------

var:library

used

var:imported

wasGeneratedBy

var:notebook



How do I extract this data from a running system?

covid.ipynb

+

✂

📄

📋

▶

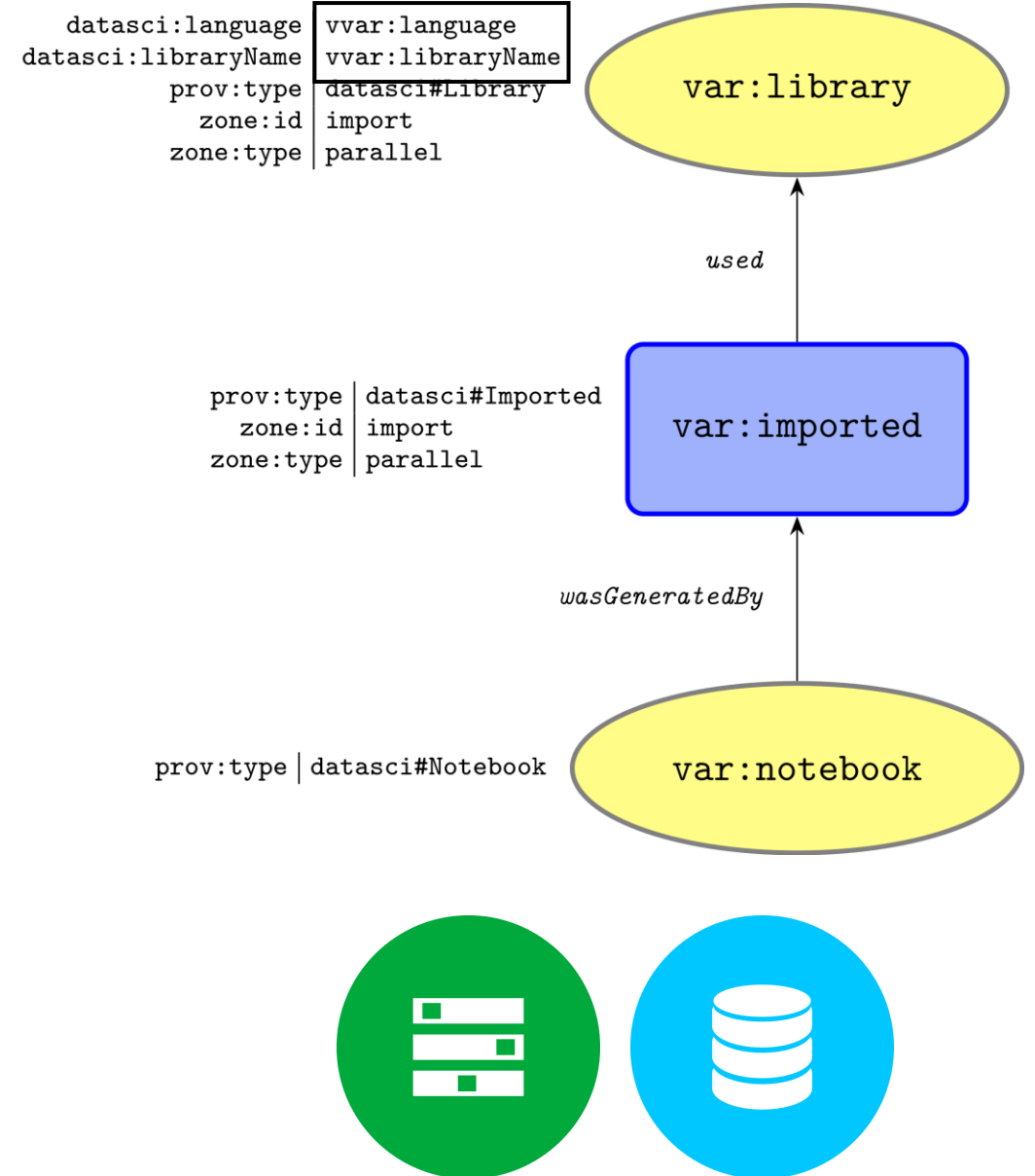
■

↺

▶▶

Code

```
[1]: import pandas as pd
from pandas import DataFrame
from pandas import Series
import numpy as np
```



How do I extract this data from a running system?

covid.ipynb

+

✂

📄

📋

▶

■

🔄

⏩

Code

```
[1]: import pandas as pd
    from pandas import DataFrame
    from pandas import Series
    import numpy as np
```

datasci:language	vvar:language
datasci:libraryName	vvar:libraryName
prov:type	datasci#Library
zone:id	import
zone:type	parallel

prov:type	datasci#Imported
zone:id	import
zone:type	parallel

prov:type	datasci#Notebook
-----------	------------------

var:library

used

var:imported

wasGeneratedBy

var:notebook



How do I extract this data from a running system?

