A SOFTWARE SEARCH INTERFACE FOR HELIOPHYSICS & SPACE PHYSICS

JULIE BARNUM[1,2], REBECCA RINGUETTE[3, 4], SHAWN POLSON[1,2], CATHERINE BYRD[4,5], AND JOE RENAUD[6,7,8]

[1]LABORATORY FOR ATMOSPHERIC AND SPACE PHYSICS
[2]UNIVERSITY OF COLORADO BOULDER
[3]ADNET SYSTEMS, INC
[4]HELIOPHYSICS DIGITAL RESOURCE LIBRARY AT NASA GSFC
[5]UNIVERSITY OF MARYLAND BALTIMORE COUNTY
[6]NASA GSFC
[7]UNIVERSITY OF MARYLAND, COLLEGE PARK
[8]SELLER'S EXOPLANET ENVIRONMENTS COLLABORATION

DASH 2024 MEETING, 15 OCTOBER 2024

HELIOPHYSICS SOFTWARE SEARCH INTERFACE (HSSI)

Why a software search interface is needed for Heliophysics

- Generic search interfaces (e.g., GitHub) do not provide science-specific filters to improve search results
- ADS's interface links to related software based on software citations, which are still mainly absent in Heliophysics publications
- No current resource adequately helps with software *discoverability*

Why what we already have in Heliophysics isn't sufficient

- PyHC search interface only includes Python centric open-source software
- Lacks software in many categories (e.g., AI/ML, models/simulations)
- Need to align metadata structure with Heliophysics' needs and international standards
- Lack of links to other important resources (e.g., related software, datasets, etc.)

WHY ISN'T SOFTWARE.NASA.GOV ENOUGH?

- Existing science software *funded by NASA* is not covered
  - AI/ML
    - general analysis
  - post processing
- Not linked to ADS/SciX
- Not linked to datasets
- Lack of representation of Heliophysics-specific software
- Lack of community awareness

From Kim Manafra's presentation at the 2024 Software for the NASA SMD Workshop

HELIOPHYSICS SOFTWARE SEARCH INTERFACE (HSSI)

- Answering the Open Science call in Heliophysics to:
  - Unify software metadata with CodeMeta recommendations
  - Improve science software search capabilities
  - Simplify citation of software in other resources
- One landing page to search for science software in Heliophysics
  - Start with software packages contained within PyHC
  - Expand out to Heliophysics code in other programming languages
  - Work to include software in other categories (e.g. AI/ML)
- Creating interoperable software metadata
  - Implement a new software metadata structure for Helio software registration
  - Build a registry of software metadata in Heliophysics with a RestAPI
- Source input from community meetings and advertising campaigns

