



# **Princeton Data Commons: Building an Ecosystem to Support Evolving Research Data Needs**

Kate Lynch and Hector Correa, Princeton University




## Introductions



Princeton University Library



Princeton **Research**  
**Data Service**



# How it started: DataSpace

Customized legacy instance of  
DSpace 5.x, inherited by Library IT

[DataSpace](#) [About](#) [Contact](#) [Help](#)

[Search](#) [Explore](#) [Login](#)

## DataSpace at Princeton University

Princeton's Research Data Repository is becoming the Princeton Data Commons! As we transition, you may not see some of the collections you were once able to submit to in DataSpace, and that means Princeton Data Commons is ready for you. If you are intending to publish from the Princeton Plasma Physics Laboratory and you have questions about the new submission process, please contact [publications@pppl.gov](mailto:publications@pppl.gov). For all other questions about getting started with submissions to Princeton Data Commons, please contact [prds@princeton.edu](mailto:prds@princeton.edu). For more information, please visit the [Princeton Research Data Service website](#).

DataSpace is a digital repository meant for both archiving and publicly disseminating digital data which are the result of research, academic, or administrative work performed by members of the Princeton University community. DataSpace will promote awareness of the data and address concerns for ensuring the long-term availability of data in the repository.

### Communities in DataSpace

<a href="#">Astrophysical Sciences</a>	<a href="#">Liechtenstein Institute on Self-Determination</a>
<a href="#">Chemical and Biological Engineering</a>	<a href="#">Mechanical and Aerospace Engineering</a>
<a href="#">Chemistry</a>	<a href="#">Molecular Biology</a>
<a href="#">Civil and Environmental Engineering</a>	<a href="#">Music and Arts</a>
<a href="#">Computational Social Science</a>	<a href="#">Office of Information Technology</a>
<a href="#">Department of Geosciences</a>	<a href="#">Physics</a>
<a href="#">Department of Slavic Languages and Literatures</a>	<a href="#">Princeton Neuroscience Institute</a>
<a href="#">Digital Humanities</a>	<a href="#">Princeton Plasma Physics Laboratory</a>
<a href="#">Economics</a>	<a href="#">Princeton School of Public and International Affairs</a>
<a href="#">Education Research Section</a>	<a href="#">Princeton University Doctoral Dissertations, 2011-2024</a>
<a href="#">Electrical and Computer Engineering</a>	<a href="#">Princeton University Library</a>
<a href="#">Faculty Publications</a>	<a href="#">Princeton University Masters Theses, 2022-2024</a>
<a href="#">Geophysical Fluid Dynamics Laboratory</a>	<a href="#">Princeton University Undergraduate Senior Theses, 1924-2023</a>
<a href="#">Industrial Relations Section</a>	<a href="#">Psychology</a>
<a href="#">Lewis-Sigler Institute for Integrative Genomics</a>	<a href="#">Seeger Center for Hellenic Studies</a>
	<a href="#">Sociology</a>



# DataSpace

- Heavily customized, legacy code base
- Out of support in the DSpace community
- No ability to develop new features



# DataSpace

- Descriptive metadata issues
  - Using Dublin Core
  - Not leveraging controlled vocabularies
- No metrics/analytics support
- No automated support for
  - DOI minting
  - Uploading files over 2.5GB
  - ORCID iDs



# How it's going: Princeton Data Commons (PDC)

A set of Open Source softwares  
supporting Research Data needs at  
Princeton

**Princeton Data Commons Discovery**  
A service of Princeton Research Data Service and Princeton University Library

Home About Features How to Submit Contact Us

Limit your search

- Domain
- Community
- Collection Tags
- Affiliation
- Type
- Year Published
- Source

