Application de la technologie ActivePapers pour la biologie computationnelle

G. Chevrot
MISC - Universités d'Orléans et de Tours

Retour d'expéRiences sur la Recherche Reproductible (R^4)
Atelier MISC / CaSciModOT
Le Studium - Orléans

What is ActivePapers?

- K. Hinsen - 2011: first paper mentioning the principles of ActivePaper. [1]
- Goal: computational science made reproducible and publishable
- ActivePaper solution:
  - a single file combining datasets and programs that works on it
  - Citable (DOI - unique and persistant identifiers)

[1] K. Hinsen. Procedia Computer Science 2011 (4) 579

Why ActivePapers?

Molecular Dynamics simulations specificities:

- Size of the data: several Gb

- Local storage in binary format (size optimization - efficient reading)

- Need to work on different computer (HPC, personal computer)

- MOSAIC - MOlecular SimulAtion Interchange Conventions [2]:

Detailed description of molecular simulation

Facilitate the exchange of data

[2] K. Hinsen. J. Chem Inf. Model. 2014 (54) 131

ActivePaper design

- HDF5 file + conventions
- HDF is supported by many commercial and non-commercial software platforms: Java, MATLAB, Scilab, Mathematica, Octave, IDL, Python, R …
- Can also be inspected using many generic HDF5 tools, in particular HDFView or HDF Compass

HDF5: Hierarchical Data Format (version 5).

- open-source format
- designed to store large amounts of data
- provide a hierarchical structure
- self-describing
- Maintained by the non-profit HDF Group (spin-off of the NCSA*)

* NCSA: National Center for Supercomputing Applications

A concrete example

THE JOURNAL OF CHEMICAL PHYSICS **139**, 154110 (2013)

Model-free simulation approach to molecular diffusion tensors

Guillaume Chevrot,[1,2] Konrad Hinsen,[1,2] and Gerald R. Kneller[1,2,3,a)]
[1]*Centre de Biophysique Moléculaire, CNRS, Rue Charles Sadron, 45071 Orléans, France*
[2]*Synchrotron Soleil, L'Orme de Merisiers, 91192 Gif-sur-Yvette, France*
[3]*Université d'Orléans, Chateau de la Source-Av. du Parc Floral, 45067 Orléans, France*

(Received 5 July 2013; accepted 18 September 2013; published online 21 October 2013)

E. Availability of software and data

An ActivePaper[29] containing all the software, input datasets, and results from this study is available as the supplementary material.[30] The datasets can be inspected with any HDF5-compatible software, e.g., the free HDFView.[31] Running the programs on different input data requires the ActivePaper software.[32] These files also contain plots for all the components of all the correlation functions and mean-square displacements we have computed and of which we show only a selection in this article.

A concrete example - figshare

figshare.com/articles/Model_free_simulation_approach_to_molecular_diffusion_tensors_Water/808595

figshare

water_diffusion.ap download

Download

216 views

0 shares

cites coming soon

Published on **26 Sep 2013 - 14:59 (GMT)**

Filesize is **78.28 MB**

Share this: Share 0 Tweet 0 G+1 0 Embed*

Cite this: Chevrot, Guillaume; Hinsen, Konrad; Kneller, Gerald R. (2013): Model-free simulation approach to molecular diffusion tensors: Water. figshare. http://dx.doi.org/10.6084/m9.figshare.808595
Retrieved 17:31, Nov 27, 2015 (GMT)

*The embed functionality can only be used for non commercial purposes... more

Description

This file contains part 2 (water) of the supplementary material for the following publication:

Title: Model-free simulation approach to molecular diffusion tensors
Authors: Guillaume Chevrot, Konrad Hinsen, Gerald R. Kneller
Journal: Journal of Chemical Physics **139**, 154110 (2013)

It contains the software implementing the computations described in the article, the input datasets for water, the resulting output datasets, and the figures. A detailed list is given below.

The file water_diffusion.ap is a HDF5 file that can be read with any HDF5-compatible software, including the free HDFView package (http://www.hdfgroup.org/hdf-java-html/hdfview/). HDFView can be used to inspect the arrays and tables contained in this file. Reading the molecular structure for water requires software that understands the Mosaic data model (http://bitbucket.org/molsim/mosaic/).

Categories

- Condensed Matter Physics
- Computational Physics

Authors

Guillaume Chevrot
Konrad Hinsen
Gerald R. Kneller

Tags

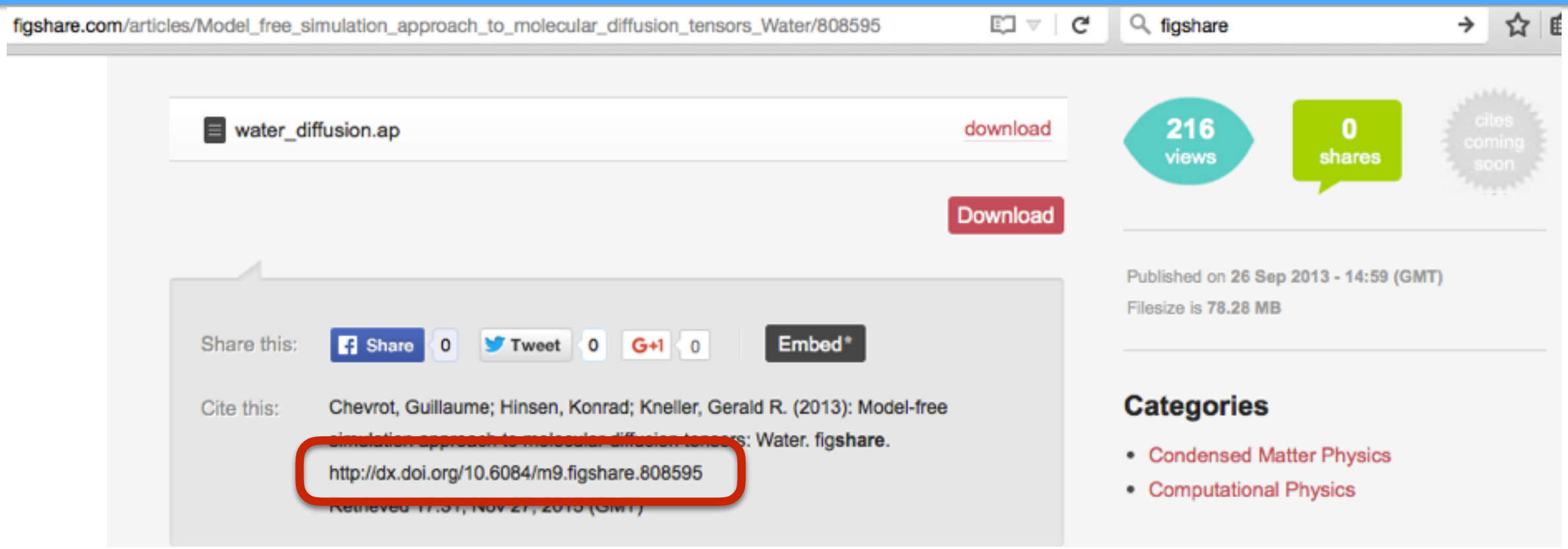
- Molecular Simulation
- ActivePapers
- diffusion tensor
- water

