Open source software and data for fusion energy sciences

Nick Murphy
Center for Astrophysics | Harvard & Smithsonian


This work has been supported by: NSF, NASA.
My background

- Graduate school in astronomy (U. Wisconsin)
  - Studied connections between laboratory & astrophysical plasma physics

- Postdoc and researcher (Center for Astrophysics)
  - Studied solar physics and fundamental plasma science

- Last $3.1 \pm 0.7$ (3σ) years
  - Research software engineering for PlasmaPy
  - Advocating for open & reproducible plasma science

- Most recent pandemic hobby
  - Reading tokamak data management plans 🥳 😫 😱 😞
The fusion energy sciences community does not yet have a culture that supports open sharing of software and data.

But we will!
Topics for the next ~70 minutes

- Data and software environment comparison
  - Solar physics
  - Fusion energy sciences
- Open science
  - Motivation
  - Barriers
- FAIR principles for data stewardship
  - Example: Fair4Fusion
- Open source software
  - Example: PlasmaPy
Case study: solar physics

● Similarities with fusion energy sciences
  ○ Plasma physics is foundational
  ○ Diagnostics (e.g., spectroscopy)
  ○ Comparisons between simulations and reality

● Differences with fusion energy sciences
  ○ Solar observations are more homogeneous
    ■ Images, spectra, and time series
  ○ No experimental control
  ○ “There’s only one Sun.” — D. Coster
The solar physics software environment

- **SolarSoft**
  - Community-developed (unclear licensing)
  - Developed since the 1990s
  - Written in IDL (proprietary language)
  - Monolithic architecture

- **SunPy**
  - Community-developed (open source license)
  - Developed since 2011
  - Written in Python (open source language)
  - Modular architecture
  - Written using modern software engineering best practices
Most observational data sets are openly available

The **Virtual Solar Observatory** (VSO) allows us to:
- Simultaneously search multiple databases
- Download data from multiple observatories

Multiple ways to use VSO
- Web interface
- SolarSoft
- SunPy

Data sets are mostly standardized
Can start downloading solar data in minutes
Consequences for solar physics

- Data sets are widely used and re-used
- Multiple groups can study the same event
- Customary to use data from multiple sources
- Data sets well-suited for machine learning studies
- Well-documented and well-tested community software
- Less software duplication
- Data access not restricted by major institutions
Fusion energy sciences data environment

- Access to most experimental data is restricted
  - Can often request permission
- User agreements often contain restrictions such as:
  - No commercial use (without prior approval)
  - No redistribution of data (without prior approval)
  - Internal approval required for presentations & papers
- Data sets usually not standardized
- Difficult to search archives
Fusion energy sciences software environment

- Dependence on legacy codes
  - Often not open source
  - Different degrees of documentation, testing, and usability

- Software and many data sets bridged together by OMFIT
  - Likely to become open source soon!

- Common license customizations
  - Restrictions on commercial use
  - Limits on redistribution rights

- Difficult to find out which codes do what
Consequences for fusion energy sciences

● Hard to find data
  ○ Cannot simultaneously search across tokamak data archives

● Hard to access data
  ○ Need to request permissions
  ○ Difficult to perform cross-device studies

● Hard to write analysis software for multiple devices
  ○ Need partially met by OMFIT

● Reduced scientific reproducibility
Increasing number of open source software projects

- PlasmaPy
- OMFIT (soon!)
- OMAS
- tofu
- PyMethes
- FIDASIM
- Aurora
- simsopt
- BOUT++
- Gkeyll
- VPIC
- bapsflib
- gptools
- profiletools
- divHretention
- MOOSE
Open science principles

- Open access data
- Open source software
- Open methodology
- Open peer review
- Open access publications
- Open educational resources
Open science principles

- Open access data
- Open source software
- Open methodology
- Open peer review
- Open access publications
- Open educational resources
Why open science?

- Reduce barriers to access
- Broaden research impact
- Improve scientific reproducibility
- Make research more transparent
- Make publicly funded research available to the public
- Maximize use of data
- Allow community review of results
- Invest in our future
Barriers to open science

- Pressure to publish
  - Fear of being scooped
- Time pressure
- Financial pressure
  - Open access publication often costs extra
- Bureaucracy and institutional inertia
  - Decisions might need approval of ITER Council
- Power imbalances
  - Early career scientists more likely to support open science
Barriers to open science

● Toxic culture
  ○ Blatant or subtle acts of racism, sexism, etc.
  ○ Bullying, harassment, and discrimination
  ○ Retaliation

● Equity gaps

● Language barriers
  ○ Scientific information sometimes only available in English

● National security and intellectual property rights
What’s good for science ≠ what’s good for scientists

Good for career
- New and exciting results
- Not “wasting time” making data & software available

Good for science
- Writing software documentation
- Investing time to make data & software available

How can we make what’s good for science also what’s good for scientists?
Open science is an investment

- It takes time and resources to:
  - Make data FAIR
  - Write documentation
  - Write tests
  - Maintain software
  - Learn necessary skills

- Open science will not happen right away…

- …but it is worth the effort!
How do we get closer open science?

