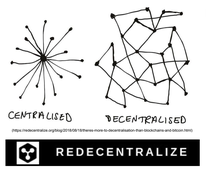
Embracing the (re)decentralized web for sustainable research collaboration cyberinfrastructure

T. Andrew Manning
National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign, IL, USA

ILLINOIS
NCSA | National Center for Supercomputing Applications

What is the (re)decentralized web?

These quotes from the **redecentralize.org** website summarize the vision pretty well:



> *We've had enough of digital monopolies and surveillance capitalism. We want a world that works for everyone, just like the original intention of the web and net.*
>
> *We seek a world of open platforms and protocols with real choices of applications and services for people. We care about privacy, transparency and autonomy. Our organisations and tools should fundamentally be accountable and resilient.*
>
> *We strive for an ecosystem of interoperable products, letting people choose their software and keep their data themselves.*

People around the world have been collaboratively developing decentralized protocols and software applications outside the scientific community for many reasons. Some include (in no particular order):

- Financial freedom and efficient systems for value exchange (cryptocurrencies, blockchain)
- Messaging and teleconferencing (**email**, Matrix, XMPP/Jabber, IRC)
- Data sharing and storage (IPFS, Filecoin, Hypercore/Dat, Storj, Arweave, MaidSafe, WebTorrent)
- Social media (Mastodon, Hubzilla, PeerTube, ActivityPub)

So many more out there!

Decentralization in this context encompasses both

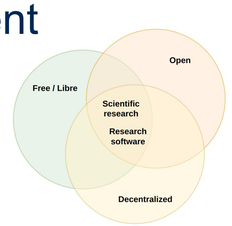
- an **ideological perspective** as well as
- a **technical approach**

to achieving capabilities that may not be readily available from the predominantly commercial and centralized solutions.

By calling a tool or service "decentralized", we typically mean that it involves a **network of nodes that are connect in federated or peer-to-peer** methods, or that it does not depend on a central, uncontrolled resource.

Science is aligned with the web redecentralization movement

- Science is **inherently decentralized**
  - Anyone can advance science
  - No human authority that can control the nature of what is learned
- Scientific results must be **reproducible**
  - Methods and results must be **open** for others to review
  - Others must have the **freedom** to actually reproduce those results
  - You must not only have open source code, you must have a license that empowers others to modify and execute that code and then share your findings with others.
- Science **thrives on inclusivity**
  - Countries of birth and ethnic identities of scientists are irrelevant to the truths scientists discover
  - Artificial limits in software can limit scientific participation



Shouldn't the technology that enables science reflect the way science actually works?

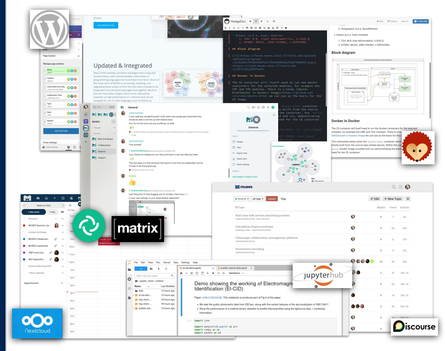
Driving principles

- Researchers should have as much control as possible over their data, services, and access control. **Responsibility should be distributed** among the collaborators effectively, never disproportionately held by one or two top-level sysadmins.
- **Free and open source software is the preferred solution** whenever possible for both ideological and practical reasons. Vendor lock-in, license/subscription price increases, and abrupt changes to terms of service in commercial solutions are a real problem for academic research.
- Embracing decentralization means that we **self-host data and web services** when advantageous, and we **select applications that prioritize data ownership** (which implies a commitment to privacy and portability).
- A collaboration's CI should be **as portable as possible for the sustainability of their project** as funding comes and goes and the CI may migrate to a new host institution.
- There is **plenty of innovation in the creative integration of existing open-source applications**, and these innovations can enable better research outcomes. We should leverage existing work whenever feasible.

Collaborative tools and services

The collaborations that benefit most from our CI typically involve

- researchers from **multiple international institutions** with dozens to hundreds of scientists including academic research professors, postdocs, graduate and undergraduate students, and staff scientists
- who find that their new projects require *data management* as well as *identity and access management* (IAM) services **independent of the participating institutions.**



Researchers need online platforms and tools to collaborate effectively. Our framework aims to provide a suite of **tightly-coupled, well-supported, collaborative tools** and services that are based upon **decentralized**, **free and open source software**:

- **File storage and sharing** (Nextcloud)
- **Organized and persistent asynchronous discussions, a.k.a. "forum"** (Discourse)
- **Real-time messaging** (Matrix)
- **Audio/video conferencing** (Matrix, Nextcloud)
- **Collaborative document editing** (HedgeDoc, Nextcloud)
- **Website content management system** (WordPress)
- **Interactive development environments** (JupyterHub)
- **Backup and bulk file transfer** (Nextcloud, Syncthing)

Single-sign-on for everything, with roles defined in Keycloak specifying permissions across all services

Identity and Access Management

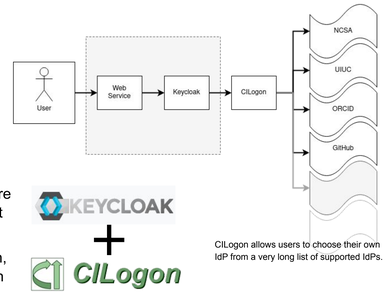
We use a Keycloak server for authentication, which uses CILogon as the identity provider (IdP).