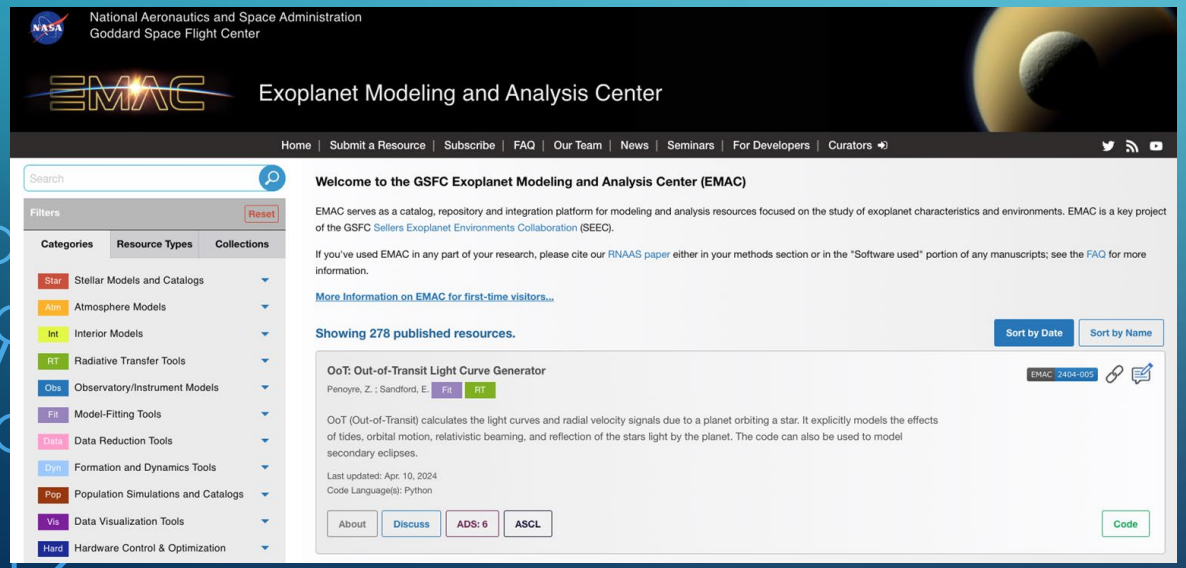
HSSI TEAM

- Core Team
  - Julie Barnum (PI, PyHC Lead), Shawn Polson (Co-I, PyHC Tech Lead), Undergraduate Developer (CU Boulder, HSSI creator), Jenny Knuth (Usability Testing Expert)
- Metadata Experts
  - Catherine Byrd (Digital Librarian), Rebecca Ringuette (Open Science Expert)
- EMAC Experts
  - Joe Renaud (EMAC Lead, providing lessons learned/metadata help), Eric Lopez (EMAC Lead, providing lessons learned), Mike Moore (EMAC Backend developer), Dylan Cristy (EMAC Frontend developer)
- Other Experts
  - Brian Thomas (NASA infrastructure leader at HDRL, providing guidance/advice for effort)

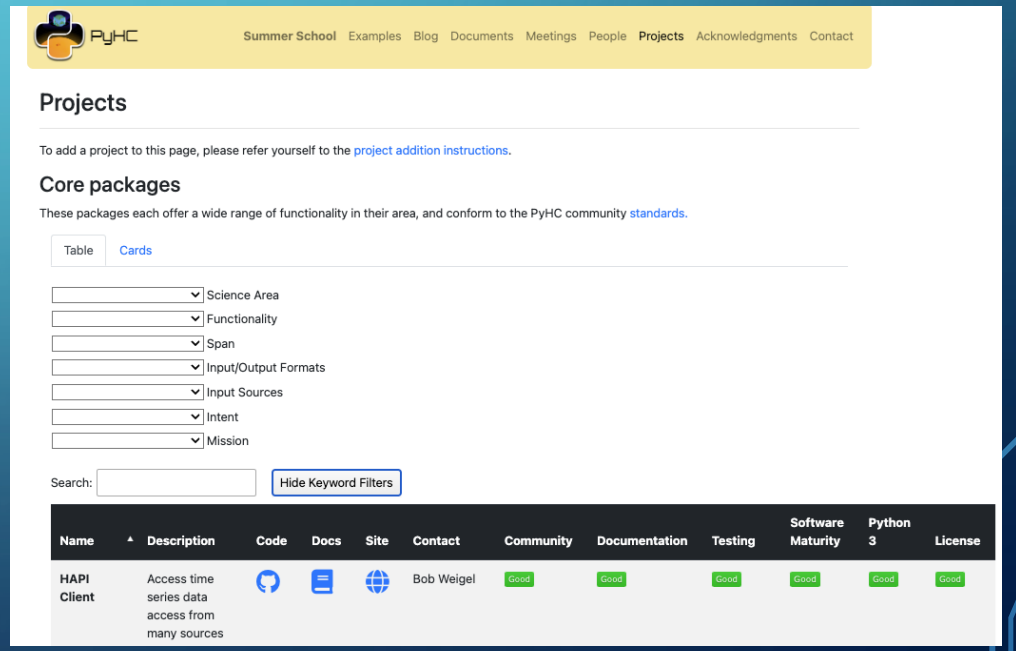
SOFTWARE SEARCH DESIGN

Build upon the designs of:

- NASA Goddard's Exoplanet Modeling and Analysis Center (EMAC)
- PyHC website
- ADS (soon, SciX)
- software.nasa.gov
- Zenodo



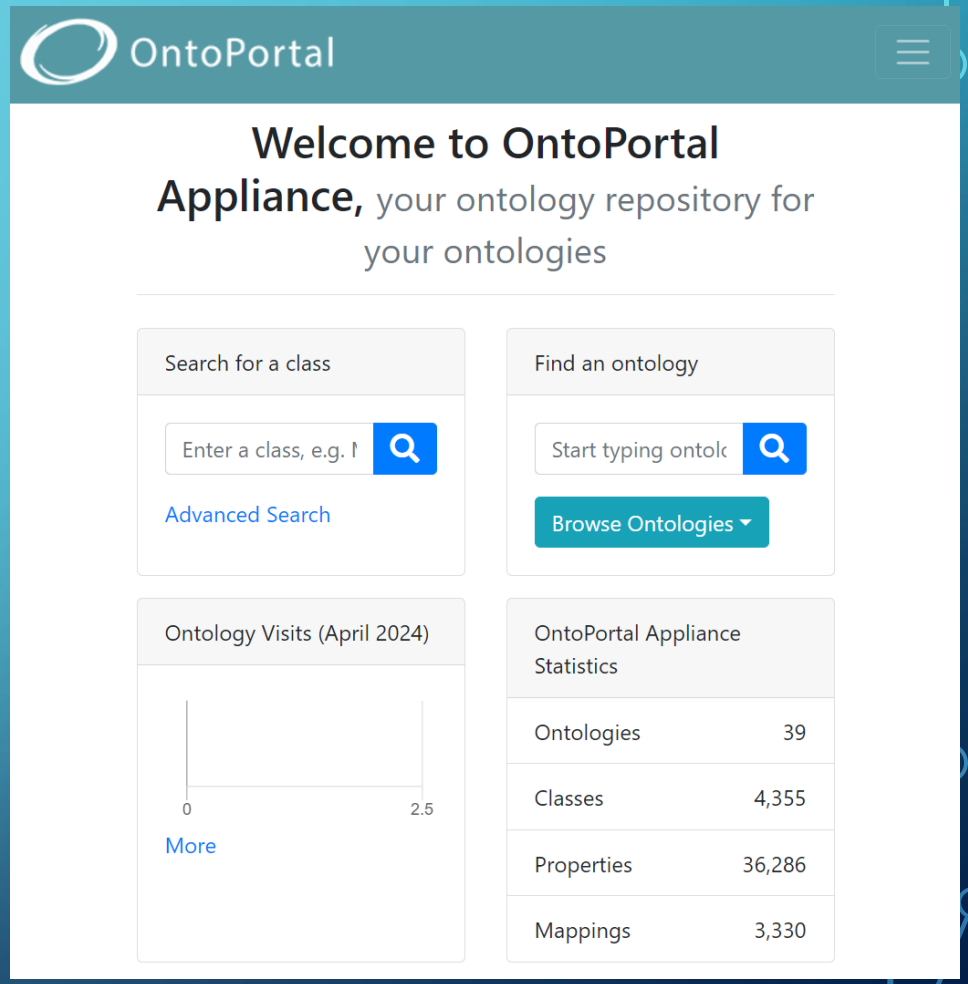
https://emac.gsfc.nasa.gov/



https://pyhc.org/

SOFTWARE METADATA

- Metadata registry to be hosted on GitHub with a RestAPI, routinely archived with Zenodo
- Metadata model to be interoperable via OntoPortal
- Begin with PyHC software metadata structure
- Study international recommendations:
  - CodeMeta
  - DataCite
  - Schema.org / Science-on-schema.org
  - RDA FAIR 4 Research Software
- Initial study on needed controlled vocabularies