covid.ipynb

+

✂

📄

📋

▶

■

🔄

⏩

Code

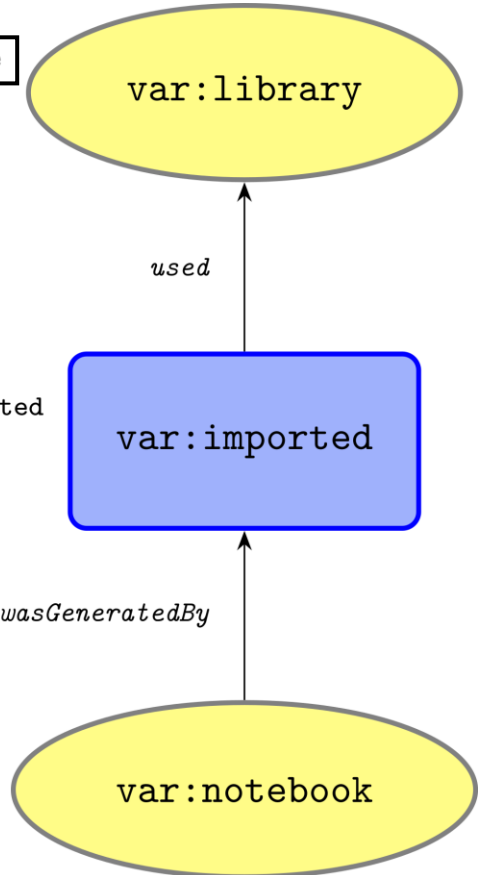
```
[1]: import pandas as pd
    from pandas import DataFrame
    from pandas import Series
    import numpy as np
```

Pattern: import []
Action: extract

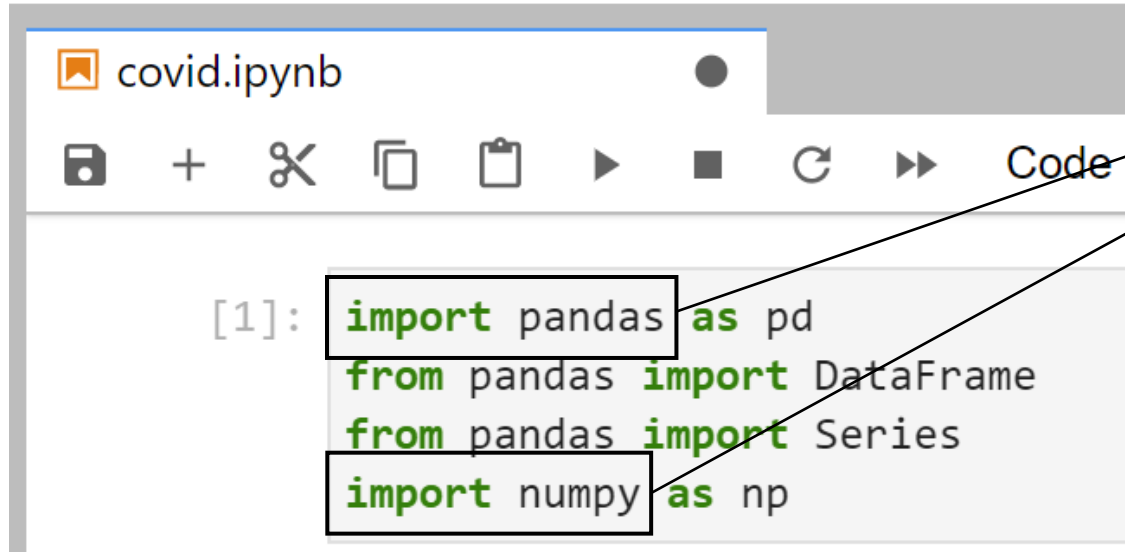
datasci:language	vvar:language
datasci:libraryName	vvar:libraryName
prov:type	datasci#Library
zone:id	import
zone:type	parallel

prov:type	datasci#Imported
zone:id	import
zone:type	parallel

prov:type	datasci#Notebook
-----------	------------------



How do I extract this data from a running system?



The screenshot shows a Jupyter Notebook titled 'covid.ipynb'. The code cell contains the following Python code:

```
[1]: import pandas as pd
      from pandas import DataFrame
      from pandas import Series
      import numpy as np
```

Arrows from the code snippets below point to the corresponding lines in this code cell.