License (what's this?)

CC-BY

Export

A concrete example - DOI



DOI
Digital Object Identifier

DOI: Digital Object Identifier

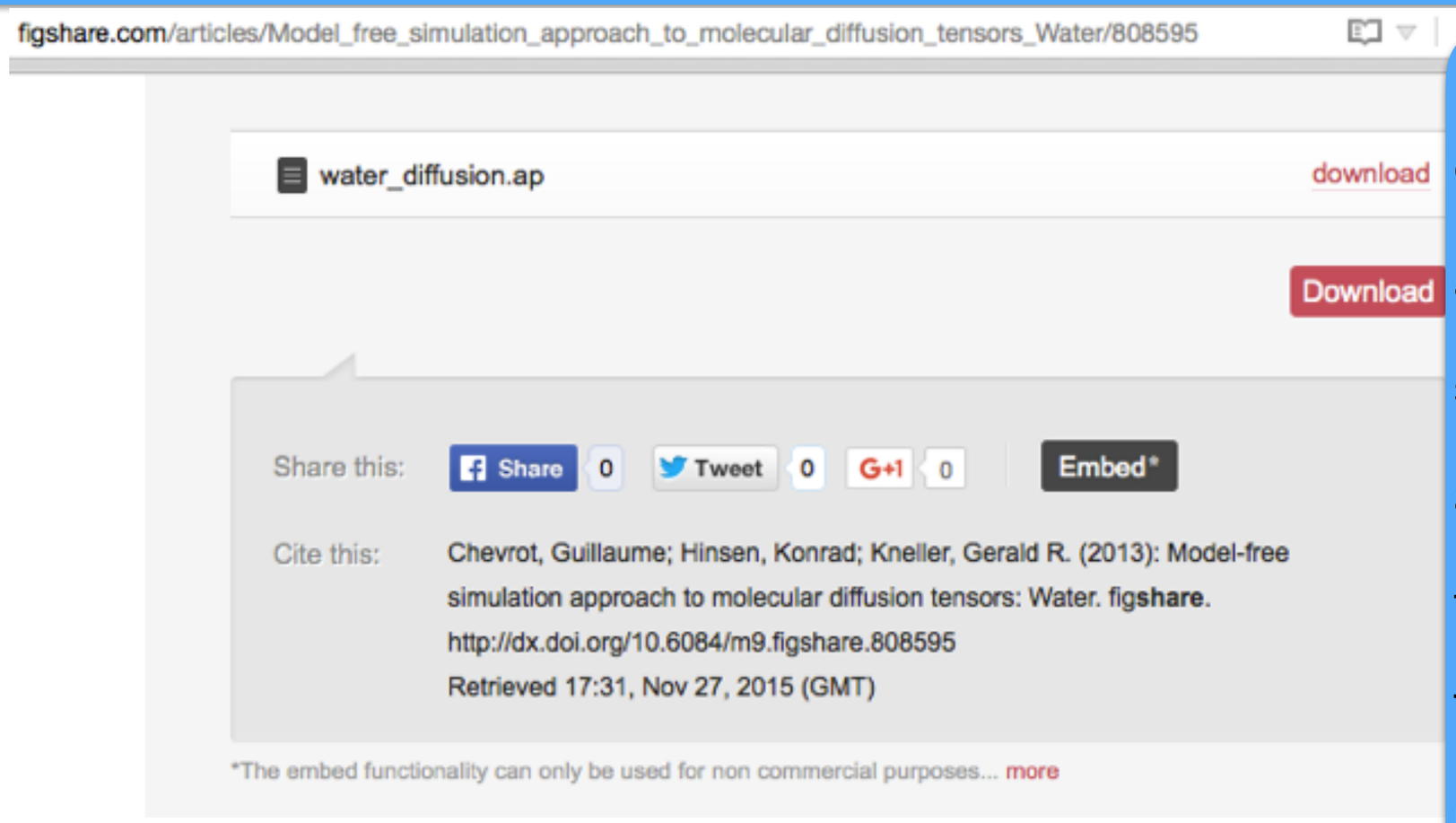
It provides unique and persistent identifiers

for electronic documents on the internet.

The uniqueness of the identifiers is

guaranteed by a central registry.

