Technical and Human Infrastructure for Open Research

Artemis Lavasa (CERN Scientific Information Service)

For the THOR consortium

PREDONx 2016

http://project-thor.eu


THOR


ORCID


EMBL-EBI


BRITISH
LIBRARY


DataCite


PANGAEA


DRYAD


CERN


PLOS


NON SOLUS
ELSEVIER


ands
AUSTRALIAN NATIONAL DATA SERVICE

Technical and Human infrastructure for Open Research

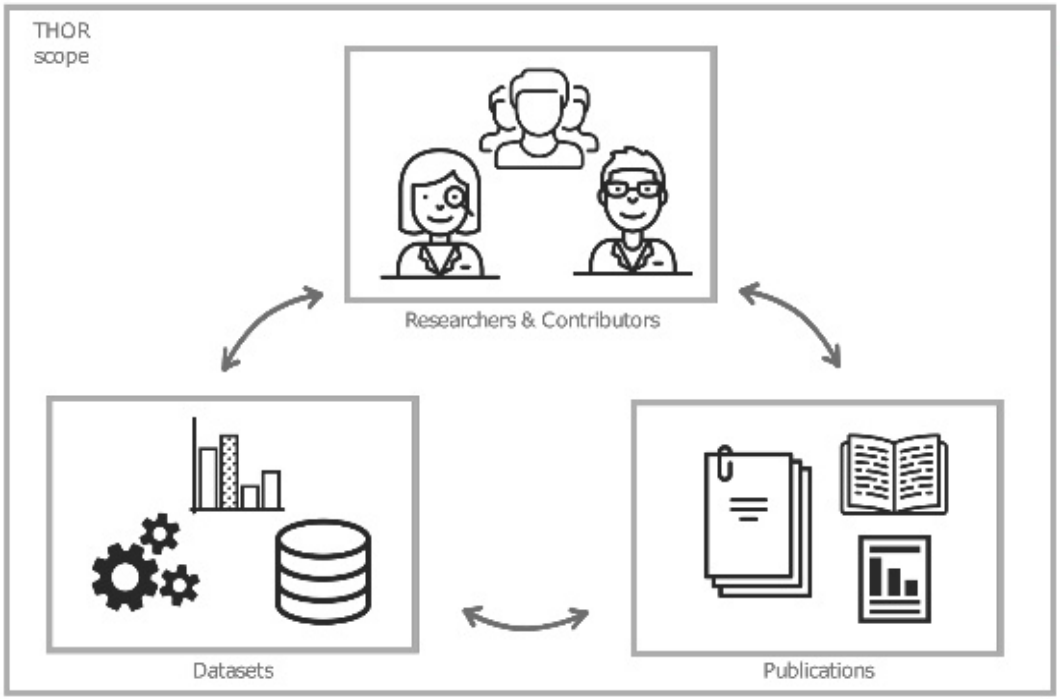
Establishing **seamless** integration between articles, data, and researchers across the research lifecycle

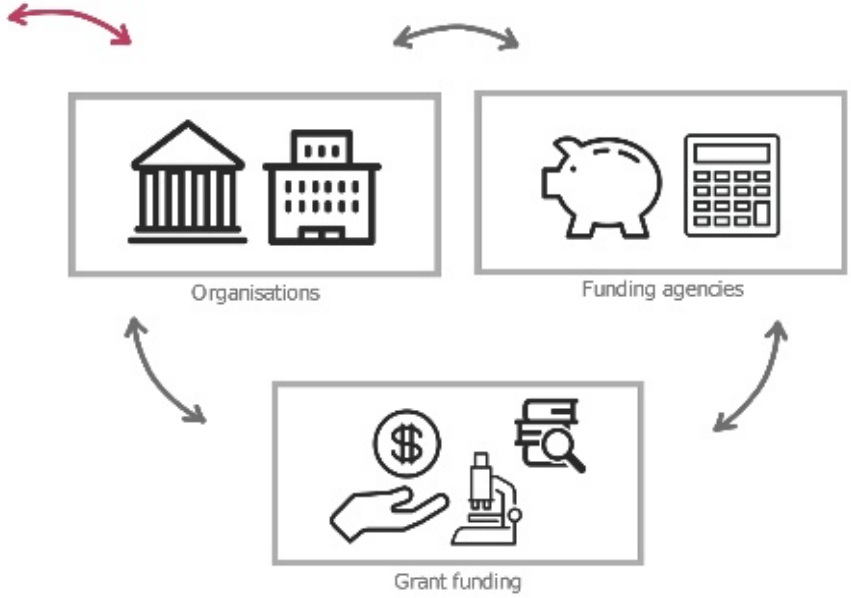
Making persistent identifier use for people and research artefacts *the default*

Comprises a technical **and** a human component

http://project-thor.eu

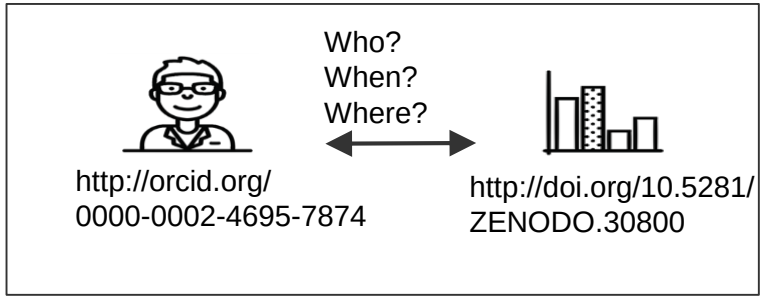
Seamless integration across the research lifecycle



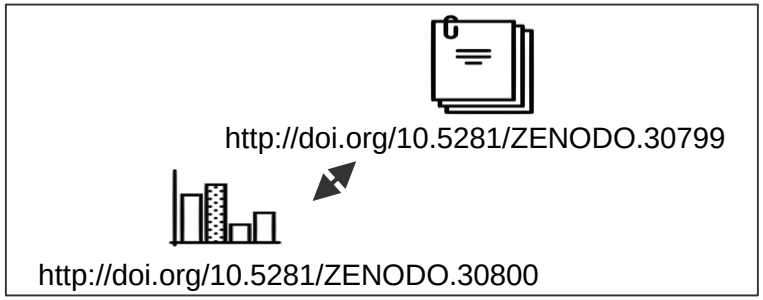


Trusted bridges across the research life-cycle

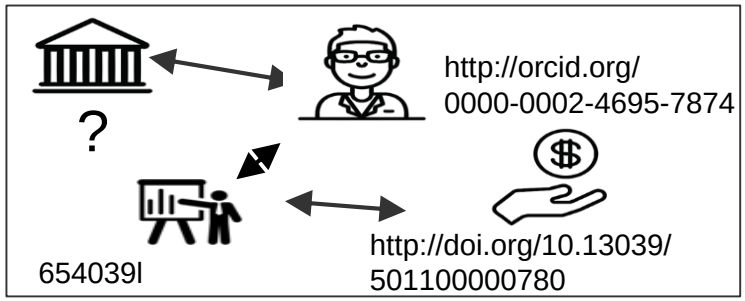
Linking data with data



Linking data with contributors



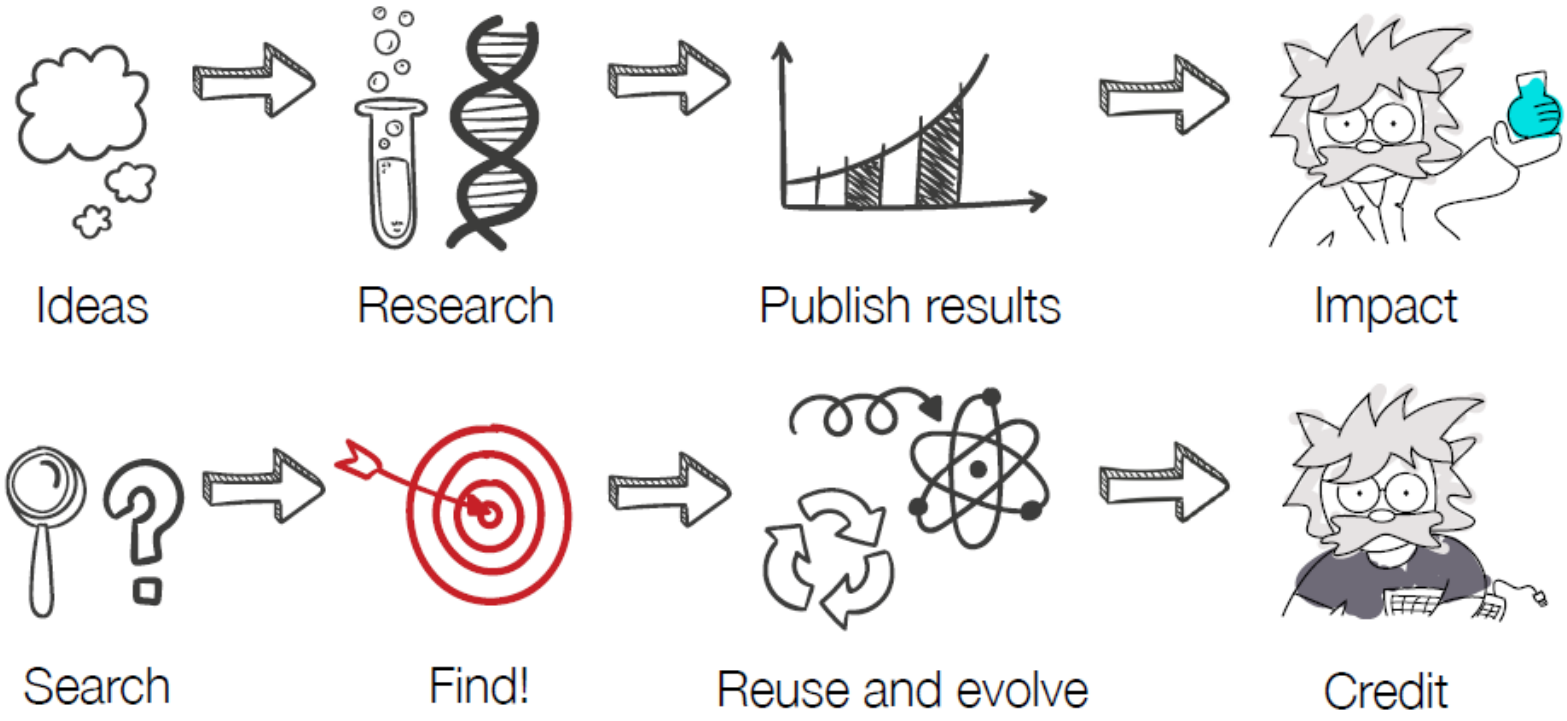
Linking data with articles



Linking data with institutions/funders

Slide credit to Trisha Cruse, Datacite

Supporting research



Persistent Identifiers (PIDs)