Pattern: import []

Action: extract

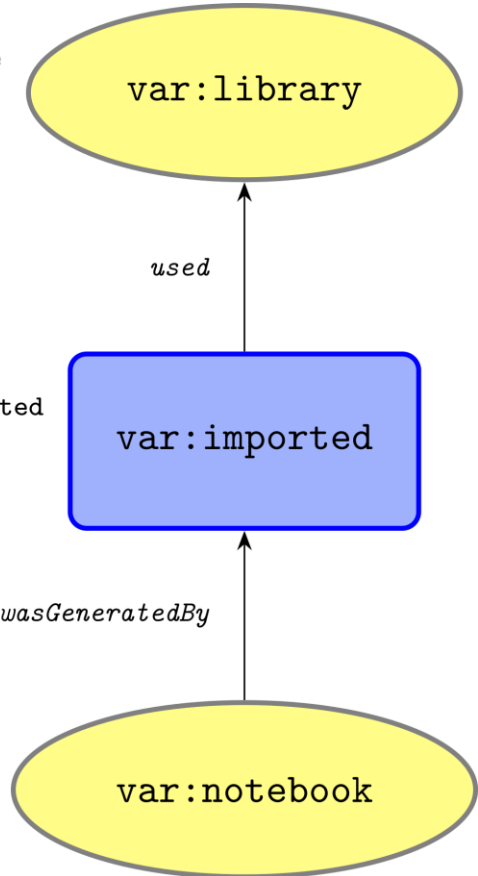
Pattern: import []

Action: use string 'python'


datasci:language	vvar:language
datasci:libraryName	vvar:libraryName
prov:type	datasci#Library
zone:id	import
zone:type	parallel










prov:type	datasci#Imported
zone:id	import
zone:type	parallel

prov:type | datasci#Notebook



How do I extract this data from a running system?

 covid.ipynb

         Code

```
[1]: import pandas as pd
    from pandas import DataFrame
    from pandas import Series
    import numpy as np
```

Pattern: import []
Action: extract

Pattern: import []
Action: use string 'python'

Template Service Client

datasci:language	vvar:language
datasci:libraryName	vvar:libraryName
prov:type	datasci#Library
zone:id	import
zone:type	parallel