A concrete example - License



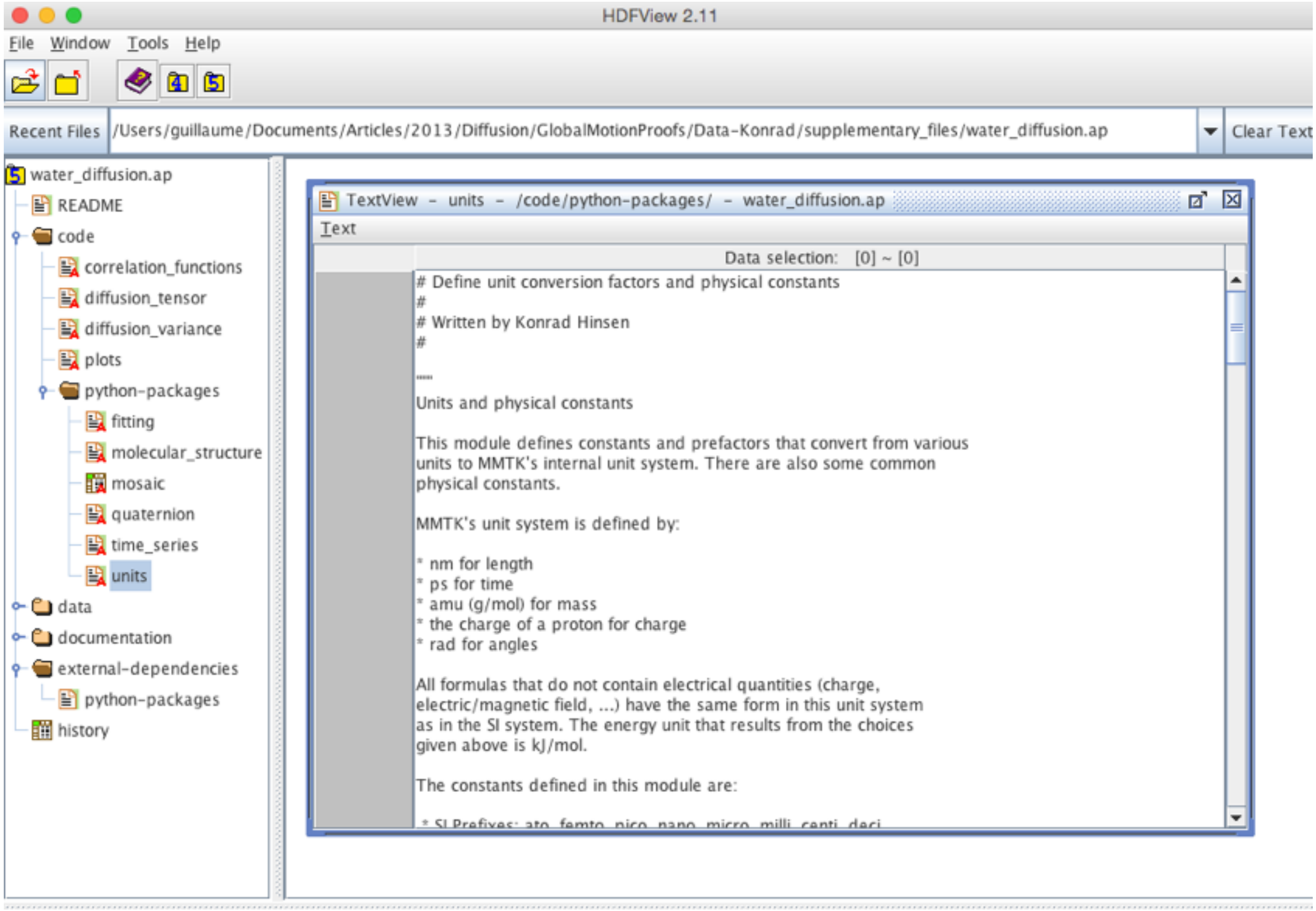
CC-BY: your research is openly available, but requires that others should give you credit, in the form of a citation, should they use or refer to the research object.

This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation.