● Change our culture
  ○ Ensure psychological safety
  ○ Eliminate equity gaps

● Change our institutions

● Collaborate on technical infrastructure
  ○ Open access data
  ○ Open source software
How do we get closer to open science?

- Change our culture
  - Ensure psychological safety
  - Eliminate equity gaps

- Change our institutions

- Collaborate on technical infrastructure
  - Open access data
  - Open source software
The FAIR principles for data stewardship

- Findability
- Accessibility
- Interoperability
- Reusability
Findability

- Before being able to reuse a data set...we have to find it!
- Assign digital resources a **persistent identifier**
  - Digital Object Identifiers (DOIs)
- Describe data sets with **rich metadata**
- Index data sets in a **searchable resource**
  - Zenodo ← online repository operated by CERN
  - Virtual Solar Observatory
Accessibility

- Can access a data set using its persistent identifier
- Can access metadata even if original data set is gone
- *Accessible* is not the same as *open*
  - Authentication and authorization sometimes required
Data should be openly available when possible

- Data sets from most fusion devices are *not* openly available
- Should each device create its own online open archive?
  - Possible duplication of effort
  - Potential limitations on cross-device searchability
- Could we create a community-wide portal to access open fusion/plasma data sets?
  - Improve findability, accessibility, & reproducibility
  - Enable cross-device studies
  - Allow wide re-use of data
Some data from MAST is now open

UKAEA Open Data

MAST Data

These pages provide an access point to publicly funded MAST research data. By selecting a Program or Objective, and then a specific experiment number, you can request the related underlying shot data if it is available for release.

MAST ≡ Mega Ampere Spherical Tokamak
Some data from MAST is now open

<table>
<thead>
<tr>
<th>UKAEA Open Data</th>
<th>Published Data</th>
<th>MAST Data</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>amh Analysed</td>
<td>Halo current Measurements (HAL), P2/P3 Halo Current Measurements (HALO)</td>
<td>amh27027.nc</td>
<td>netcdf4/hdf5</td>
</tr>
<tr>
<td>amm Analysed</td>
<td>Output from EFIT's wall model: calculated induced currents in toroidal vessel elements for input to EFIT</td>
<td>amm27027.nc</td>
<td>netcdf4/hdf5</td>
</tr>
<tr>
<td>ane Analysed</td>
<td>CO2 Interferometry</td>
<td>ane27027.nc</td>
<td>netcdf4/hdf5</td>
</tr>
<tr>
<td>anu Analysed</td>
<td>Neutron measurement</td>
<td>anu27027.nc</td>
<td>netcdf4/hdf5</td>
</tr>
<tr>
<td>asb Analysed</td>
<td>Spectroscopy CII, OII</td>
<td>asb0270.27</td>
<td>IDA3</td>
</tr>
</tbody>
</table>
Interoperability

● Fusion facilities often have their own way of storing and organizing data
  ○ Hard to perform cross-device studies
  ○ Hard to develop shared software

● Approach #1: develop software to serve as an interface
  ○ OMFIT bridges different data types and software packages

● Approach #2: adopt shared standards for data
Why do we need data interoperability?

- Suppose we are doing experiments at two facilities
  - Basic Plasma Science Facility (BaPSF)
  - Wisconsin Plasma Physics Laboratory (WiPPL)
- We’re studying the same physical process…
- …but data from BaPSF & WiPPL are structured differently!
- We need to write separate software to perform the same analysis
- A common data model would enable shared software and promote cross-device collaborations
  - Examples: IMAS, Plasma-MDS, OpenPMD, SPASE, MetaSat
Plasma science needs open metadata standards

- A **metadata standard** describes an agreed-upon way to structure and understand data
- A meaning is assigned to each variable name
  - Reduces ambiguity
- Greatly improves interoperability
  - Allows different groups to use and interpret data
- **Metadata crosswalks** convert between different standards
Why should metadata standards be open?

- Some standards like IMAS (for ITER) are not open
- Open standards allow for wider adoption
  - PDF, Blu-Ray, etc.
- Restricting access to metadata standards limits adoption
  - May lead to creation of competing standards
Reusability

- Describe data with sufficient **metadata**
- Metadata meet **community standards**
- Data sets have a clear **license**
- Data sets include detailed **provenance**
  - Where did the data come from?
Benefits of reusable data

● Broaden access to research data
● Maximize knowledge gained from data sets
● Improve scientific reproducibility
● Enable machine learning studies
Objective: Make European-funded fusion data more widely available & FAIR

- Raise awareness of open data within the fusion program
- Develop tools needed for an open data approach
- Lay foundations for an open data policy

Subset of tasks

- Outreach to community
- Define use cases
- Create blueprint architecture for open fusion data
- Build data foundation for open access
- Develop open data demonstrators

https://www.fair4fusion.eu/
Fair4Fusion is laying the groundwork to improve the data environment for fusion
Use cases from F4F show need for open access

- As a member of the general public, I would like to know how many shots per day and per year are performed by each of the experiments.

- As a researcher, I want to locate all H-mode shots that had a flat-top phase of longer than 0.5 seconds, across a selection of devices.