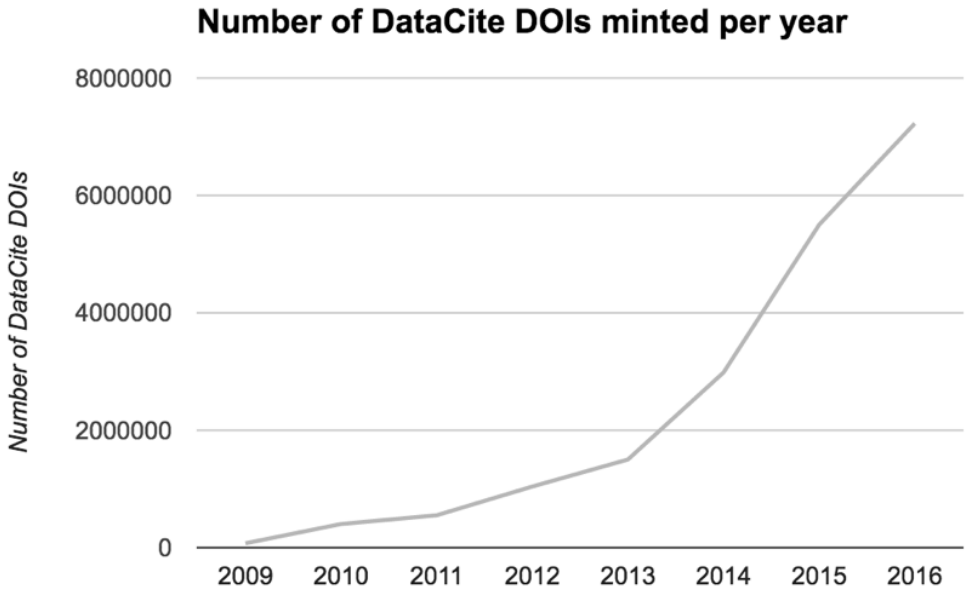
- Key technical infrastructure to connect services, authors, contributors
- The current need for persistent identifiers came out of problems with the way URLs were used in the early days of the World Wide Web.
  - In the long term URLs have proven to be fragile:
- THOR focuses on: DataCite DOIs and ORCID iDs
- But it also covers many other systems:
  - Handles, ARKs, ISNIs, Open Organisation Identifiers…

Many different PIDs

- Object indentifiers:
  - DOI - Digital Object Identifier
    - Crossref and DataCite are the main registration agencies for assigning DOIs for scholarly communication
  - Handle
  - URN - Uniform Resource Name
  - ARK - Archival Resource
- Author identifiers:
  - ORCID iD
  - Scopus ID
  - ResearcherID
  - ArXiv Author ID
  - PubMed Author ID
  - Microsoft academic research ID

DataCite

- Not-for-profit global initiative since 2009
- • > 40 members worldwide
- • > 800 data centres
- • Almost 8 million DOIs created
- • More than 8 million resolutions/month

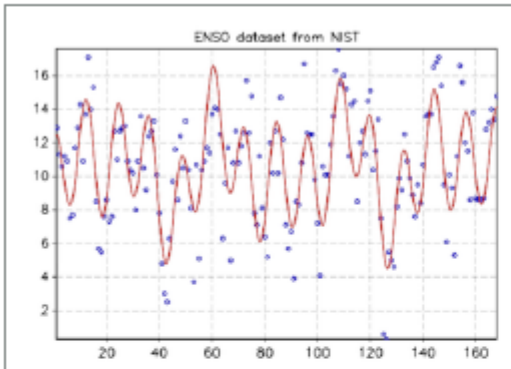


DataCite: Mission

- DataCite develops and support methods to:
  - Locate
  - Identify
  - Cite
- data and other research objects to:
  - Establish easier access
  - Increase acceptance
  - Foster reuse

Why use a DOI?

1. Take a dataset



2. Describe it

- Title
- Authors
- Year
- Description
- And others…

3. Assign a DOI

10.1234/exampledata

4. Reuse and reference!

ATLAS Collaboration, "Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC: $H \to \gamma\gamma$," http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

☑ Unique ☑ Persistent

5. Enjoy the benefits!

- Findability
- Track citations
- Reusability
- Measure impact

ORCID (Open Researcher and Contributor ID)

- Independent non-profit membership organization
- Non-proprietary and platform-neutral
- 556 Members worldwide
- Over 2.7 million ORCID iDs
- International service that integrates with other researcher identifiers
- Registry use is free for individuals
- Open data, software, APIs, and documentation

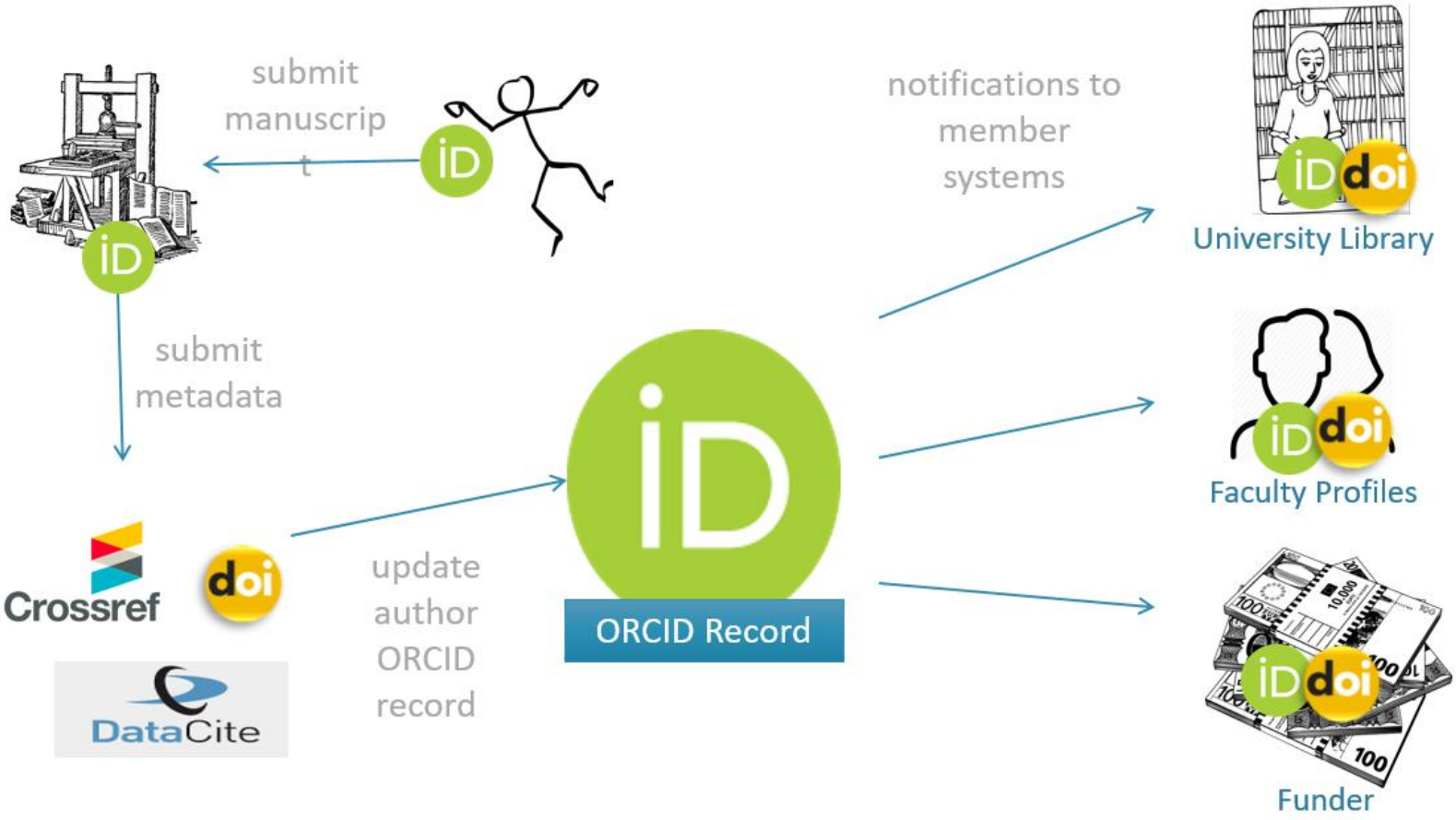
ORCID: Challenge

- Names are:
  - Messy
  - Not unique
  - Change
  - Difficult to transliterate
- To solve this, ORCID provides:
  - Persistent digital identifiers to distinguish researchers from each other
  - Member-built integrations that connect researchers and their activities/affiliations
  - A hub for synchronizing machine-readable connections between identifiers for people, organizations, and research activities

Why use an ORCID iD?