Creative Commons Licenses

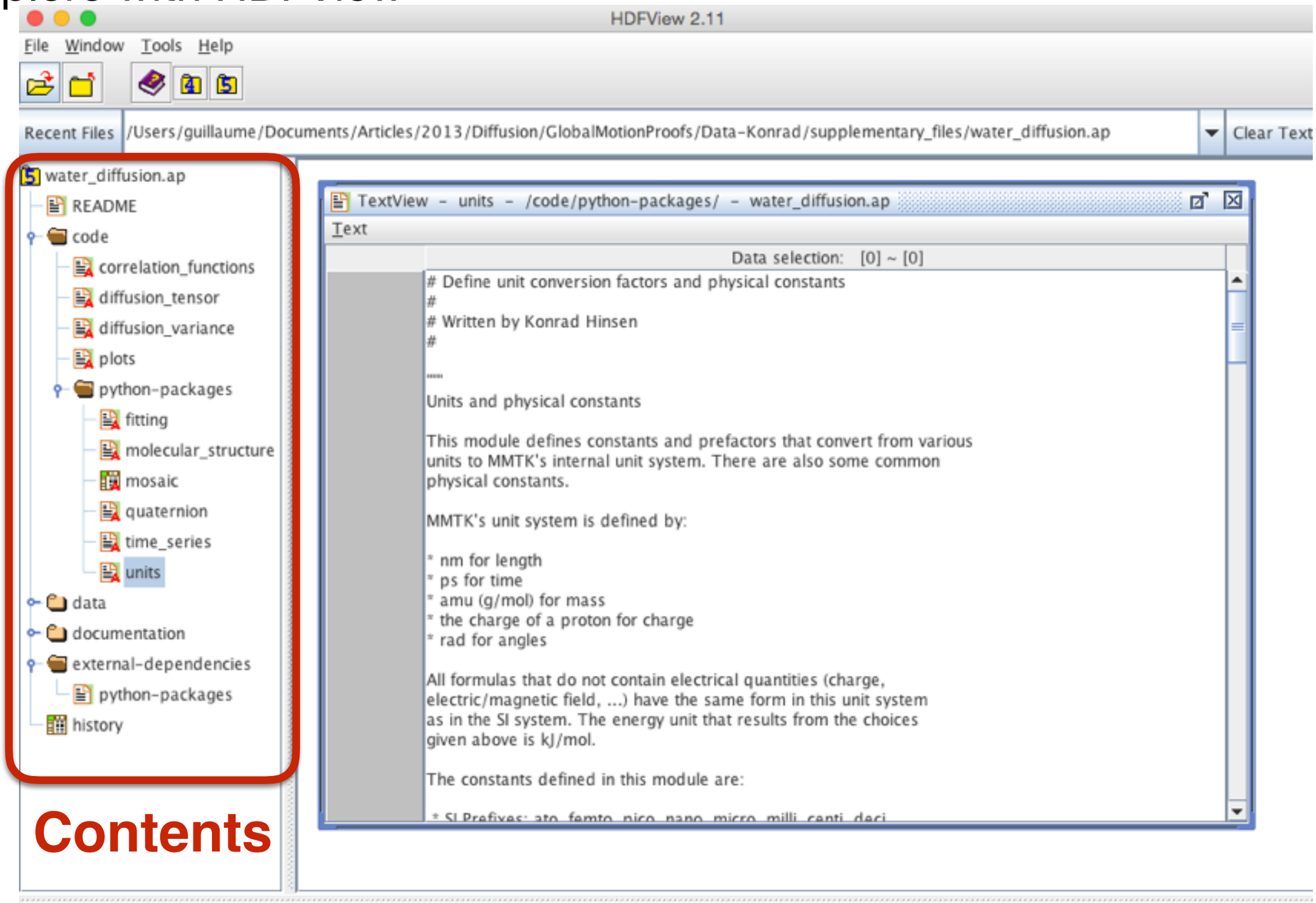
A concrete example - exploration

- Explore with HDFView



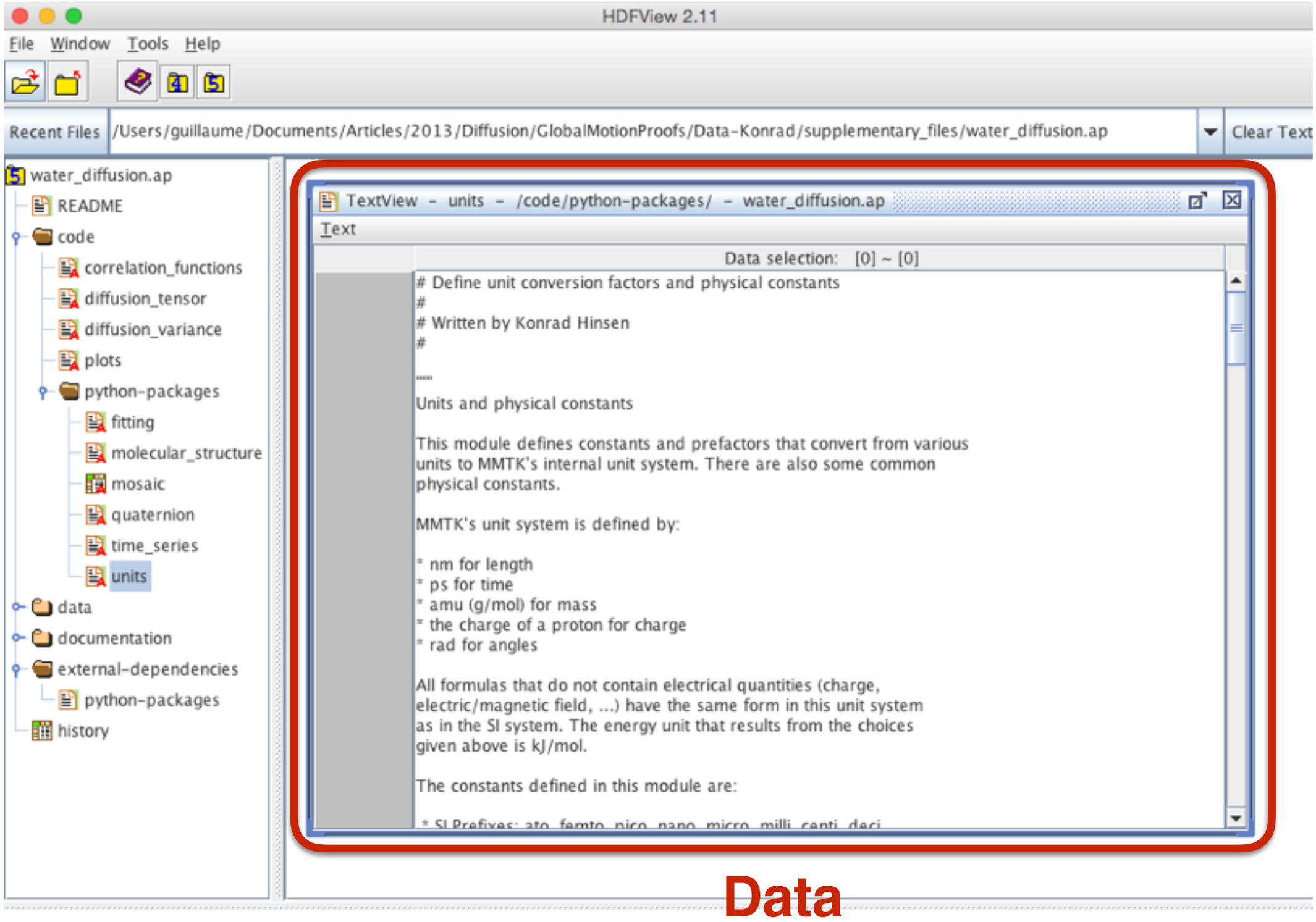
A concrete example - exploration

- Explore with HDFView



A concrete example - exploration

- Explore with HDFView



A concrete example - exploration

- Explore with HDFView

HDFView 2.11

File Window Tools Help

Recent Files /Users/guillaume/Documents/Articles/2013/Diffusion/GlobalMotionProofs/Data-Konrad/supplementary_files/Test/water_diffusion.ap Clear Text

water_diffusion.ap
- README
- code
  - correlation_functions
  - diffusion_tensor
  - diffusion_variance
  - plots
  - python-packages
    - fitting
    - molecular_structure
    - mosaic
    - quaternion
    - time_series
    - units
- data
- documentation
- external_dependencies
- history

history at / [water_diffusion.ap in /Users/guillaume/Documents/Articles/2013/Diffusion/GlobalMotionProofs/Data-Konrad/supplementary_files/Test]