prov:type	datasci#Imported
zone:id	import
zone:type	parallel

var:library

used

var:imported

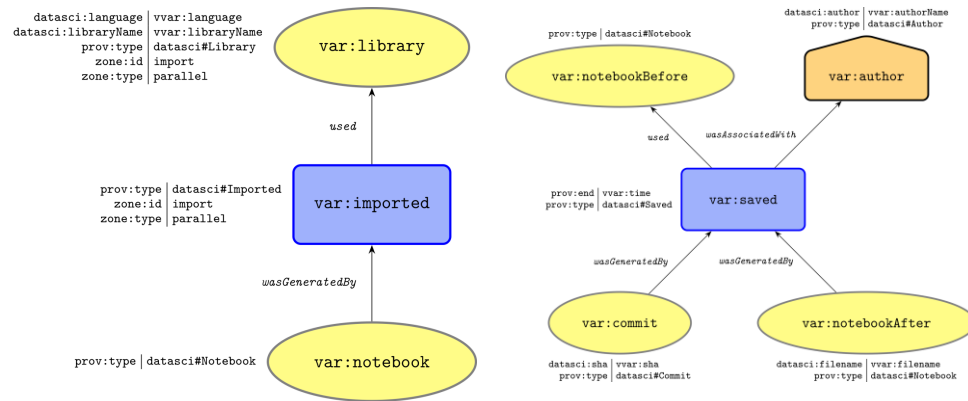
wasGeneratedBy

var:notebook

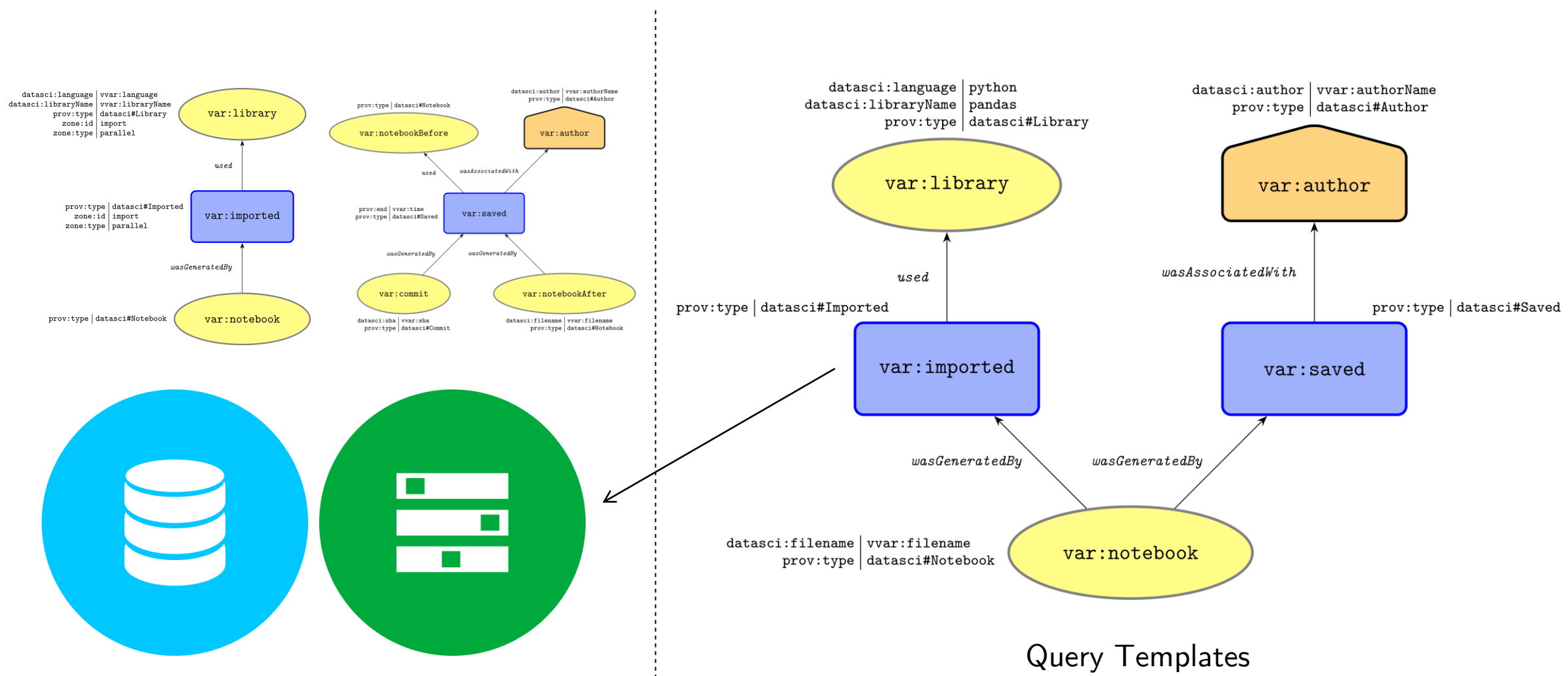
prov:type | datasci#Notebook



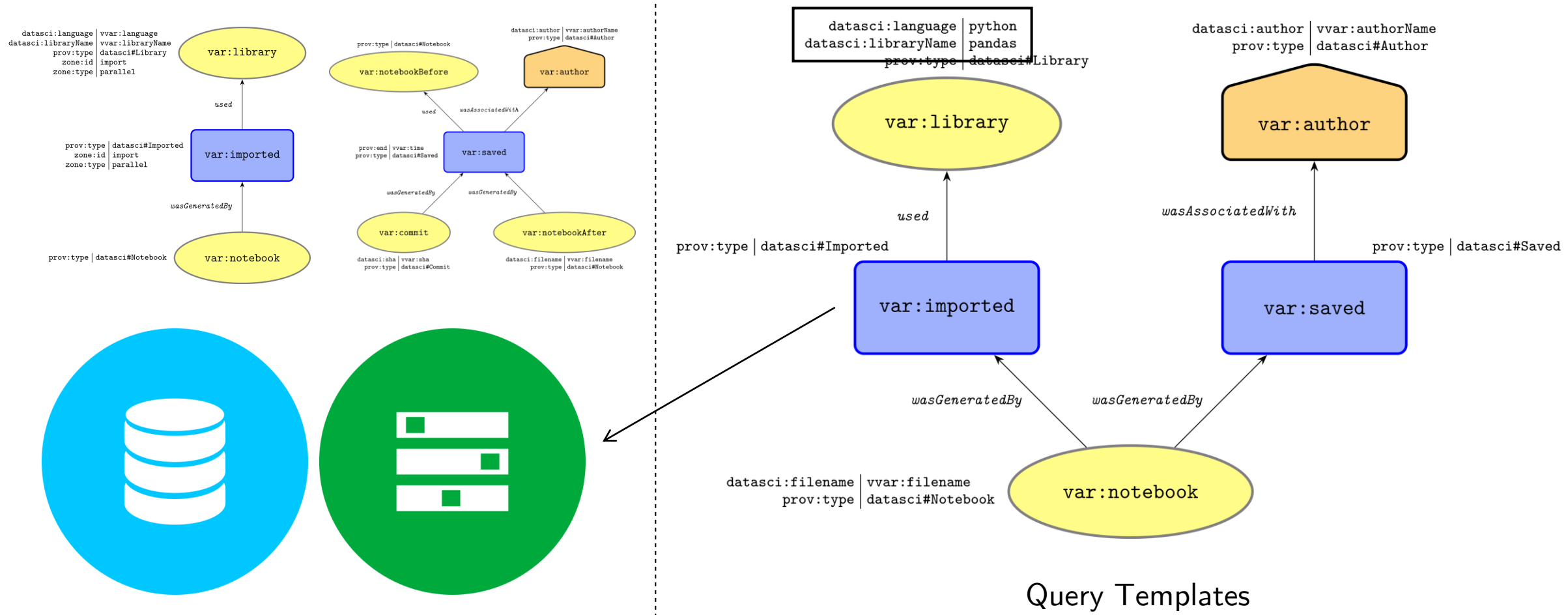
How do I get my data back, once it's been stored?