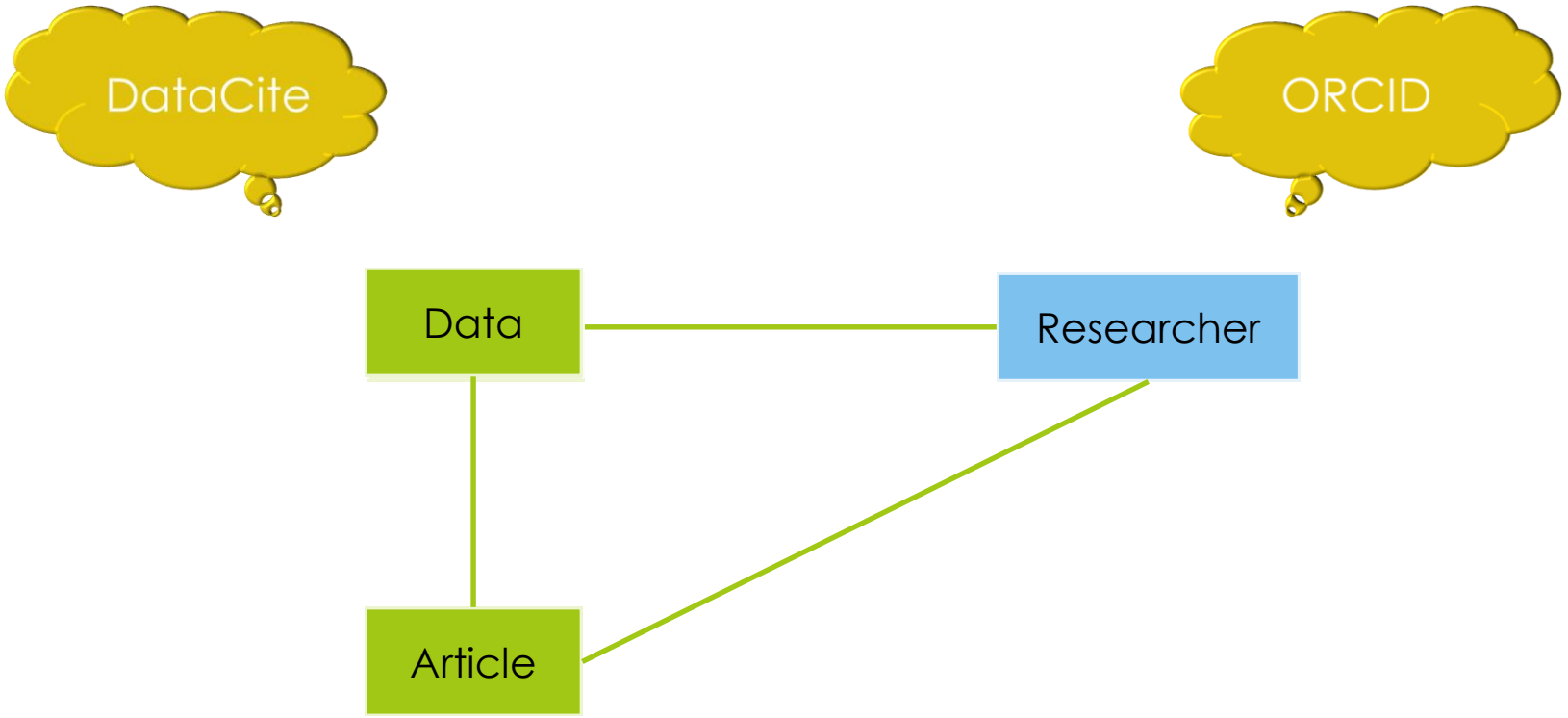
- Used by all major publishers and many repositories worldwide
- Enter information once, re-use often
- Report your activities easily
- Disambiguate your name
- Prove your affiliation
- Link your identities
- = everyone and everything gets credited for the work they share

Why use an ORCID iD?