Table

0-base

	opened	closed	platform	hostname	username	activepap...	python_ve...	numpy_ve...	h5py_vers...	hdf5_versi...	matplotlib_version
0	1370247400257	1370247400263	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
1	1370247400844	1370247400865	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
2	1370247401362	1370247401372	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
3	1370247401881	1370247401891	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
4	1370247402387	1370247402396	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
5	1370247402899	1370247402914	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
6	1370247403420	1370247403438	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
7	1370247403934	1370247403938	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
8	1370247404420	1370247404425	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
9	1370247404913	1370247404921	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
10	1370247405414	1370494691493	linux2	isei000.hpc	hinsen	0.1.0	2.7.2	1.6.1	2.1.2	1.8.9	1.1.0
11	1370500745854	1370500746130	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x
12	1370500746675	1370500746681	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x
13	1370500747217	1370500747223	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x
14	1370500747757	1370500747779	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x
15	1370500748315	1370500800627	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x
16	1370500802132	1370500802151	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x
17	1370500802688	1370500803034	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x
18	1370500803586	1370500803602	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x
19	1370500804170	1370500804252	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x
20	1372849295699	1372849295951	darwin	Konrad-H...	hinsen	0.1.0	2.7.3	1.6.2	2.1.3	1.8.11	1.2.x

History

Date

Platforms

User

Packages versions

A concrete example - exploration with AP

- Explore with ActivePapers
  - open a shell window - works with the command line. You can also use a Jupyter notebook.
  - ‘unix command concept’:

    aptool -h or aptool --help

```
 [49] → aptool -h
usage: aptool [-h] [-p PAPER] [--log LOG] [--logfile LOGFILE] [--version]
              {create,ls,rm,dummy,set,group,extract,calclet,importlet,import,run,update,checkin,c
heckout,ln,cp,refs,edit,console,ipython}
              ...

Management of ActivePapers

positional arguments:
  {create,ls,rm,dummy,set,group,extract,calclet,importlet,import,run,update,checkin,checkout,ln,c
p,refs,edit,console,ipython}
                        commands
    create              Create a new ActivePaper
    ls                  Show datasets
    rm                  Remove datasets and everything depending on them
    dummy               Replace datasets by dummies
    set                 Set dataset to the value of a Python expression
```

A concrete example - exploration with AP

aptool ls -l

```
 [51] → aptool ls -l
2013-05-28/10:07:06  calclet    code/correlation_functions
2013-05-23/14:40:23  calclet    code/diffusion_tensor
2013-05-31/08:54:25  calclet    code/diffusion_variance
2013-06-03/18:24:00  calclet    code/plots
2013-05-22/11:26:53  module     code/python-packages/fitting
2013-05-21/11:34:06  module     code/python-packages/molecular_structure
2013-06-03/10:16:42  reference  code/python-packages/mosaic
2013-05-21/11:39:30  module     code/python-packages/quaternion
2013-06-03/10:16:15  module     code/python-packages/time_series
2013-05-21/11:34:06  module     code/python-packages/units
2013-06-06/06:58:09  data       data/averaged/correlation_function_laboratory_frame
2013-06-06/06:58:09  data       data/averaged/correlation_function_molecular_frame
```

Date/time of the last modification

Type of datasets

Hierarchical structure

A concrete example - datasets

8 types of datasets:

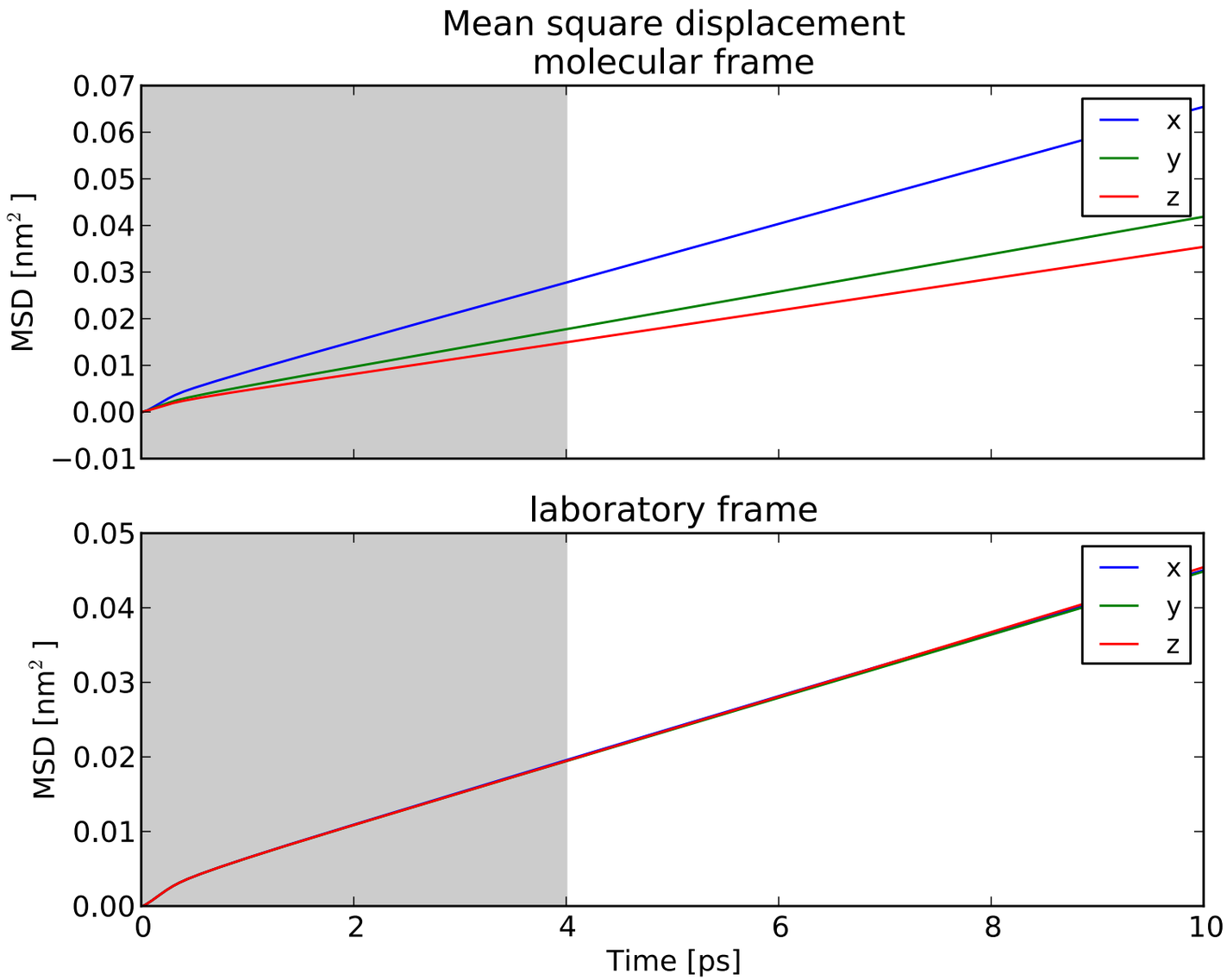
- **module**: a Python source code file to be imported by other Python code.
- **calclet**: Python script with no I/O outside of the ActivePaper.
- **importlet**: Python script without restriction (Internet, data on your computer, Python libraries).
- **reference**: to refer to datasets in other ActivePaper files.
- **data**: HDF5 dataset (array).
- **text**
- **file** (ex: PDF)
- **dummy**: a deleted dataset (to reduce the size of the file, can be restored with a calclet).

A concrete example - documentation

Extracting the documentation / plots:

- aptool checkout documentation

Get a directory containing all the PDF files. Example:



A concrete example - code

Extracting the code:

- aptool checkout code

Get a directory with the code

```
 [82] → ls -lR code/
total 64
-rw-r--r--  1 guillaume  staff  8329 May 28  2013 correlation_functions
-rw-r--r--  1 guillaume  staff  2061 May 23  2013 diffusion_tensor
-rw-r--r--  1 guillaume  staff   758 May 31  2013 diffusion_variance
-rw-r--r--  1 guillaume  staff  9195 Jun  3  2013 plots
drwxr-xr-x  8 guillaume  staff   272 Nov 28 22:31 python-packages

code//python-packages:
total 40
-rw-r--r--  1 guillaume  staff   273 May 22  2013 fitting
-rw-r--r--  1 guillaume  staff   655 May 21  2013 molecular_structure
-rw-r--r--  1 guillaume  staff     0 Nov 28 22:31 mosaic
-rw-r--r--  1 guillaume  staff  2741 May 21  2013 quaternion
-rw-r--r--  1 guillaume  staff  2983 Jun  3  2013 time_series
-rw-r--r--  1 guillaume  staff  3772 May 21  2013 units
```

A concrete example - code

Modifying the code.

Update the ActivePaper with the new code:

aptool checkin code

```
 [101] → aptool ls -l | grep \*
2013-06-06/08:39:38  file      *documentation/averaged/c_rr_diagonals.pdf
2013-06-06/08:39:39  file      *documentation/averaged/c_rr_off_diagonal.pdf
2013-06-06/08:39:36  file      *documentation/averaged/c_tt_diagonal.pdf
2013-06-06/08:39:37  file      *documentation/averaged/c_tt_off_diagonal.pdf
2013-06-06/08:39:40  file      *documentation/averaged/c_vr.pdf
2013-06-06/08:39:59  file      *documentation/averaged/msd_rr_diagonal.pdf
2013-06-06/08:39:59  file      *documentation/averaged/msd_rr_off_diagonal.pdf
2013-06-06/08:39:58  file      *documentation/averaged/msd_tt_diagonal.pdf
2013-06-06/08:39:58  file      *documentation/averaged/msd_tt_off_diagonal.pdf
```

Some files in the documentation are now marked as stale

Then you can update the files: aptool update

```
2015-11-30/17:15:18  file      documentation/averaged/c_rr_diagonals.pdf
2015-11-30/17:15:19  file      documentation/averaged/c_rr_off_diagonal.pdf
2015-11-30/17:15:16  file      documentation/averaged/c_tt_diagonal.pdf
2015-11-30/17:15:17  file      documentation/averaged/c_tt_off_diagonal.pdf
2015-11-30/17:15:20  file      documentation/averaged/c_vr.pdf
2015-11-30/17:15:43  file      documentation/averaged/msd_rr_diagonal.pdf
2015-11-30/17:15:44  file      documentation/averaged/msd_rr_off_diagonal.pdf
2015-11-30/17:15:42  file      documentation/averaged/msd_tt_diagonal.pdf
2015-11-30/17:15:43  file      documentation/averaged/msd_tt_off_diagonal.pdf
```

New dates

The files in the documentation are now updated

A concrete example - parameters

Modifying parameters.