Search and access Princeton's research data

All Fields Search... Search

Recently published

- Dataset  
Raw data for "OmpA controls order in the outer membrane and shares the mechanical load"  
Published 2024, Benn, Georgina
- Dataset  
Data repository for "Toward electron temperature profiles in hot-dense plasmas from x-ray spectral ensembles"  
Published 2024, Kraus, B. Frances et al.
- Waves from Shabansky Orbits in the Dayside
- Beam Emission Spectroscopy on NSTX-U"

**Princeton Data Commons Describe**

Welcome, Kate Administrator

My Dashboard My Reports

Welcome to Princeton Data Commons: Describe

Princeton Data Commons Describe is the submission portal for our data repository. Here you will find everything you need to know to successfully submit your research data and code to the Princeton Data Commons.

Who can submit to this repository?

All faculty, staff, and students of Princeton University are welcome to submit their **digital research data and code** to Princeton Data Commons. (Want to submit a manuscript, paper, or other scholarly work? Please see Princeton's [Open Access Repository](#)).

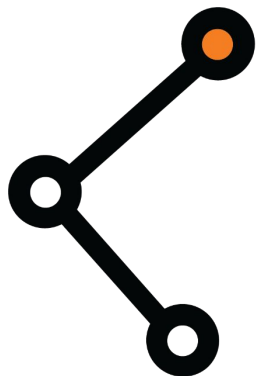
How to Submit

- Plan Ahead**
  - All submissions are reviewed (curated) by a team of data management experts (curators) in order to make the data and code submitted to PDC as findable, accessible, interoperable, and reusable (FAIR) as possible. The review process typically takes **5-10 business days** once a submission has been completed in the webportal.
- Prepare your files**
  - All files should be free of any information that is unsuitable for public release (e.g., sensitive personal information or intellectual property you do not own)
  - If at all possible, data files should be in open and widely accessible formats (e.g., CSV)
  - Provide adequate documentation for others who may re-use your data or code (e.g., at minimum a README.txt file explaining variable codes and relations among files; feel free to use [our template](#))



## Princeton Data Commons (PDC)

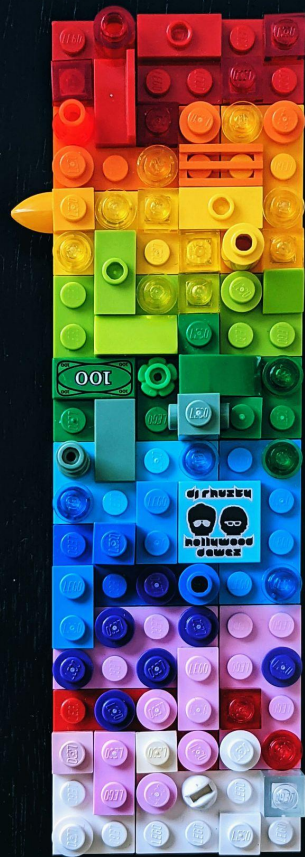
A project to create an ecosystem of **online resources** and **tools** in support of the needs of Princeton-affiliated research communities, especially surrounding the **research data lifecycle** from ingestion to publication, citation and beyond