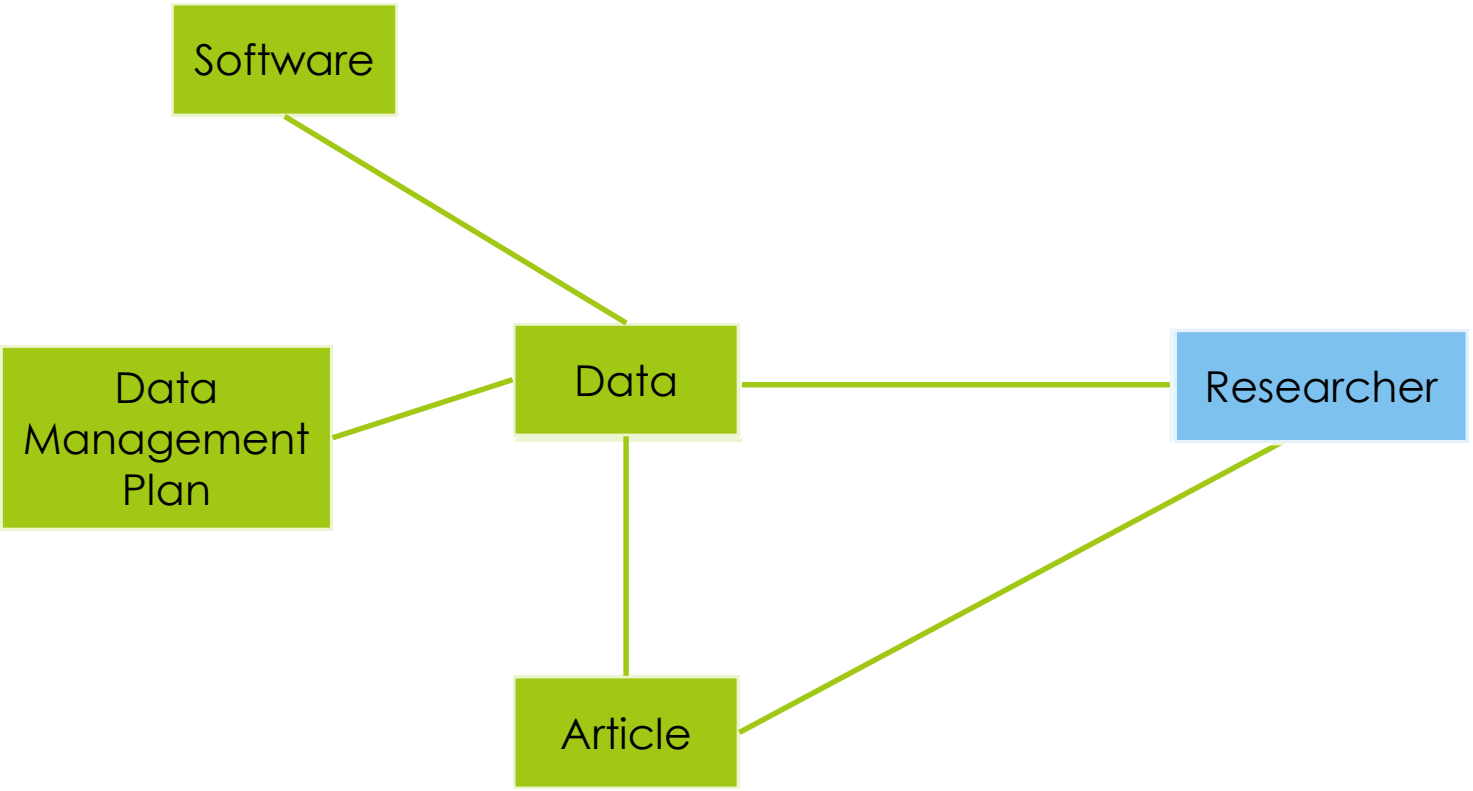


The Research Lifecycle Landscape

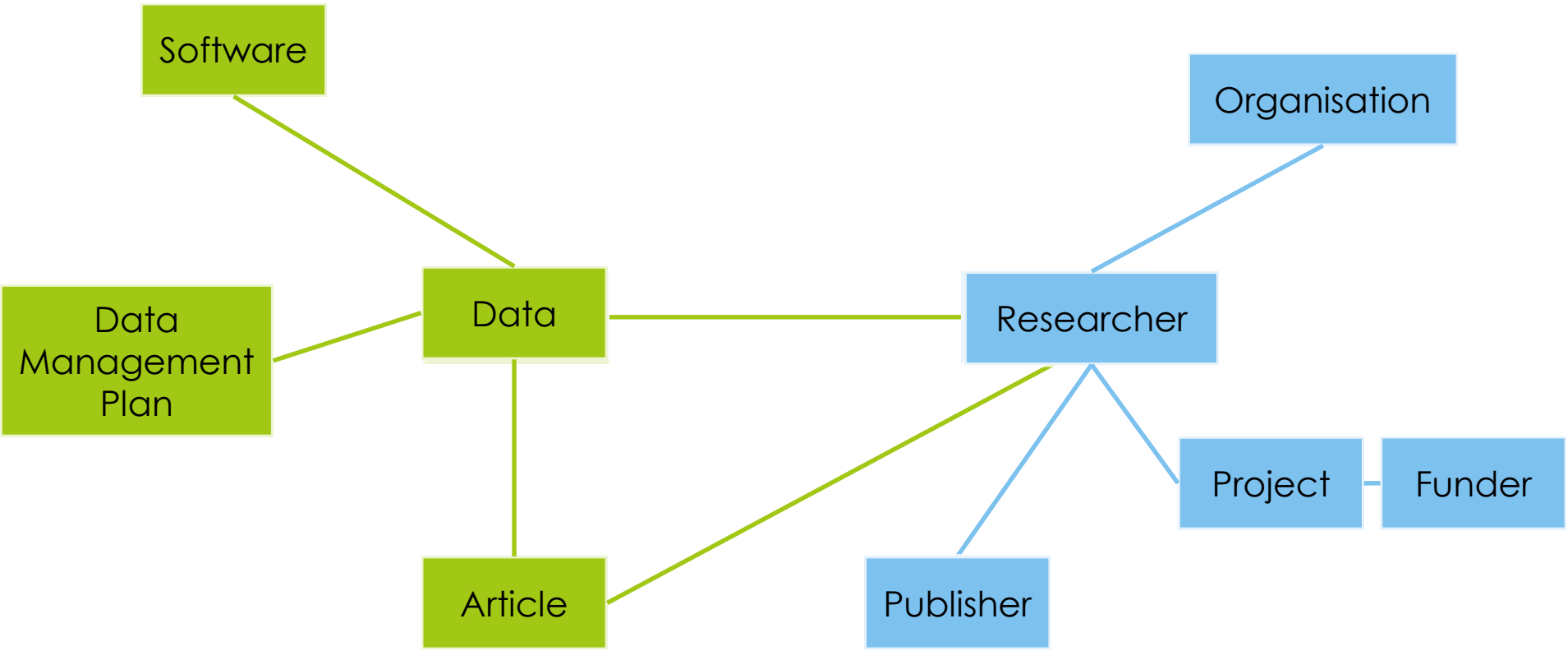
The Research Lifecycle Landscape



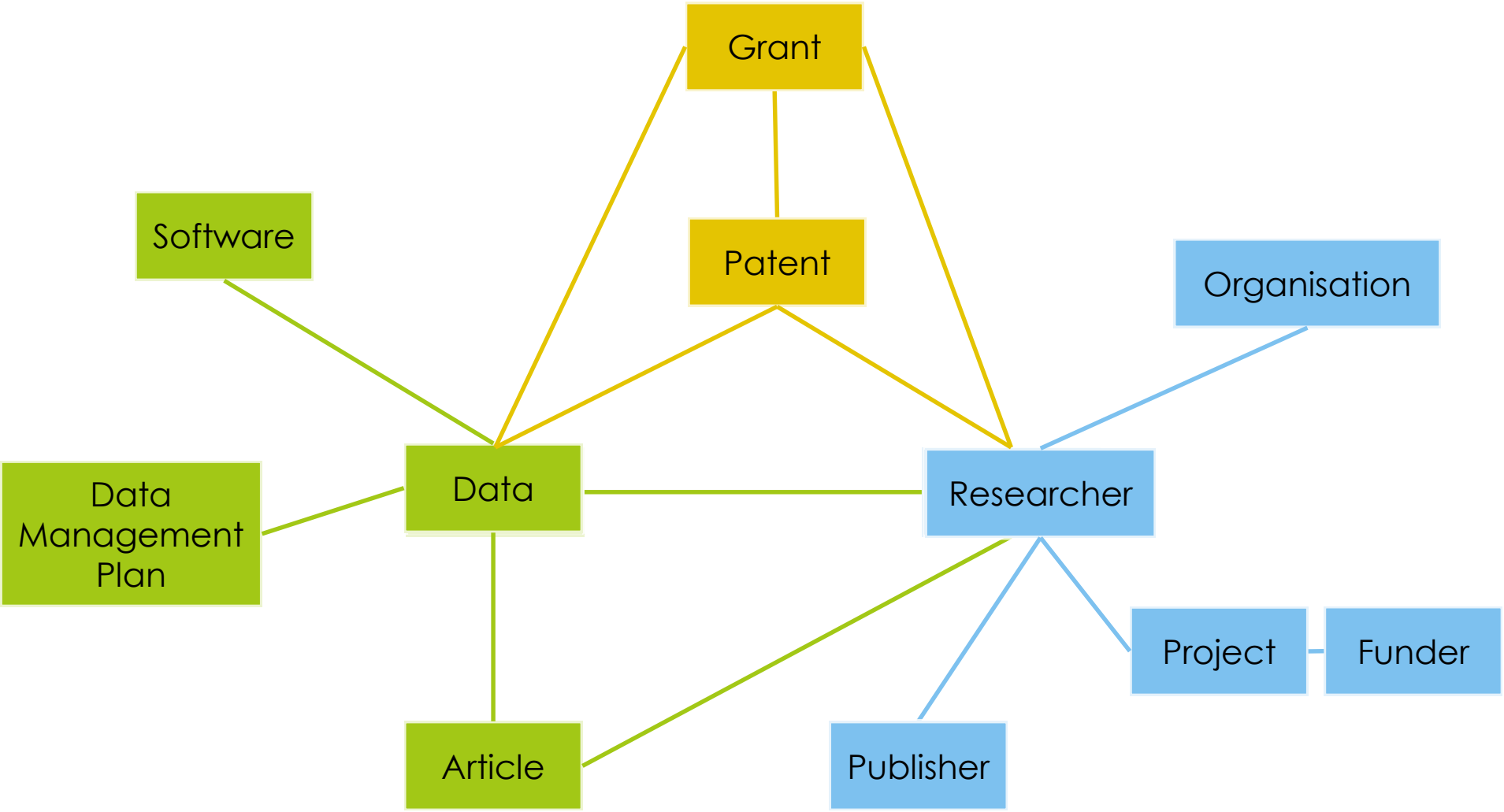
The Research Lifecycle Landscape



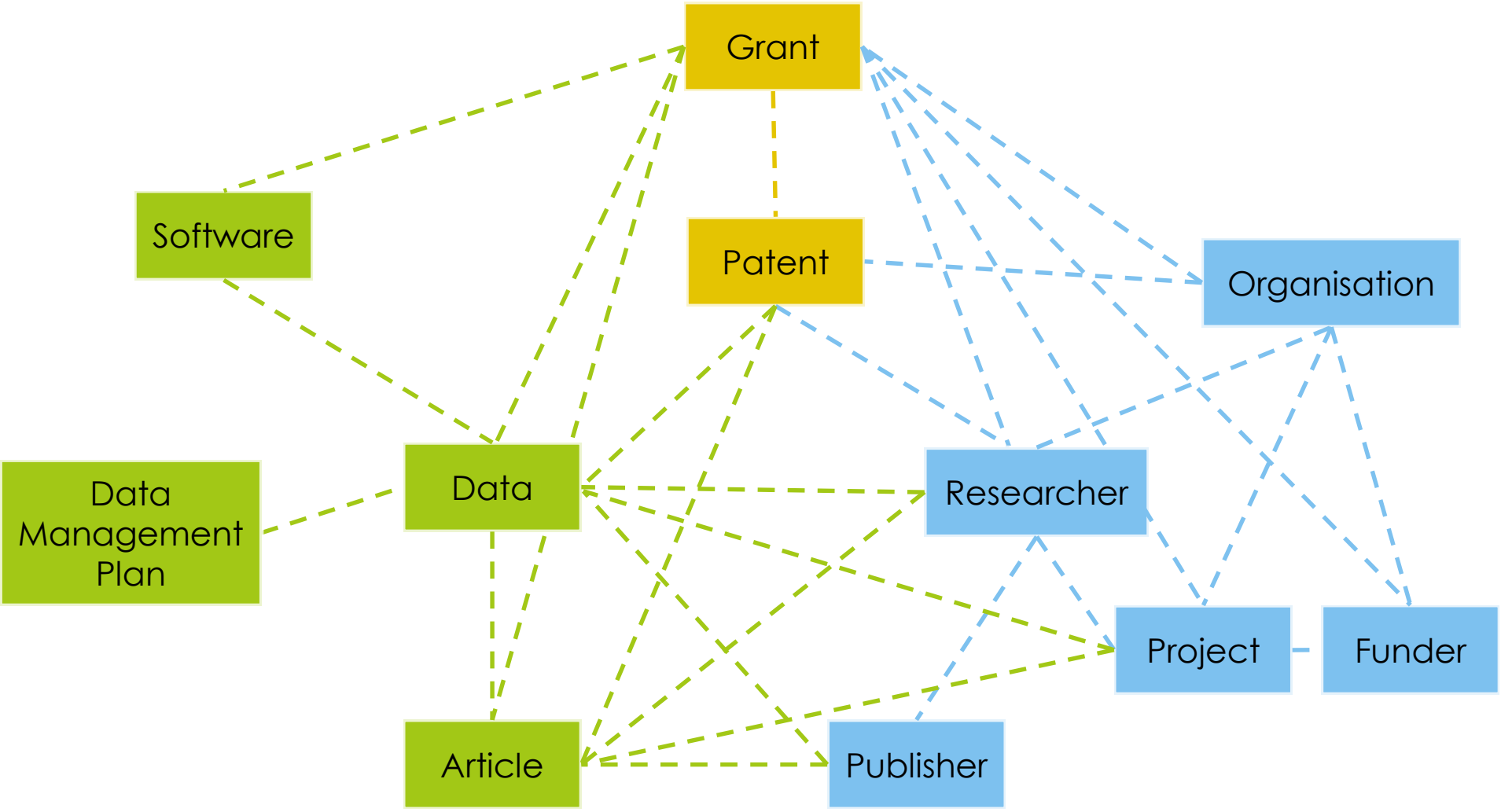
The Research Lifecycle Landscape



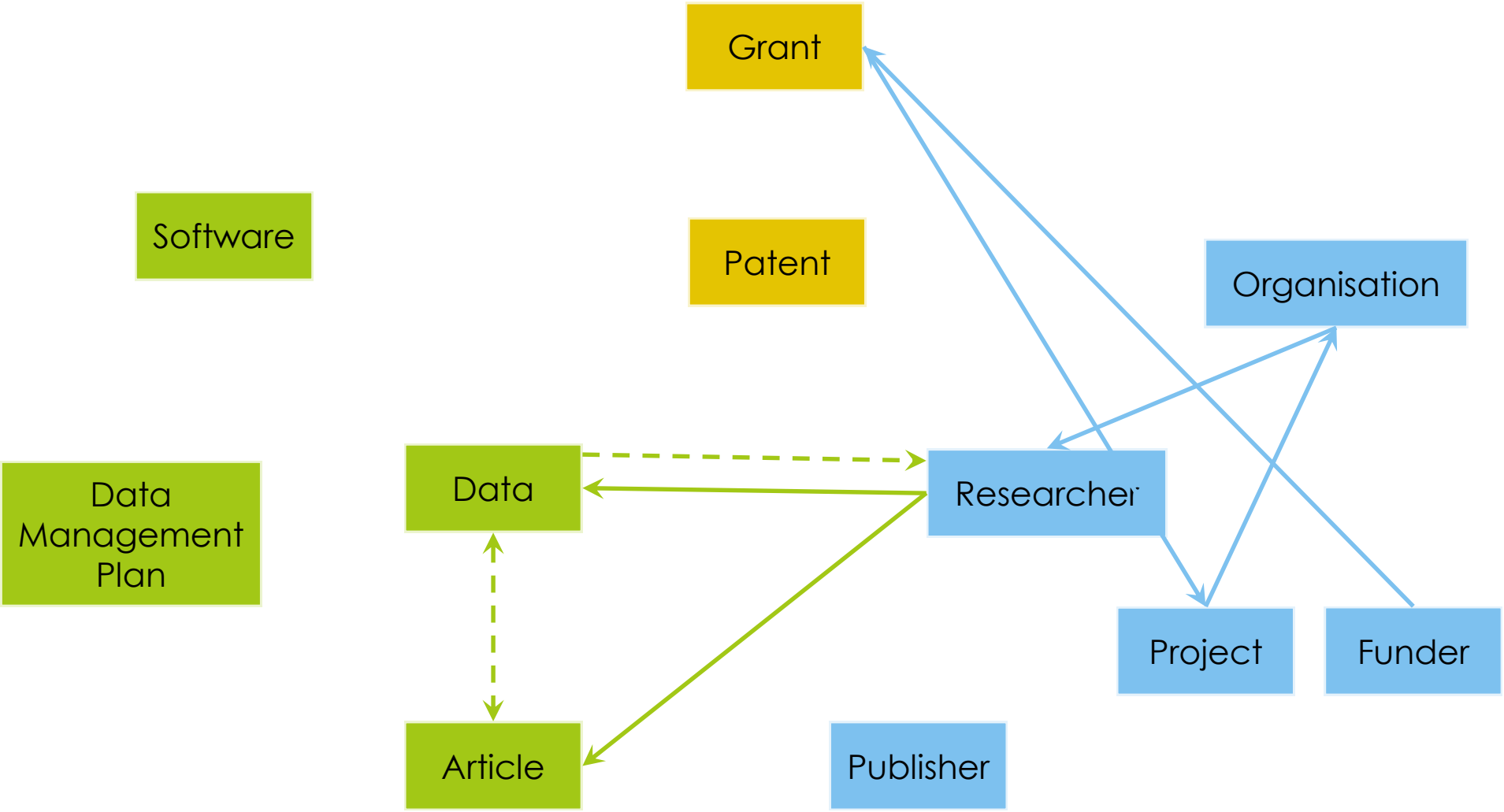
The Research Lifecycle Landscape



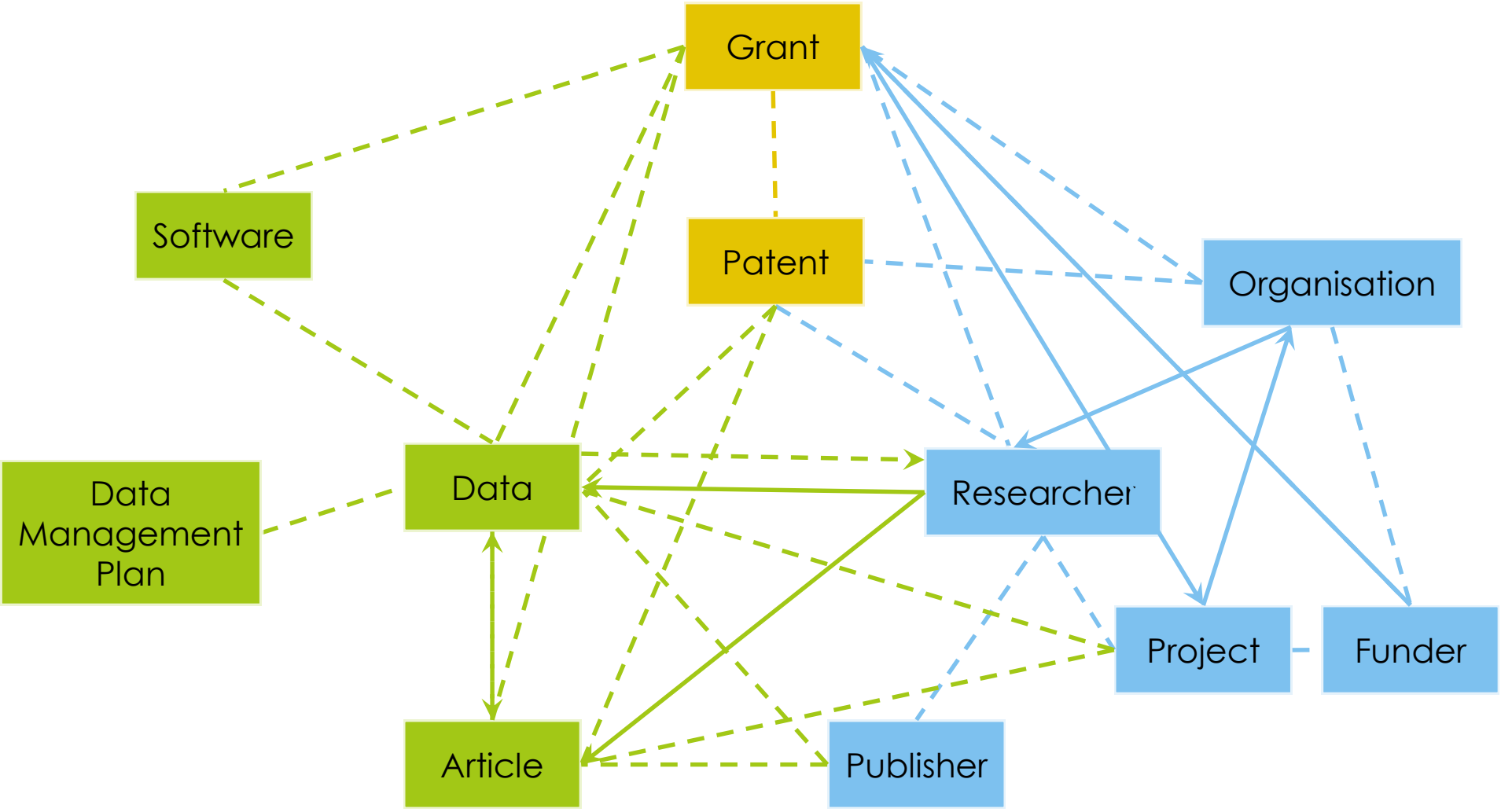
The Research Lifecycle
Many Interesting Relationships



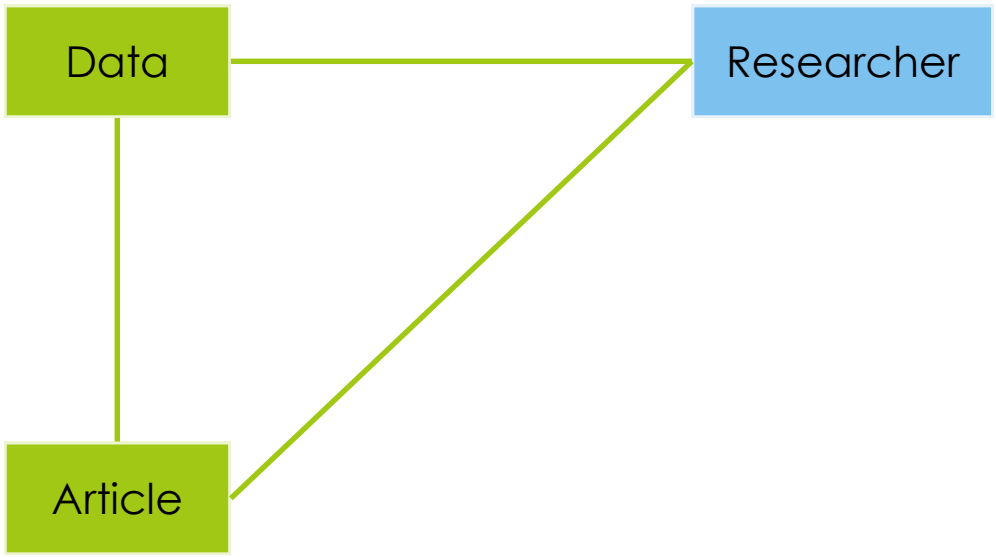
The Research Lifecycle
Few Well-Supported Relationships



The Research Lifecycle Landscape



Back to the Core Entities



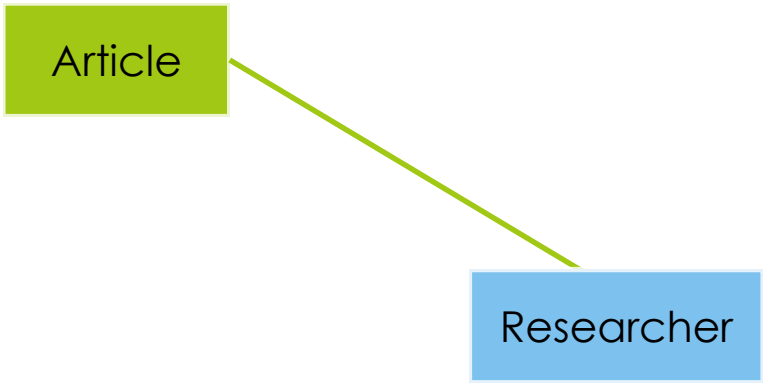
Connecting Researchers and Articles

Metadata does not disambiguate:

- **Different people with the same name**
- **Different spellings of a name**
- **Name changes over time**
- **Different transliterations**
- **…**

Does a metadata update result in a different instance of the Entity?

Local identifiers are brittle when they are used in a foreign environment



Persistent identifiers (PIDs)

- **disambiguate**
- **provide persistence**
- **even if metadata changes**

Connecting Researchers and Data

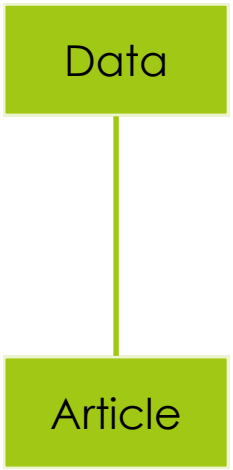
No tradition of crediting people in data sets

Implement dedicated data citation services with persistent identification

Connecting Articles and Data

Traditional data citation

- **references data sources ambiguously. Data cannot be identified.**
- **does not dereference to the actual data**