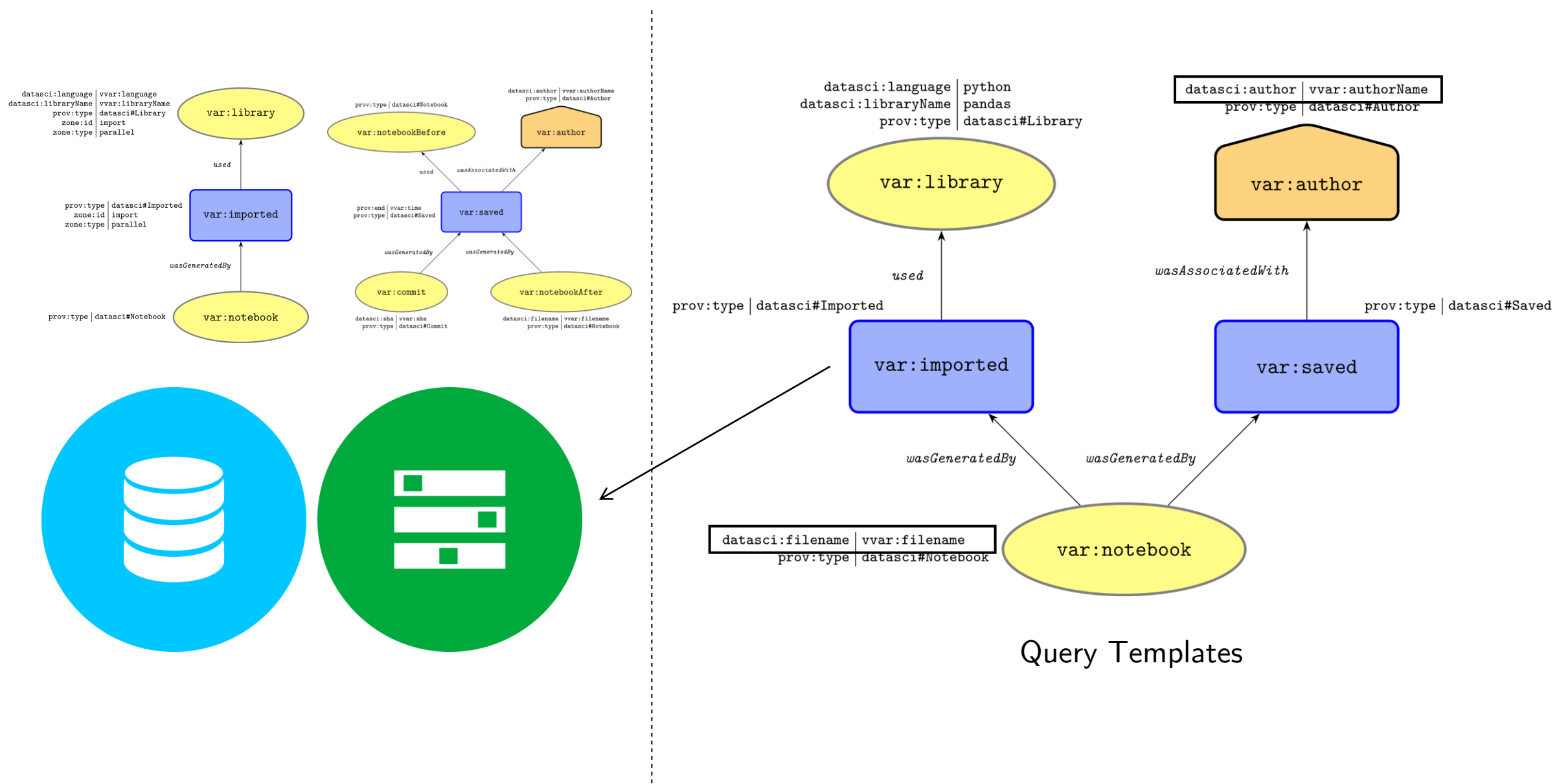
How do I get my data back, once it's been stored?