http://voparis-ontoportal-dev.obspm.fr

SOFTWARE METADATA CURRENT STATUS

- Initial term prioritization based on EMAC and PyHC search interfaces and experiences
- Terms are now mapped to CodeMeta, DataCite, and schema.org
- Controlled vocabularies identified:
  - Software functionality categories
  - Helio missions/observatories
  - Operating system/architecture
- M/R/O drafted, motivated by the expected layout of the interface and Open Science practices (including FAIR)

Function	Concept	Codemeta property	DataCite name/structure	Schema.org name/structure	Mandatory(M), Recommended (R), or Optional (O)
button	Programming language	programmingLanguage	Format	programmingLanguage	R
display, text search	Release/Publish date	datePublished	publicationYear	datePublished	R
display, text search	Authors / Credits	author	creators	author	M
display, text search	publisher	publisher	publisher	publisher	R
display, text search	Related Missions / Instruments	relatedLink	subject	relatedLink	O
display, text search	software Name	name	title	name	M
display, text search	Version	version	version	softwareVersion	R
display, text search	PID	identifier	DOI	identifier	R
display, text search	description (used in side-by-side comparisons)	description	description	description	M
display, text s…	concise description	description	description	description	O
button	Type (database, model, simulation, etc), software category, functionality	applicationCategory	subject	applicationCategory	R
display, text s…	documentation link	buildInstructions	relatedItem	buildInstructions	R
display, text s…	inputs	supportingData	relatedItem	supportingData	O
button	input/output file format ONLY for convention file formats (CDF, netCDF, FITS, HDF5, ASCII, csv, cloud-optimized, other)	keywords	subject	keyword	R
display, text search	Related publications	relatedLink	relatedItem	relatedLink	O

UPCOMING PYHC SESSIONS AND AGU CONTRIBUTIONS

- PyHC hybrid fall 2024 meeting metadata working sessions
  - Led by Joe Renaud (EMAC), Catherine Byrd (HDRL), and Rebecca Ringuette (HDRL).
  - Sourcing feedback from the PyHC on specific metadata keywords
  - Registration (in-person by Oct 25th; virtual suggested Nov 8th)
- Broader AGU contributions
  - "The metadata structure for the Heliophysics Software Search Interface" (Catherine Byrd in PyHC poster session SM43C)
  - "The Python in Heliophysics Community: a "north star" for open-source software communities" (Julie Barnum in Open Science Success Stories poster session IN053)

SUMMARY

- A Heliophysics-specific software search interface will significantly improve FAIR for research software in Heliophysics
- A NASA ROSES HPOSS proposal has been awarded to fund this effort
- Initial work between PyHC, EMAC, and HDRL are progressing
- The metadata framework is now ready for more detailed efforts
- Seeking collaborations with the international science software metadata community
- Dedicated sessions at the PyHC fall meeting and more generally at AGU
- Logo to come?

BACK UP SLIDES

WHY NOT SPASE?

- SPASE is not compliant/interoperable with the international best practices for software metadata embodied in the CodeMeta recommendations
  - Important to support efforts to create an international software search capability that stretches beyond science
- SPASE does not support (some examples):
  - External FAIR vocabularies
  - SPDX licenses (yet)
  - Related products* more general than dataset documentation
- SPASE is meant for datasets not software (not aligned with int'l standards)
- SPASE is denormalized – has multiple fields for the same information
- Known errors with SPASE mapping to schema.org

* Related products = links to related software, publications, and datasets, whether used in creation of the software (e.g. provenance) or items that cite the software at a later point. We're prioritizing the mapping to schema.org because the CodeMeta terms are aligned most closely with schema.org.