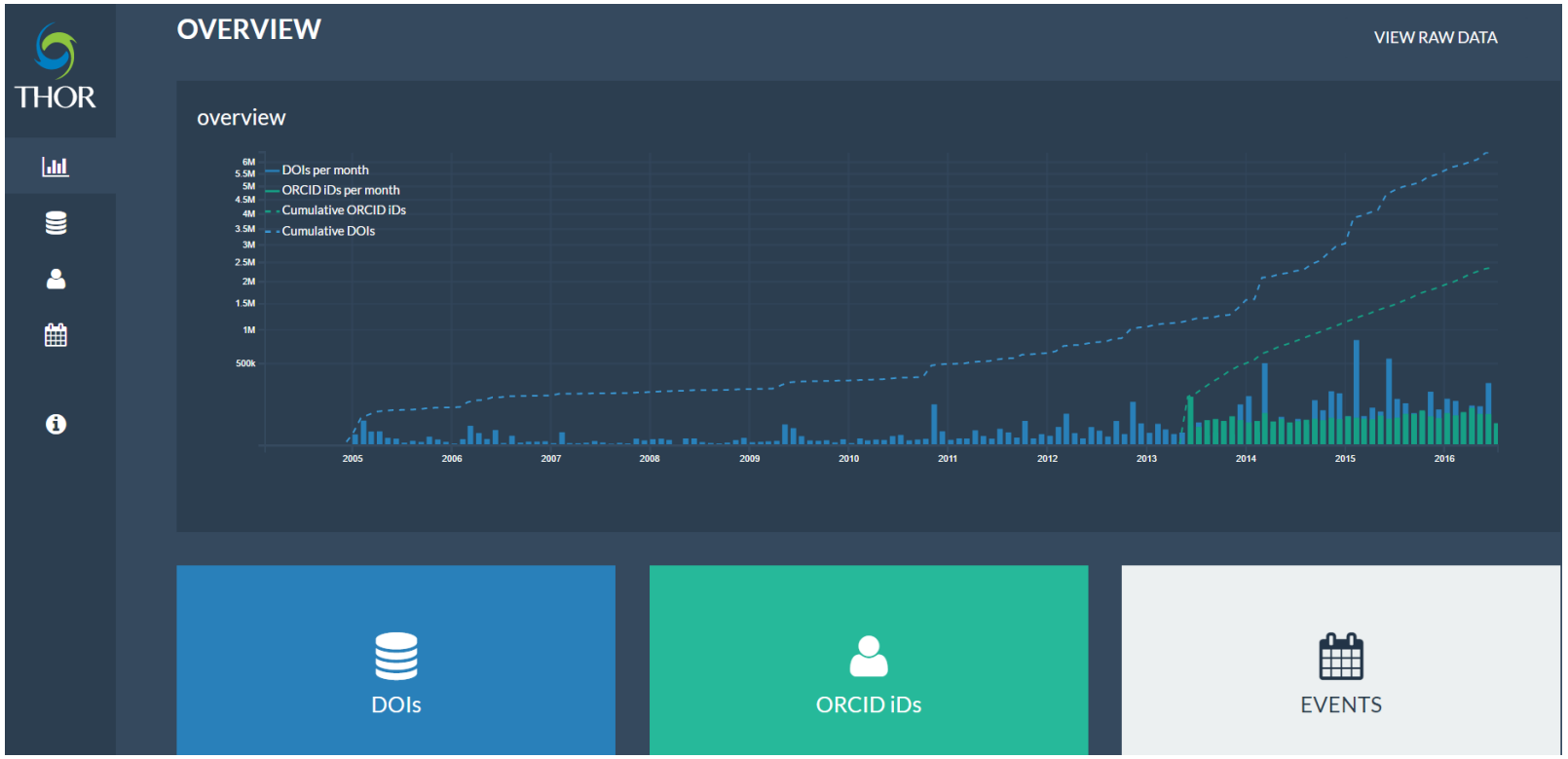
Implement dedicated data citation services with persistent identification and dereferencing

THOR Project - Approach

- **Research**: understanding challenges, supporting standards, designing workflows
- **Development**: building tools, setting up services, connecting platforms
- **Outreach**: getting others involved, develop skill sets
- **Sustainability**: making sure it lasts, providing evidence of improvements

Monitoring progress: metrics dashboard

http://dashboard.project-thor.eu



http://project-thor.eu

How does THOR impact data intensive research?

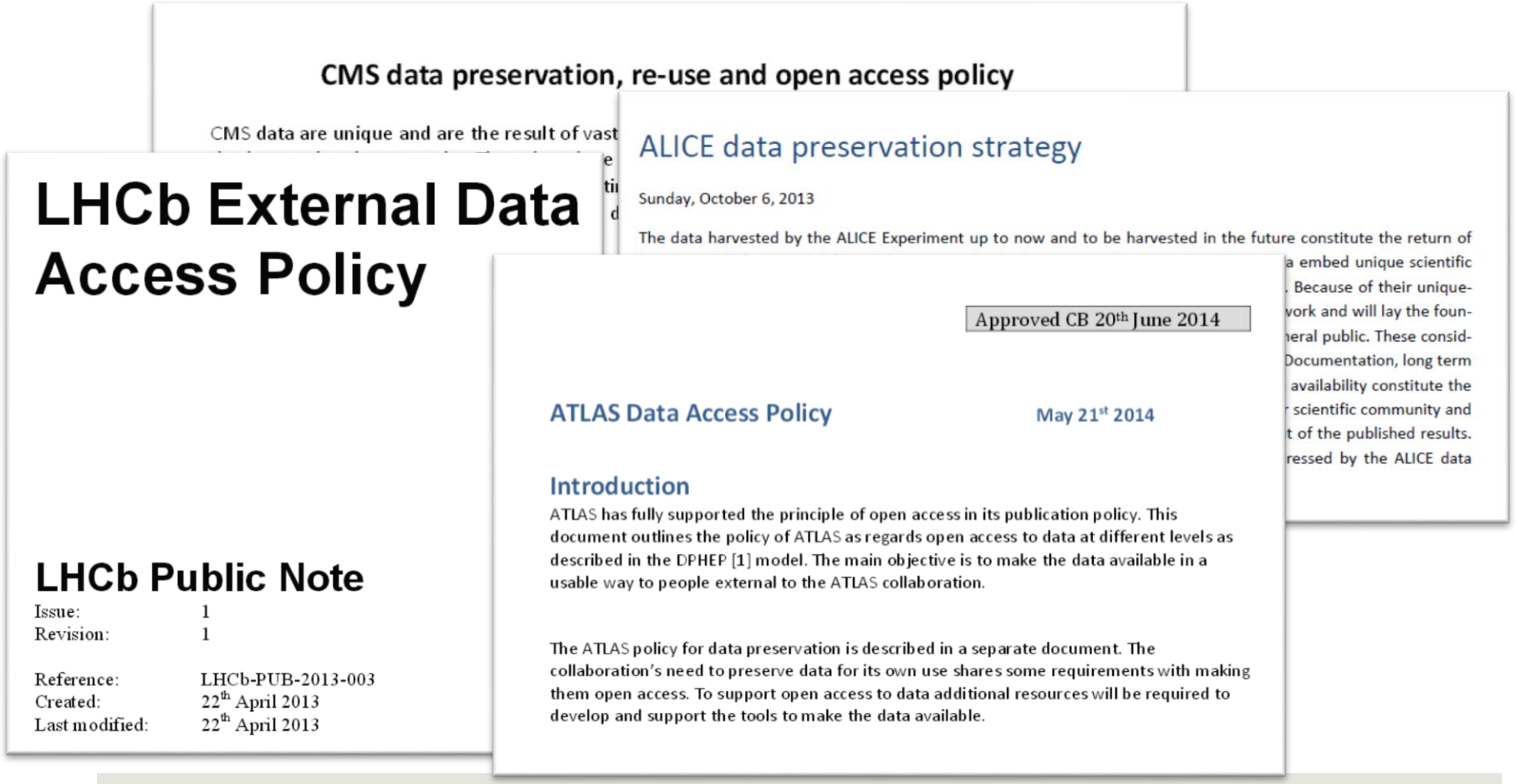
Example of one community

http://project-thor.eu

High-Energy Physics services

- INSPIRE-HEP (HEP literature, article-data links and citation tracking)
- HEPData (publication-ready data, article-data links)
- CERN Open Data (standalone, big data, public)
- CERN Analysis Preservation (preservation and reproducibility of analysis-level data, closed tool)

LHC policies



CMS data preservation, re-use and open access policy

CMS data are unique and are the result of vast

ALICE data preservation strategy

Sunday, October 6, 2013

The data harvested by the ALICE Experiment up to now and to be harvested in the future constitute the return of
a embed unique scientific
. Because of their unique-
work and will lay the foun-
neral public. These consid-
Documentation, long term
availability constitute the
scientific community and
t of the published results.
ressed by the ALICE data

LHCb External Data Access Policy

LHCb Public Note

Issue: 1
Revision: 1

Reference: LHCb-PUB-2013-003
Created: 22th April 2013
Last modified: 22th April 2013

Approved CB 20th June 2014

ATLAS Data Access Policy

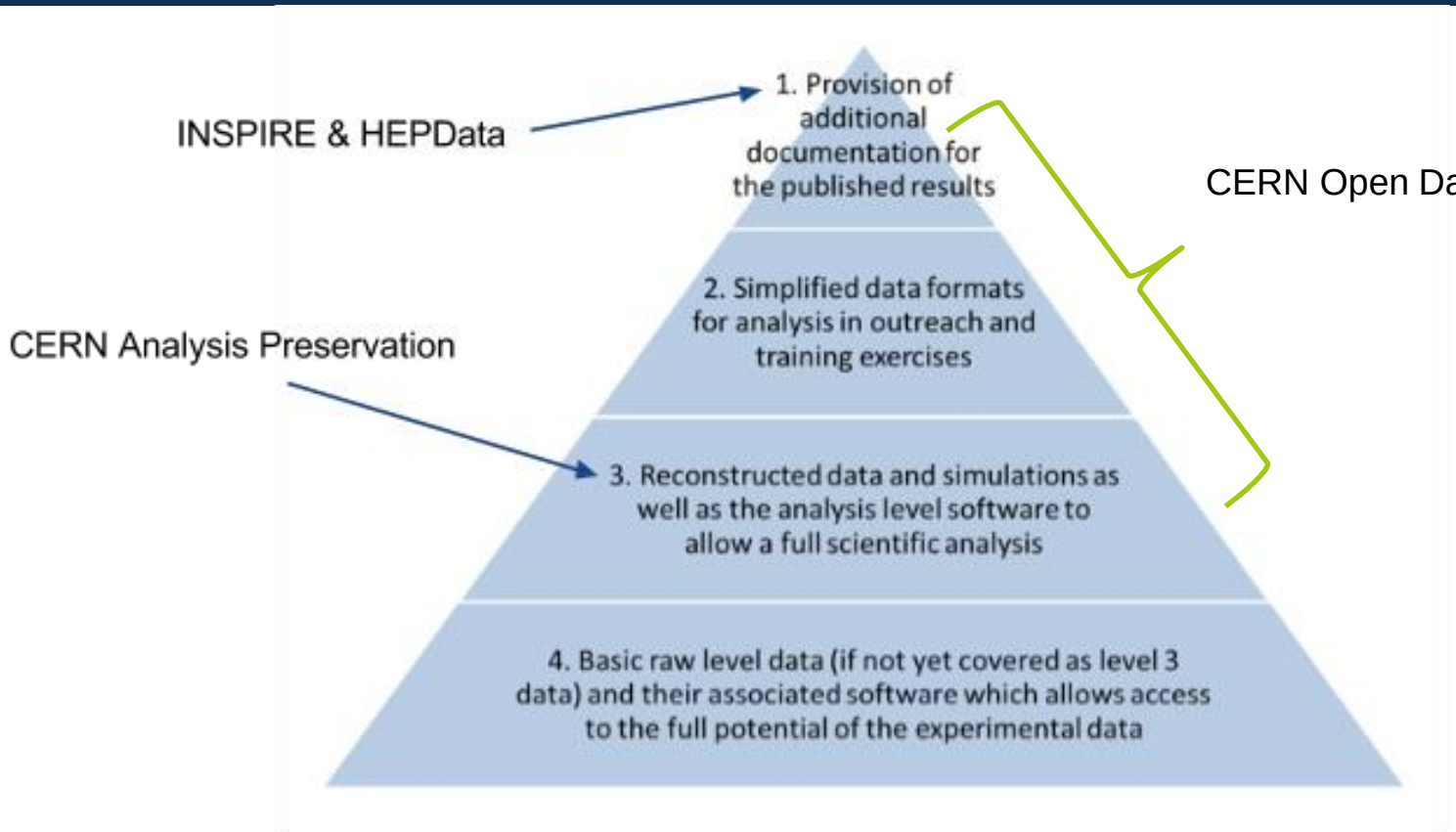
May 21st 2014

Introduction

ATLAS has fully supported the principle of open access in its publication policy. This document outlines the policy of ATLAS as regards open access to data at different levels as described in the DPHEP [1] model. The main objective is to make the data available in a usable way to people external to the ATLAS collaboration.