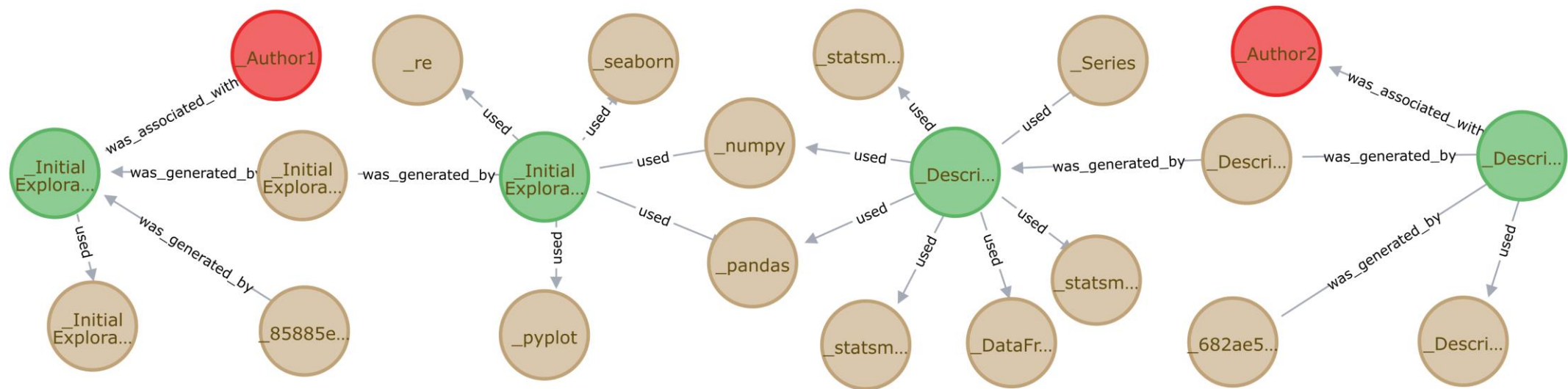
How do I get my data back, once it's been stored?



How do I get my data back, once it's been stored?