Example: change the range of the MSD plot:

aptool set msd_plot_range ‘array([0.,200.])'

```
 [107] → aptool ls -l | grep \*
2015-11-30/17:15:43  file      *documentation/averaged/msd_rr_diagonal.pdf
2015-11-30/17:15:44  file      *documentation/averaged/msd_rr_off_diagonal.pdf
2015-11-30/17:15:42  file      *documentation/averaged/msd_tt_diagonal.pdf
2015-11-30/17:15:43  file      *documentation/averaged/msd_tt_off_diagonal.pdf
2015-11-30/17:15:46  file      *documentation/averaged/msd_vr.pdf
```

Some files in the documentation are now marked as stale

Then you can update the files: aptool update

```
2015-11-30/17:58:03  file       documentation/averaged/msd_rr_diagonal.pdf
2015-11-30/17:58:03  file       documentation/averaged/msd_rr_off_diagonal.pdf
2015-11-30/17:58:01  file       documentation/averaged/msd_tt_diagonal.pdf
2015-11-30/17:58:02  file       documentation/averaged/msd_tt_off_diagonal.pdf
2015-11-30/17:58:04  file       documentation/averaged/msd_vr.pdf
```

New dates

The files in the documentation are now updated

http://www.activepapers.org/python-edition/tutorial.html

A concrete example - interactive exploration

jupyter ActivePapers Last Checkpoint: 2 minutes ago (autosaved)

File Edit View Insert Cell Kernel Help Python 3

Code Cell Toolbar: None

```
In [2]: %matplotlib notebook
        from activepapers.exploration import ActivePaper
        import matplotlib.pyplot as plt
        import numpy as np

In [3]: ap = ActivePaper("doi:10.6084/m9.figshare.808595")

In [4]: ?ap

In [ ]:
```

Download and open the ActivePaper via its DOI

```
Type:        ActivePaper
String form: <activepapers.exploration.ActivePaper object at 0x103940b00>
File:        ~/anaconda/envs/activepapers/lib/python3.4/site-packages/ActivePapers.Py-0.1.4-py3.4.egg/activepa
pers/exploration.py
Docstring:
Datasets in this file
=====================

The input data to the computations in this file are rigid-body
trajectories for each of the 511 water molecules in the original MD
trajectory. This input data is too big to be provided here.  The
correlation functions and mean-square displacements are first computed
individually for each water atom and then averaged. The
single-molecule functions are also too big to be provided here, exceot
for a subset of five water molecules for demonstration.

All single-molecule data has a numerical suffix. In the following
list, only the first one ("_0") is shown.
```

Explore the README

A concrete example - interactive exploration

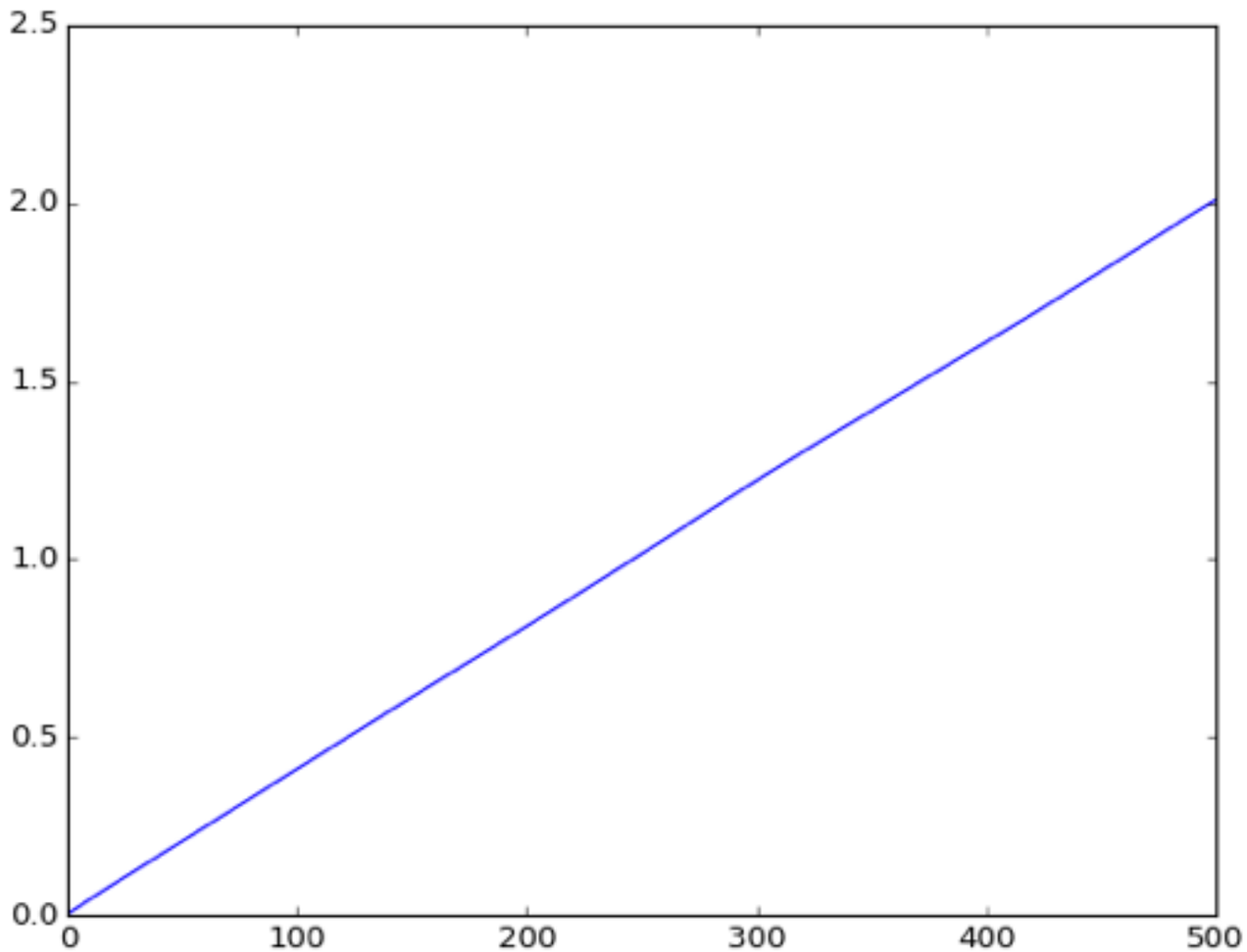
jupyter ActivePapers Last Checkpoint: 3 minutes ago (unsaved changes)