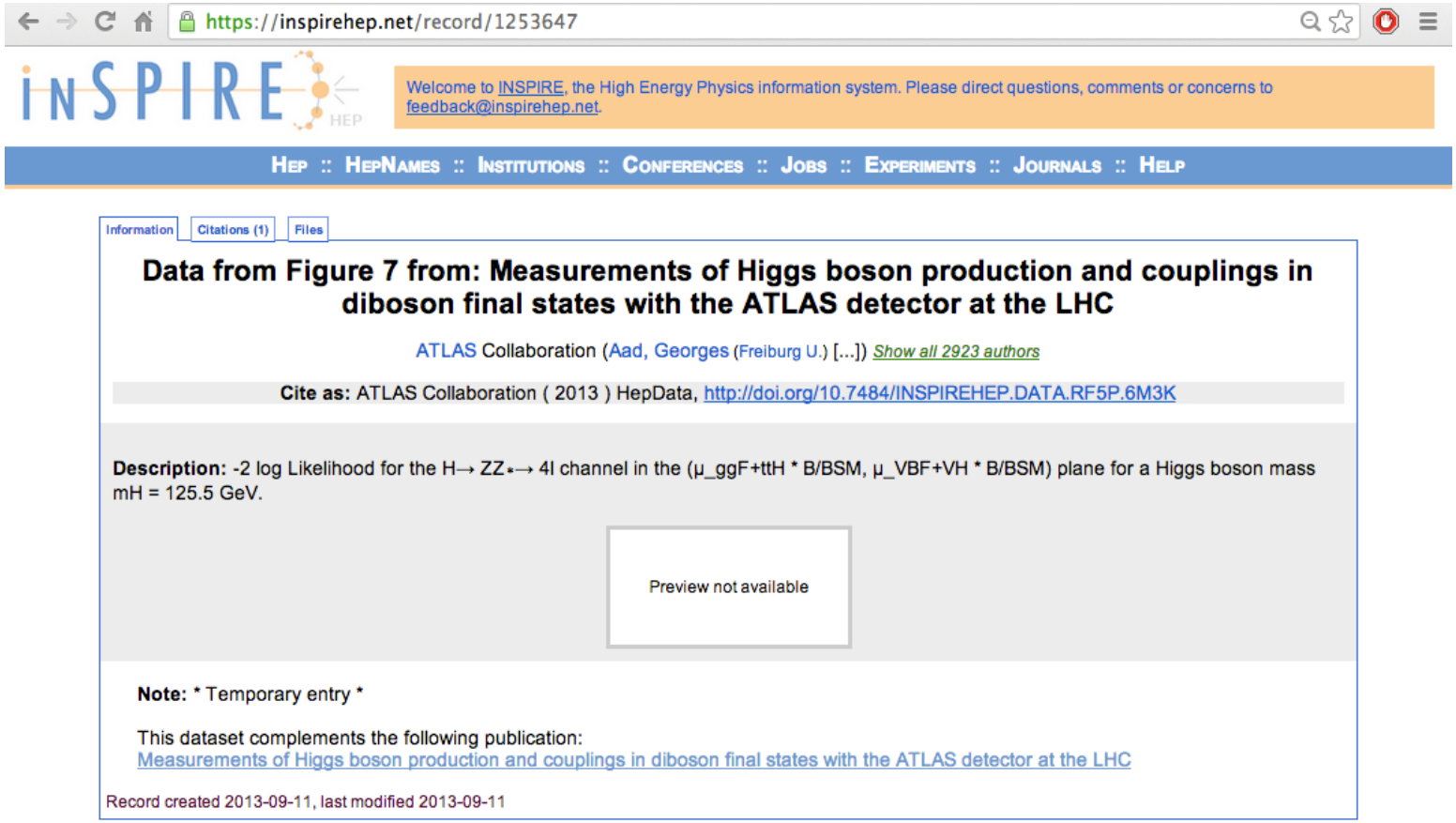
The ATLAS policy for data preservation is described in a separate document. The collaboration's need to preserve data for its own use shares some requirements with making them open access. To support open access to data additional resources will be required to develop and support the tools to make the data available.

Data in HEP



Pyramid of HEP data stages from: Herterich, P., & Dallmeier-Tiessen, S. (2016). Data Citation Services in the High-Energy Physics Community. *D-Lib Magazine*, *22*(1/2). http://doi.org/10.1045/january2016-herterich

Example DOI use



Counting citations to data: INSPIRE

Information | Citations (3) | Files

Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC - ATLAS Collaboration (for the collaboration)

Cited by: 3 records

(10) **On the presentation of the LHC Higgs Results** - Boudjema, F. *et al.* arXiv:1307.5865 [hep-ph]

(0) **Constraints on Higgs Couplings and Physics Beyond the Standard Model** - Belusca-Maito, Hermes *et al.* arXiv:1311.1113 [hep-ph]

(0) **A Novel Approach to Higgs Coupling Measurements** - Cranmer, Kyle *et al.* arXiv:1401.0080 [hep-ph]

more

.. of which self-citations: 0 records

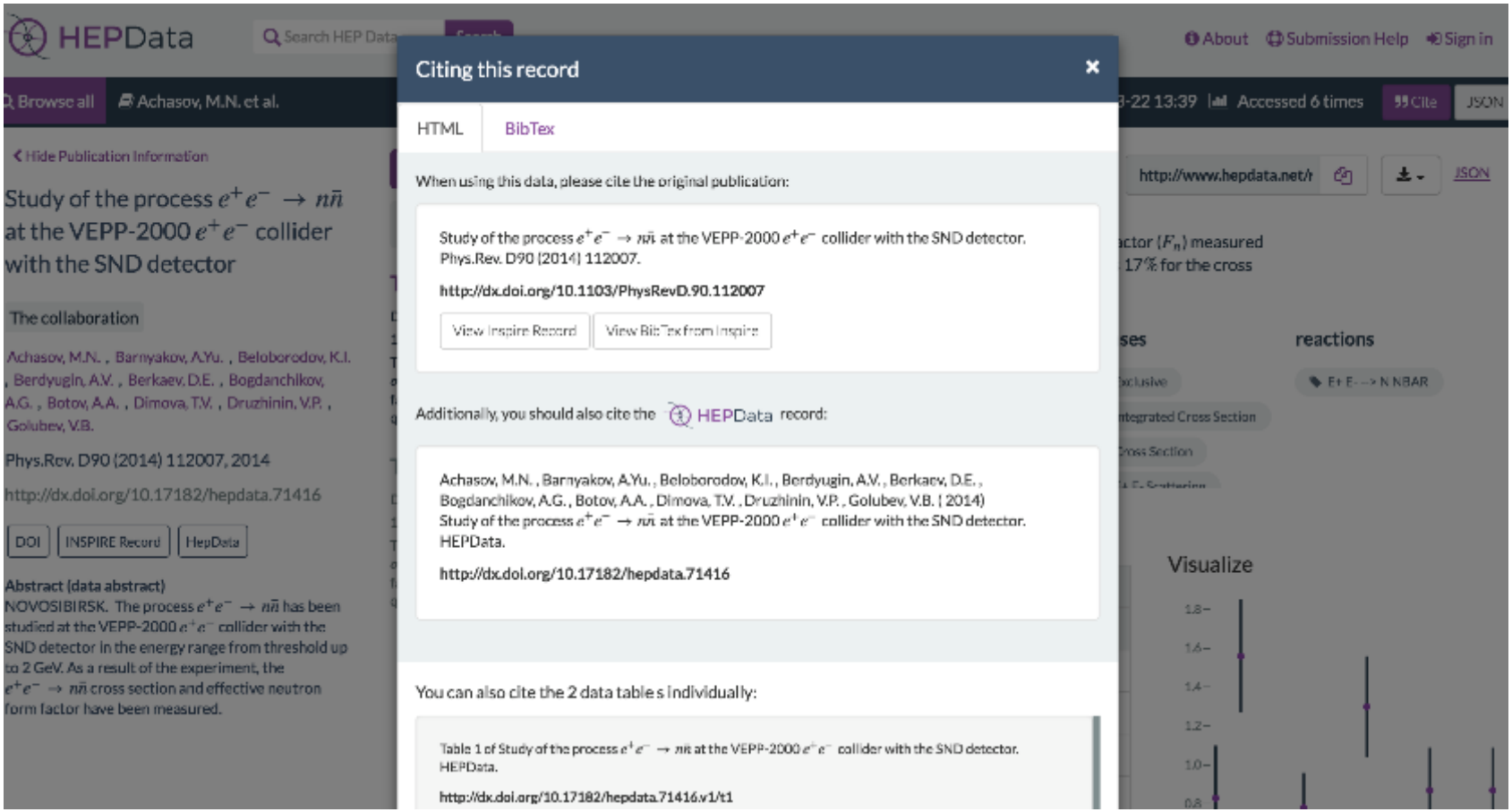
Co-cited with: 193 records

(3) **Higgs at last** - Falkowski, Adam *et al.* JHEP 1311 (2013) 111 arXiv:1303.1812 [hep-ph]

(3) **Combined coupling measurements of the Higgs-like boson with the ATLAS detector using up to 25 fb^{-1} of proton-proton collision data** - ATLAS Collaboration ATLAS-CONF-2013-034, ATLAS-COM-CONF-2013-035