Princeton **Data**  
**Commons**

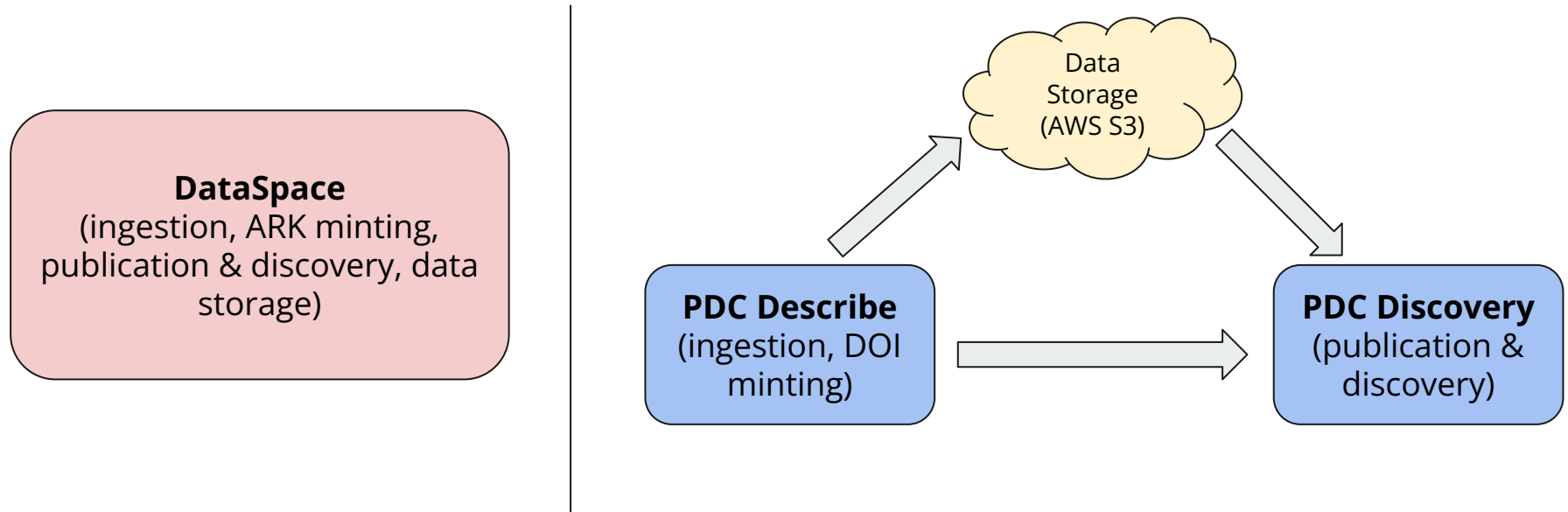
# Princeton Data Commons (PDC)

- PDC Discovery for discovery/dissemination
  - <https://datacommons.princeton.edu/discovery>
  - Code base:
    - [https://github.com/pulibrary/pdc\\_discovery](https://github.com/pulibrary/pdc_discovery)
- PDC Describe for submission
  - <https://datacommons.princeton.edu/describe/>
  - Code base:
    - [https://github.com/pulibrary/pdc\\_describe/](https://github.com/pulibrary/pdc_describe/)