The beauty of this system is that we can register a single client application with CILogon for use by Keycloak.

Then, we can create as many client applications in Keycloak as we need for our various services.

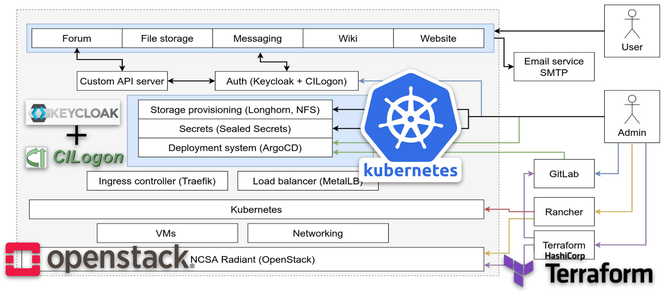
All of our services only depend on Keycloak; they are **agnostic to the identity provider -** i.e. can support multiple surveys/groups/universities

Can link identities e.g. ORCiD + University login, so if a postdoc moves to new position, transition on CI framework is seamless



Roles + access lists define permissions in one place across all services!

Cyberinfrastructure stack



Full-stack, industry standard components

- Terraform to provision hardware (real or virtual)
- OpenStack-compatible
- Kubernetes-based

Flexible, modular deployment

- Each layer of the stack is mostly independent of the others
- CI can **adapt** quickly and **migrate** easily

Project showcase

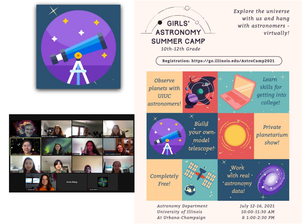
The flexibility of our system is demonstrated in the ways that it has been adapted to the different needs of independent projects, including:

- DESaccess science gateway and HPC backend for the Dark Energy Survey
- Adapting the MMA alert processing CI from ANTARES (NOIRLab) for hosting and development at NCSA
- Science gateway and HPC backend for the prototype in development for SPT-3G, with an eye toward CMB-S4
- Collaboration services and system architecture design for the NSF-funded MUSES collaboration

Education and outreach activities including Girls' Astronomy Summer Camp 2021, 2022, 2023

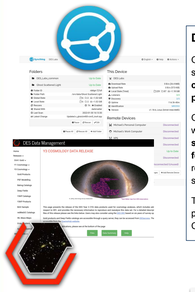
Bootstrapped a new Kubernetes cluster on the NCSA Radiant system and deployed a JupyterHub environment to support Girls' Astronomy Summer Camp at UIUC in few hours.

Dozens of teens concurrently launched their own JupyterLab servers - used to learn and experiment with astrophysics programming in Jupyter notebooks.



This kind of "ephemeral CI" for educational outreach activities/hack weeks/testing is one of many potentially impactful use cases for the system.

Integrations

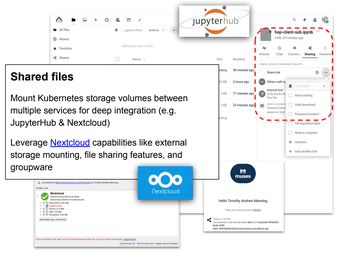


Data transfer and synchronization

Online platforms like Globus offer a wealth of services related to data transfer, but there are **large costs associated with these services that can be prohibitive** for smaller organizations.

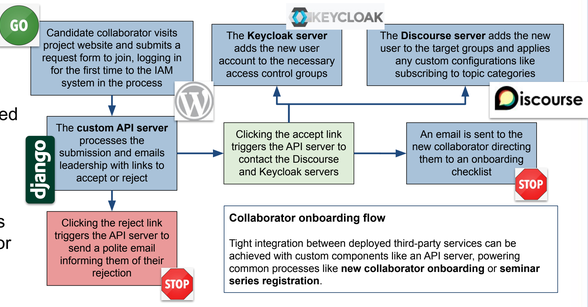
Our framework could integrate Globus services as well, but **we support deploying peer-to-peer solutions like Synching in concert with federated services like Nextcloud** to provide robust, cutting-edge and completely free and open source solutions for arbitrary data transfers.

The **DES project has utilized these tools** to prepare data releases, including the recent Y3 Cosmology public data release.



Spend time developing integrations to create powerful solutions tailored to your researchers' needs.

These are just a few examples of integrations we have implemented for our research collaborations.



Connect with us

We'd love for you to stop by our **community forum** or our **Matrix space** and say hello. Scan the QRcode or visit

`https://decentci.ncsa.illinois.edu/#connect`

Acknowledgements

Many people from different projects were directly or indirectly involved in what is presented here. Although listing them all is infeasible, I would like to thank Rob Kooper at NCSA for his material support and mentorship.

DES Data Management System is supported by the National Science Foundation under Grant NSF AST 07-15036, NSF AST 08-13543, and NSF AST 15-36171. The MUSES project is supported by National Science Foundation under Cooperative Agreement OAC-2103680.