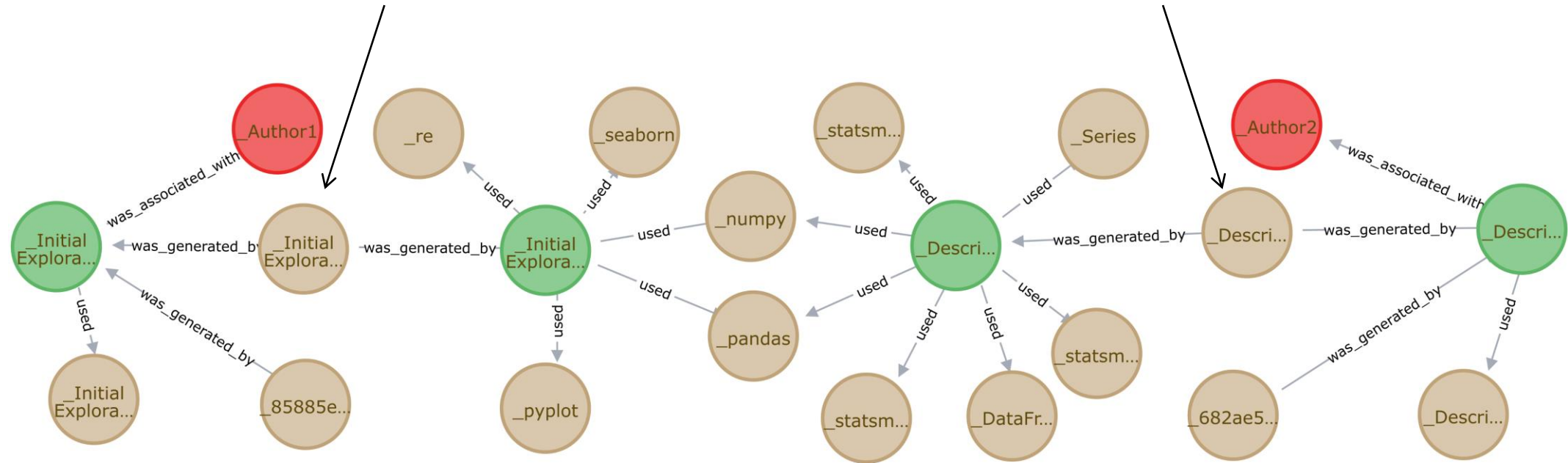


COVID-19 Analytics in Jupyter

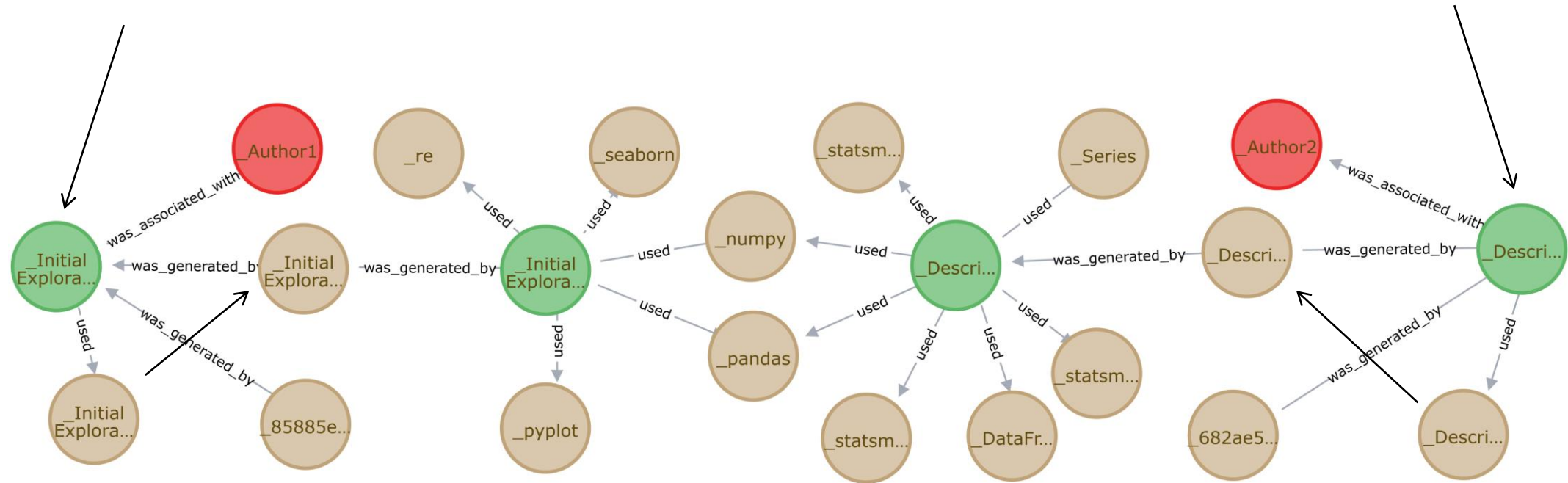


Data relating to 1468 patients who tested positive at Guy's and St. Thomas' NHS Foundation Trust (GSTT), analysed by a group of researchers between April 2020 and February 2021 at King's College London.