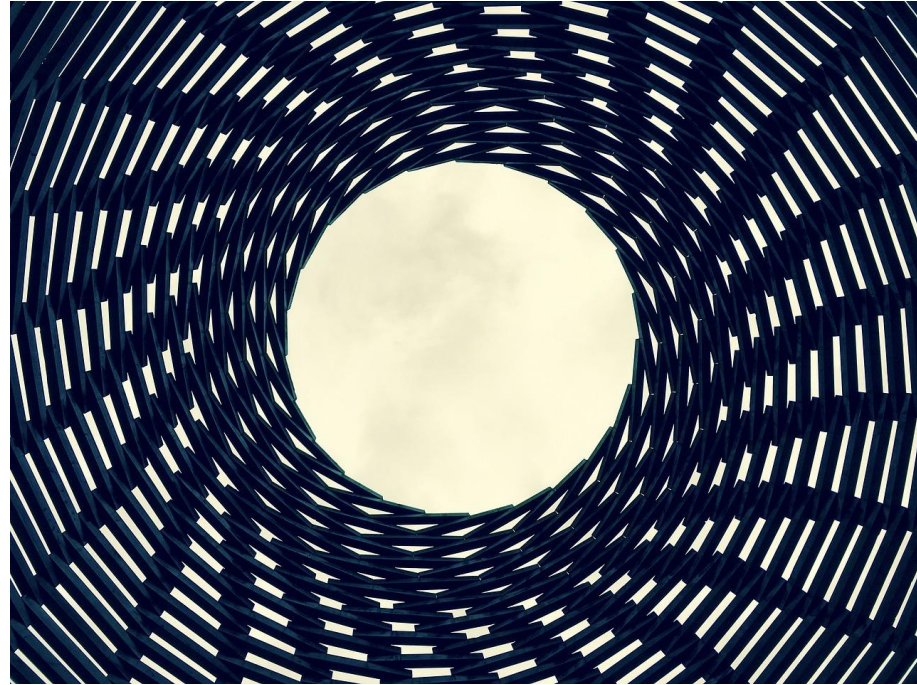
## DataSpace to Princeton Data Commons





# How we are building it

- Stakeholder support
- Many applications
- Iterative and incremental



# Stakeholder support

- This is not an “IT project”
  - It’s a Library project
- Library support at all levels
  - Administrators, curators, users, developers
- Key user groups
  - Princeton Plasma Physics Lab (PPPL)



Photo from [Pixabay](#)

## Stakeholder support

“Fund curation, not just storage.  
Invest in structures, standards, and  
capacity that allow data to not just be  
shared, but published.” - Wind Cowles &  
Mikala Narlock in [Beyond compliance: Curation as  
essential open science infrastructure](#)



Photo from [Pixabay](#)

# Many applications

- Applications
  - One for discovery (PDC Discovery)
  - One for curation (PDC Describe)
- Services
  - Storage (Amazon S3)
  - Data transfers via Globus
  - DataCite for DOI minting
- Plus other systems for non-research data

See also [Unboxing the Repository Model](#)

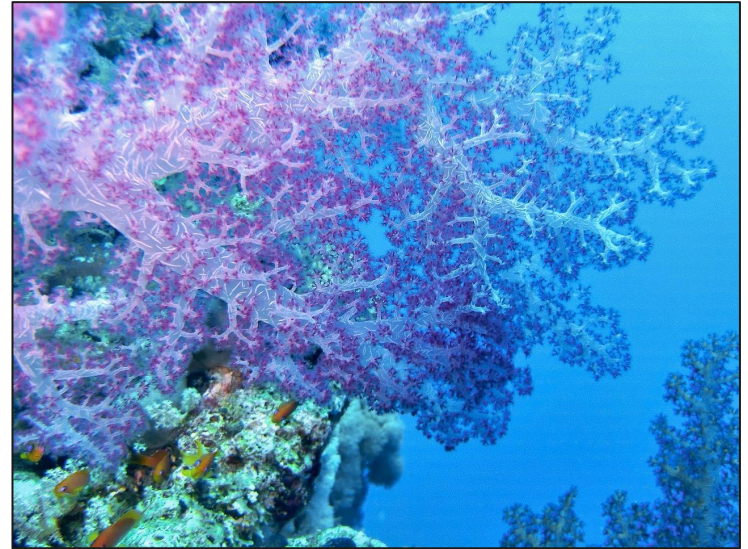


Photo by [Giustiliano Calgaro](#) from [Pixabay](#)



# Iterative and incremental

- Iterative and incremental (agile)
  - Discovery first, then Describe
  - ORCID support
  - Analytics (views & downloads)
  - Larger datasets
- Collaboration
  - Stakeholders, developers
  - On-going
  - Product Owner
  - Common language

See also [Agility, Growth, and Cooperative Service Design](#)