File Edit View Insert Cell Kernel Help Python 3

Code Cell Toolbar: None

```
In [4]: ?ap
```

```
In [5]: msd_time = ap.data['mean_square_displacement_time'][...]
        msd_averaged_lab_xx = ap.data['averaged/mean_square_displacement_laboratory_frame'][0, 0]
        plt.plot(msd_time, msd_averaged_lab_xx)
```



Make a plot from the data inside the ActivePaper

A concrete example - interactive exploration

jupyter ActivePapers Last Checkpoint: 16 hours ago (unsaved changes)

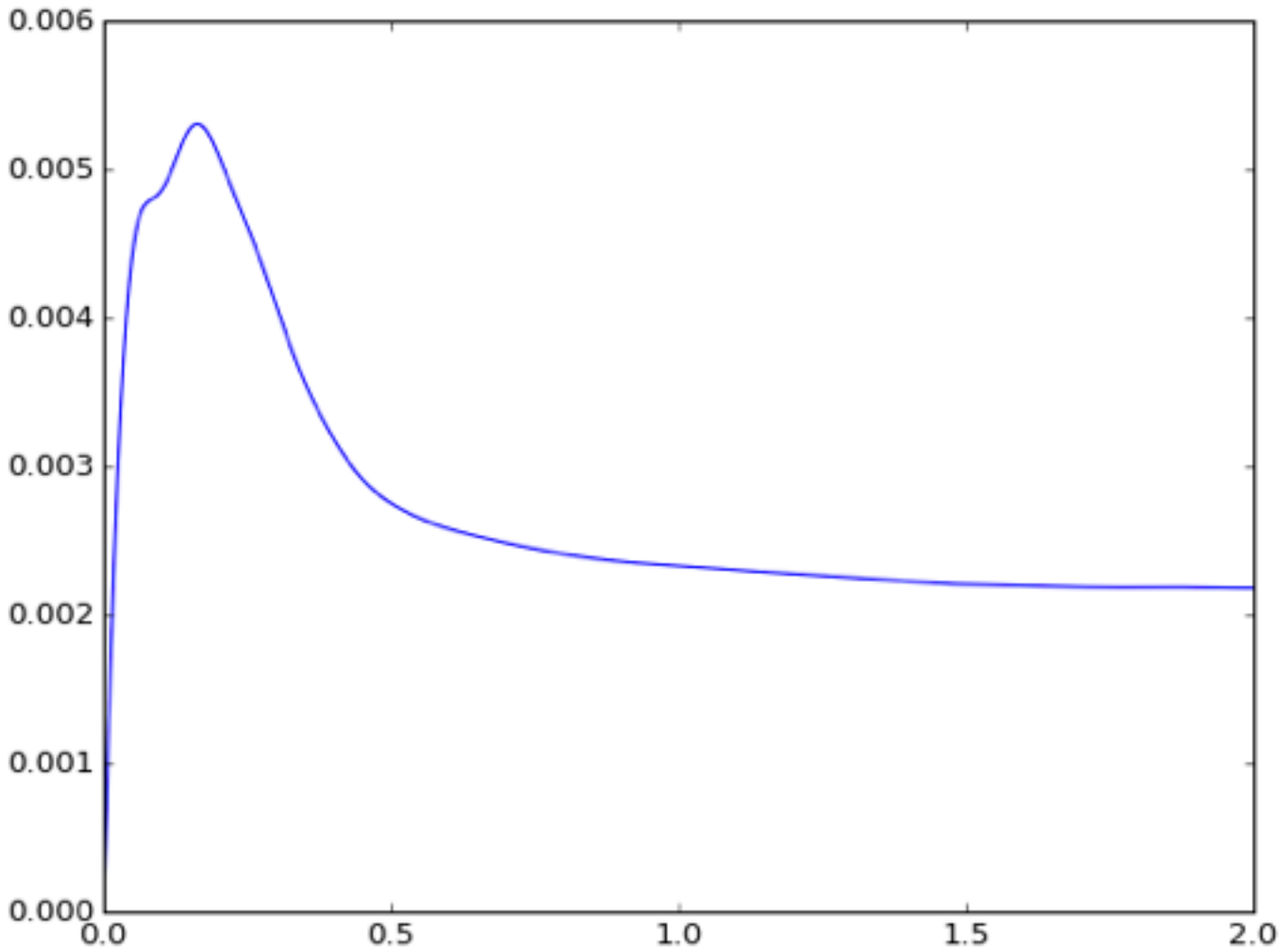
File Edit View Insert Cell Kernel Help | Python 3

Code | Cell Toolbar: None

```
In [6]: import time_series

In [7]: vacf_time = ap.data['correlation_function_time'][:2000]
        vacf_averaged_lab_xx = ap.data['averaged/correlation_function_laboratory_frame'][0, 0, :2000]

In [10]: dt = vacf_time[1]-vacf_time[0]
         plt.plot(vacf_time, time_series.integral(vacf_averaged_lab_xx, dt))
```



Import Python modules stored in the ActivePaper

Interact with the module

Limitations

- Data limit that can be stored online. In the example, we start from the simulations data (>10Gb)

- Python only

- Dependences: Numpy, h5py + external libraries

- How many years your ActivePaper will be stable? (Scientific Python ecosystem has proven to be moderately stable in the past)

Conclusions

- ActivePaper store all your data in a single file

- Python edition is operational - Number of publications available as an ActivePaper: 4

- Next developments:
  - interaction with notebooks
  - other language

- Reproducible research on the web: ActivePaper and Exec&Share
  http://www.univ-orleans.fr/misc-orleans-tours

  http://www.activepapers.org/

  GitHub https://github.com/activepapers/activepapers-python

  twitter @ActivePapers