- As a data provider, I want to ensure the appropriate availability of my data without breaking the law (e.g. GDPR).

Adapted from D. Coster et al., doi: 10.5281/zenodo.4337222, CC BY 4.0
Efforts to make fusion data FAIR need to continue

- The Fair4Fusion project ends later this year
  - All deliverables to be completed on schedule

- Lesson: *It’s very difficult to change an established community with “working” practices...but the astronomical and meteorological communities show it can be done.*

- Significant potential for international collaboration
  - As with International Heliophysics Data Environment Alliance
Open Source Definition

- Software is **open source** if anyone is free to use, modify, and/or redistribute it
  - Including source code
- An open source license does not discriminate against persons, groups, or fields of endeavor
  - No restrictions on commercial use
- The Open Source Initiative maintains a list of approved licenses
  - Customizing licenses causes problems
Two categories of licenses

- **Permissive** licenses have few restrictions
  - Examples: MIT and BSD 2-clause license

- **Copyleft** licenses require derived works to be released under the same license
  - Example: GNU General Public License version 3 (GPLv3)

- Two licenses are compatible if they allow both programs to be combined into a single program
  - Permissive licenses maximize license compatibility
Pain points with scientific software

- Lack of user-friendliness
- Difficult to compile & install
- Inadequate documentation
- Unreadable code
- Cryptic error messages
- Licensing issues
- Packages not written to work with each other
- Unvalidated code
Consequences of pain points

- Beginning research is hard
- Collaboration is difficult
- Duplication of functionality
- Research is less reproducible
- Research can be frustrating
How can we address these pain points?

- Make our software open source
- Write readable, usable, & maintainable code
- Use a high-level language, where appropriate
- Prioritize documentation
- Create an automated test suite
- Develop code as a community
- Build a shared software framework...

A software ecosystem!
What is PlasmaPy?

Mission

To grow an open source software ecosystem for plasma research & education
Many ways to be part of the community

- Come to PlasmaPy’s...
  - Community meeting (Tuesdays at 18 UTC)
  - Office hours (Thursdays at 18 UTC)
- Join our Element chat
- Request new features on GitHub
- Organize community events
- Contribute!
Current & planned PlasmaPy subpackages

**plasmapy.particles**
- Object-oriented representations of ions, electrons, and fundamental particles

**plasmapy.formulary**
- Commonly needed formulae for plasma parameters and transport coefficients

**plasmapy.simulation**
- To include building blocks of plasma simulations and an improved particle tracker
Current & planned PlasmaPy subpackages

**plasmapy.analysis**
- Analysis techniques for data from simulations, experiments, and observations

**plasmapy.diagnostics**
- For representations of plasma diagnostics such as Langmuir and magnetic flux probes, as well as synthetic diagnostics

**plasmapy.dispersion**
- To contain dispersion relation solvers for plasma waves
Current & planned PlasmaPy subpackages

**plasmapy.plasma**
- Base classes to represent different plasmas

**plasmapy.utils**
- Helpful tools for the rest of the package

**plasmapy.addons**
- Entry point for affiliated packages to be put in PlasmaPy namespace
Biggest challenges for PlasmaPy

● Changing the culture
  ○ Community-wide software development

● Building community
  ○ Supporting new contributors

● Lack of a community-wide portal for experimental data

● Lack of open metadata standards 😳
Plasma Hack Week 2021 (28 June – 2 July)

● Mix of a summer school and a hackathon

● Tutorials for research software engineering
  ○ How to contribute to open source projects
  ○ Writing clean scientific code
  ○ Testing scientific software

● Tutorials for different plasma packages
  ○ PlasmaPy, OMFIT, BOUT++, Gkeyll, etc.

● Videos to be posted online soon

● Expected to be annual event
The nascent field of research software engineering

- Research software engineers (RSEs) include
  - Researchers who spend most of their time programming
  - Software engineers developing scientific software
  - Everyone in between

- The term “research software engineer” was coined in ~2012

- Problems
  - Unclear career paths for RSEs
  - Insufficient training for scientists to become RSEs

- University courses on research software engineering?
Building a healthy and innovative workforce

Developed by Dr. Kimberly Young-McLear, U.S. Coast Guard Academy [1]
Psychological safety is necessary for open science

● Members of the plasma community should be able to:
  ○ Be their authentic selves;
  ○ Share their perspectives and make mistakes;
  ○ Without fear of bullying, retribution, or discrimination

● Psychological safety is foundational for diversity, equity, and inclusion

● Diversity, equity, and inclusion are foundational for open science

Codes of conduct

- All collaborative software projects need a code of conduct
  - Describe unacceptable behaviors (e.g., harassment)
  - Promote positive behaviors (e.g., demonstrating empathy)
- Each code of conduct should have an enforcement policy
- Example: Contributor Covenant Code of Conduct
Summary

- Fusion energy & plasma sciences are becoming more open
- Open science requires cultural change, institutional change, and technical infrastructure
- Scientific data should be findable, accessible, interoperable, and reusable
- Work has begun on an open source software ecosystem for plasma research and education