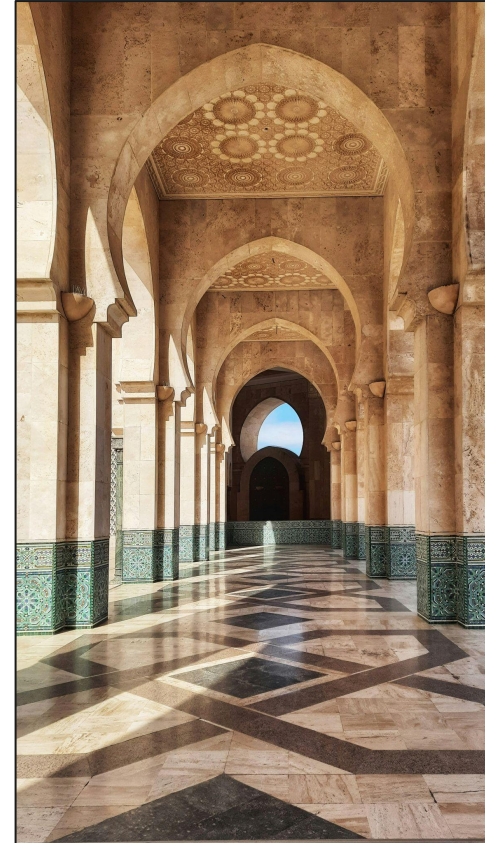


Photo by [Mueed Ahmed](#)



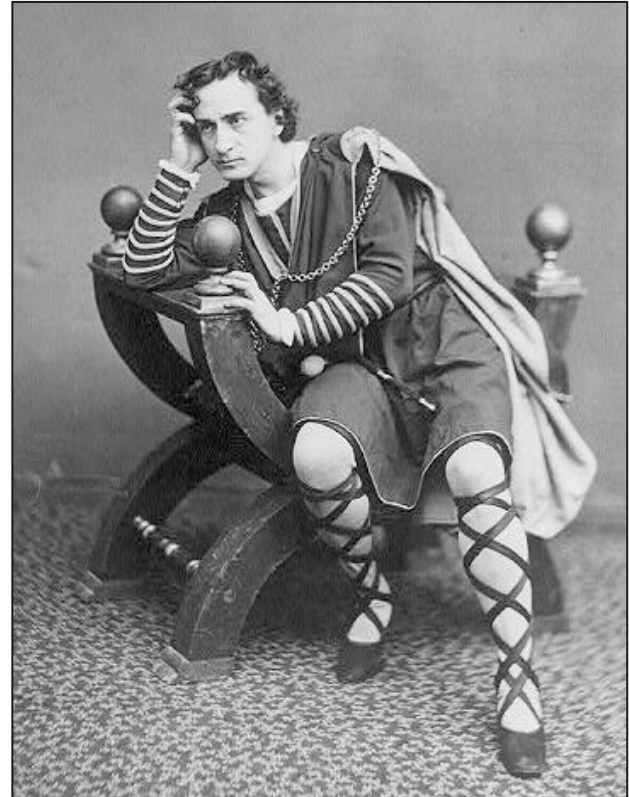
# Data Migrations

- Redescription vs migration
- Non-research data



## Redescription vs migration

"To migrate or to redescribe,  
that is the question" -  
Certainly not Shakespeare





# Redescription vs migration

- Metadata Schema
  - Dublin Core (old system)
  - DataCite (new system)
- About 400 datasets in DataSpace
  - Half of them by PPPL
  - Half from other groups
- Metadata manually redescribed by curators
- “Migration” party - all hands on deck
- We are almost done (fingers crossed)



Photo by [Shane Aldendorff](#) on [Unsplash](#)

## Non-research data

- Library PDFs
  - Migrating to another in-house repository, made discoverable in our library catalog
- Theses and Dissertations
  - Hosted DSpace 7.x instance
- This data will live somewhere else



Photo by [Melanie](#) on Pixabay



# What's next

With new software comes  
possibility



# What's next

- Building support for data versioning
  - Meets [NSTC Desirable Characteristics](#)
- Analytics strategy
  - Tracking citation behaviors
- Ongoing improvements for preservation
- Potential integrations
  - ORCID@Princeton
  - TigerData





# Thank you!

- Questions?
  - Kate Lynch
    - [katelynch@princeton.edu](mailto:katelynch@princeton.edu)
  - Hector Correa
    - [hector\\_correa@princeton.edu](mailto:hector_correa@princeton.edu)