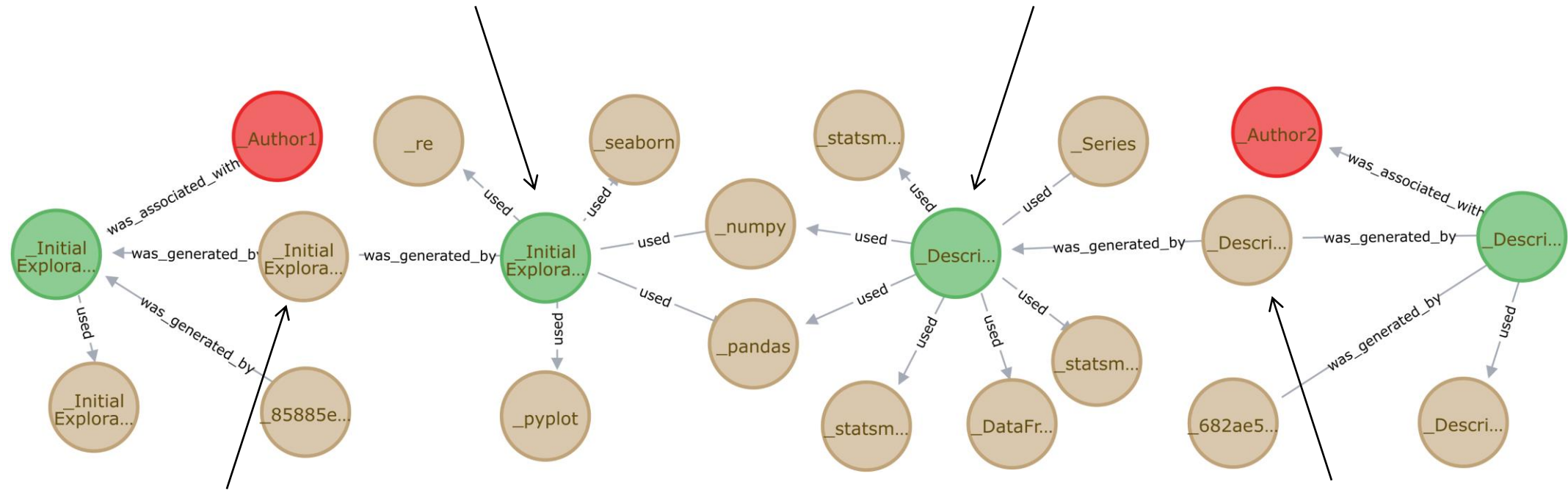
COVID-19 Analytics in Jupyter



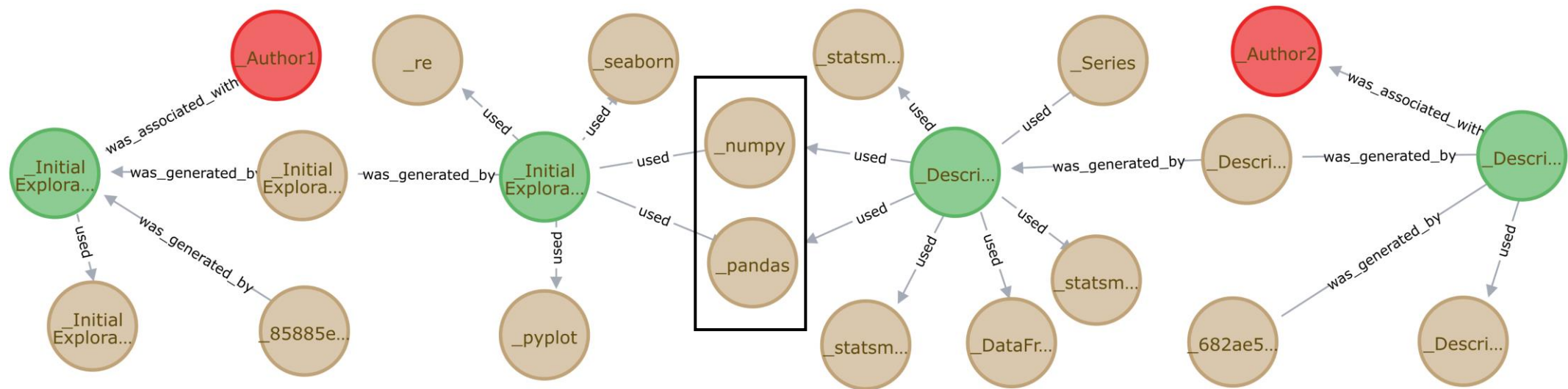
COVID-19 Analytics in Jupyter



COVID-19 Analytics in Jupyter



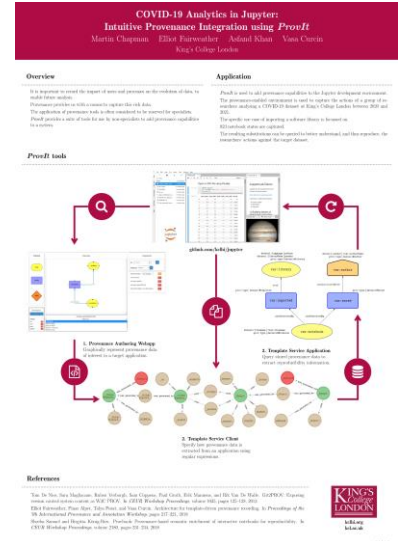
COVID-19 Analytics in Jupyter



When this is expanded across multiple saves, we can see what was imported when, and by whom; we can also see more complex information, such as common libraries

COVID-19 Analytics in Jupyter: Intuitive Provenance Integration using *ProvlIt*

Martin Chapman, Elliot Fairweather, Asfand Khan and Vasa Curcin
King's College London



Provenance-enabled Jupyter environment: <https://github.com/kclhi/jupyter>