

# tmtk and the Arborist

ETL working group



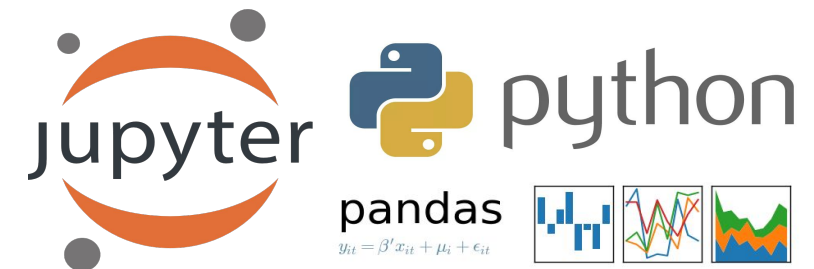
the hyve

Jun 28, 2018

# tmtk and the Arborist

- What
  - **IranSMART ToolKit**  
Python package for modelling data in Jupyter Notebook
  - **The Arborist**  
Visual ontology tree editor
- Why
  - Data modelling is hard. Let's automate where possible.
  - Data modelling requires close collaboration between data manager and data owner.

Made possible by:

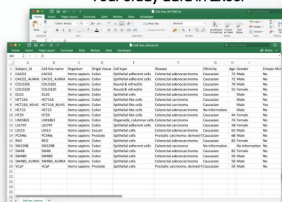


● <https://tmtk.readthedocs.io>


## From Excel to tranSMART

in five simple steps


**Your study data in Excel**



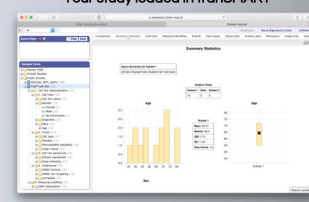
**Import:** start the import wizard to create a study based on your study data.



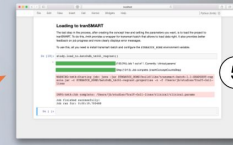
**Validate:** let the toolkit check the tranSMART-specific requirements.




**Your study loaded in tranSMART**



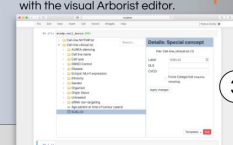
**Load:** use `smart-batch` to load your data to tranSMART.



**Save:** store the study on disk as tranSMART-ready staging files.



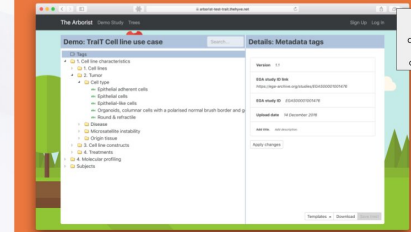
**Edit:** make changes to your tree with the visual Arborist editor.



**The Arborist ↓ Visual editor**

Collaborate on data modelling with non-technical data experts in the secure Arborist web application.

- Restructure the tranSMART tree with drag and drop
- Rename variables and values
- Add and edit metadata for any tree node
- Work with both low and high dimensional data



Send to the Arborist web application for easy collaboration!

Try it at <http://arborist-test-frait.thehyve.net/demo>  
 Code at <https://github.com/thehyve/arborist> under GPL v3 license.

**tmtk ↑ Python library**

Library that allows users to create and load studies without the need for tranSMART specific knowledge:

- Quickstart studies from tabular files (e.g. XLS, TSV, CSV)
- Extensive dataset validation
- Use **The Arborist** directly embedded into Jupyter Notebook
- Load studies to **The Arborist** web application for collaboration
- Many functions to work with **low and high dimensional data**
- **Minimal** technical and tranSMART specific knowledge required

Install for Python3: `$ pip install tmtk`

Documentation: <https://tmtk.readthedocs.io>

Code at <https://github.com/thehyve/tmtk> under GPL v3 license.

**tmtk notable python commands**

The main object in the tmtk workflow is the `Study`. It provides an API for modifying and validating your data. Below are the key methods and features provided by tmtk.

Starting a study	Validation of the data	TranSMART arborist
<code>create_study_from_template()</code> Create study from <b>TEXT</b> template. A way to create an entire study from filled in templates.	<code>validate_all()</code> Many of the objects in tmtk have validating methods, these methods can easily be extended by adding more.	<code>Visual</code> drag and drop editor for the tranSMART concept tree. Use it to change the concept tree, change word mappings, add metadata, and map concepts to ontologies.
<code>wizard.create_study()</code> Create Study from tabular files. Quickstart your tranSMART study.	Loading the data into tranSMART is provided a wrapper for <code>smart-batch</code> for easy use and better progress bars!	<code>call_boris()</code> Launch the Arborist embedded into Jupyter.
<code>randomstudy()</code> Generate fully randomised Study objects. Great for testing stuff!	<code>load_to()</code> Load your study to tranSMART from Jupyter or the console.	<code>publish_to_base()</code> Send data tree to Arborist web application for easy collaboration.



We empower scientists by building on open source software