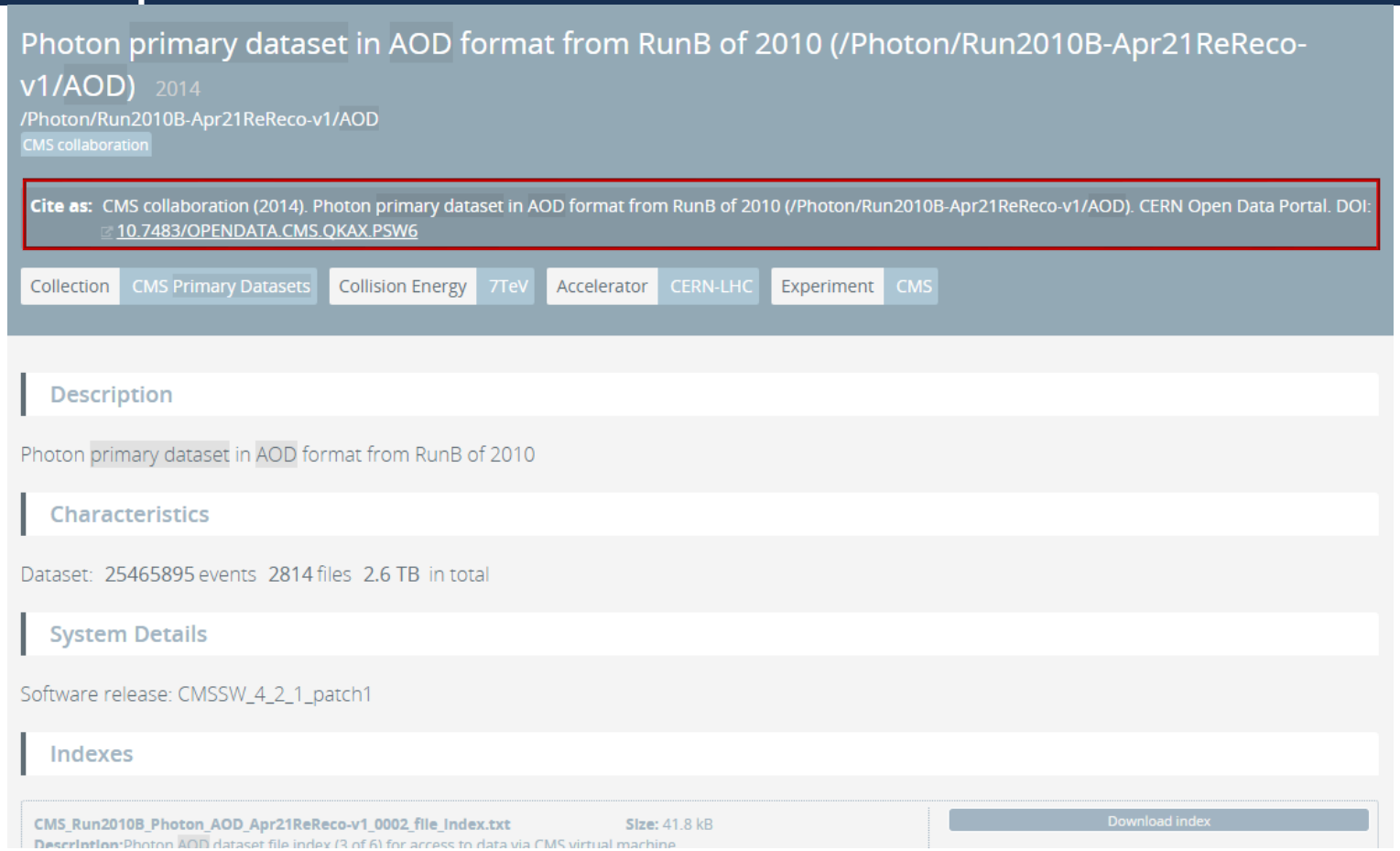
(3) **Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC** - ATLAS Collaboration (Aad, Georges *et al.*) Phys.Lett. B726 (2013) 88-119 arXiv:1307.1427 [hep-ex] CERN-PH-EP-2013-103

Facilitating data citation: HEPData



http://project-thor.eu

Facilitating data citation: CERN Open Data



Photon primary dataset in AOD format from RunB of 2010 (/Photon/Run2010B-Apr21ReReco-v1/AOD) 2014

/Photon/Run2010B-Apr21ReReco-v1/AOD

CMS collaboration

Cite as: CMS collaboration (2014). Photon primary dataset in AOD format from RunB of 2010 (/Photon/Run2010B-Apr21ReReco-v1/AOD). CERN Open Data Portal. DOI: 10.7483/OPENDATA.CMS.QKAX.PSW6

Collection: CMS Primary Datasets | Collision Energy: 7TeV | Accelerator: CERN-LHC | Experiment: CMS

Description

Photon primary dataset in AOD format from RunB of 2010

Characteristics

Dataset: 25465895 events 2814 files 2.6 TB in total

System Details

Software release: CMSSW_4_2_1_patch1

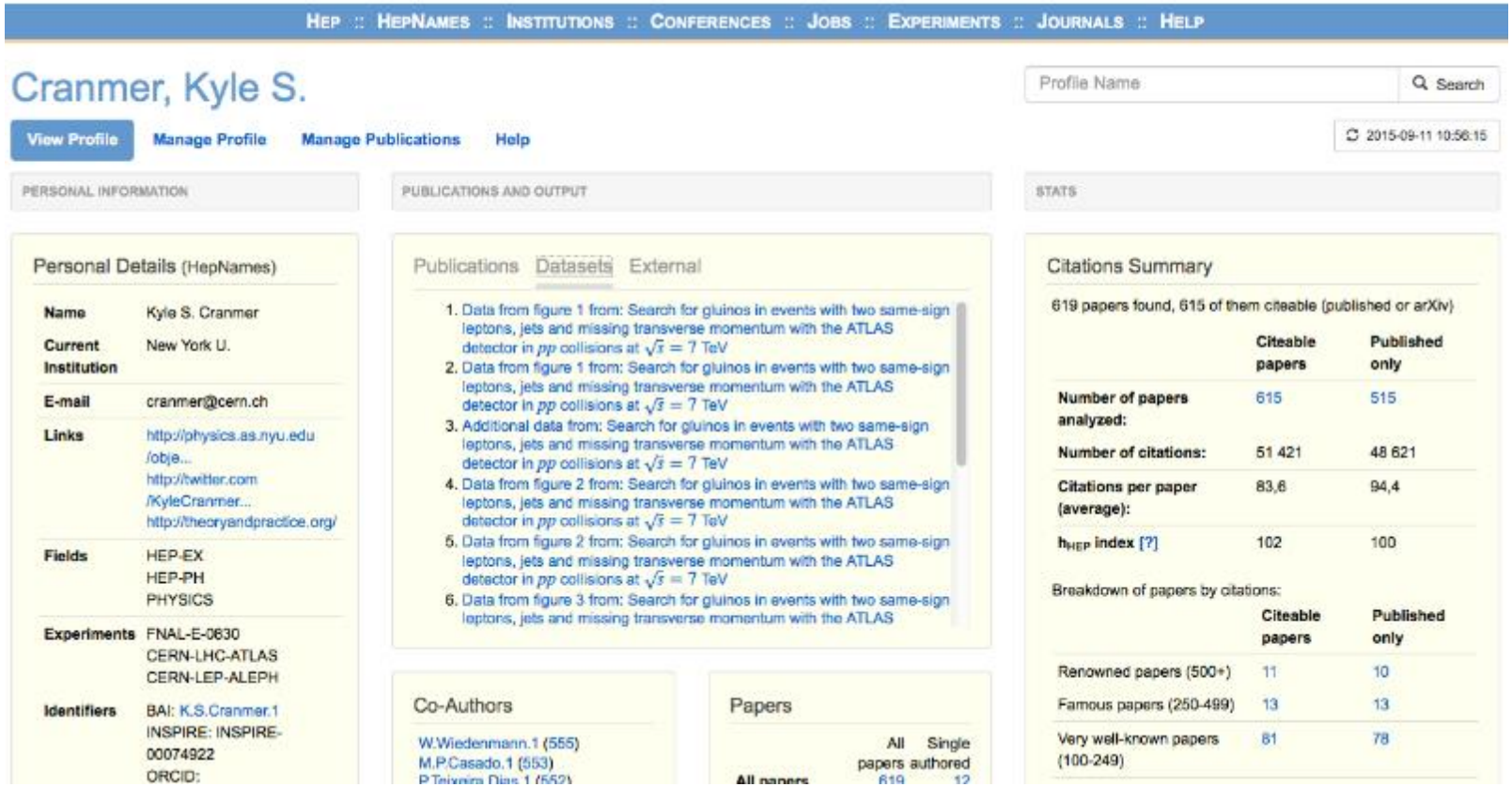
Indexes

CMS_Run2010B_Photon_AOD_Apr21ReReco-v1_0002_file_index.txt Size: 41.8 kB

Description: Photon AOD dataset file index (3 of 6) for access to data via CMS virtual machine

Download index

Showing impact: author profiles



http://project-thor.eu

Connecting to ORCID

Cranmer, Kyle S.

View Your Profile | **Manage Your Profile** | Manage Your Publications | Search Profiles | Help

Login with your arXiv.org account

You have succesfully logged in via arXiv.
You can now manage your profile.

Manage publications

Assign publications to your INSPIRE profile to keep it up to date.

Manage publication list

Automatically assigned publications

The following publications have been successfully assigned to your profile:

Publication title

1. Potential for Higgs physics at the LHC and super-LHC
2. Asymptotic distribution for two-sided tests with lower and upper boundaries on the parameter of interest
3. Frequentist hypothesis testing with background uncertainty
4. Multivariate analysis from a statistical point of view
5. RECAST: Extending the Impact of Existing Analyses
6. Challenges in moving the LEP Higgs statistics to the LHC
7. Maximum significance at the LHC and Higgs decays to muons
8. Search for neutral Higgs bosons decaying into four taus at LEP2
9. Kernel estimation in high-energy physics

Connect this profile to an ORCiD

This profile has not been connected to an ORCiD yet.

Connect an ORCiD to this profile

Merge profiles

If your or somebody else's publications in INSPIRE exist in multiple profiles, you can fix that here.

Merge profiles

HepNames data

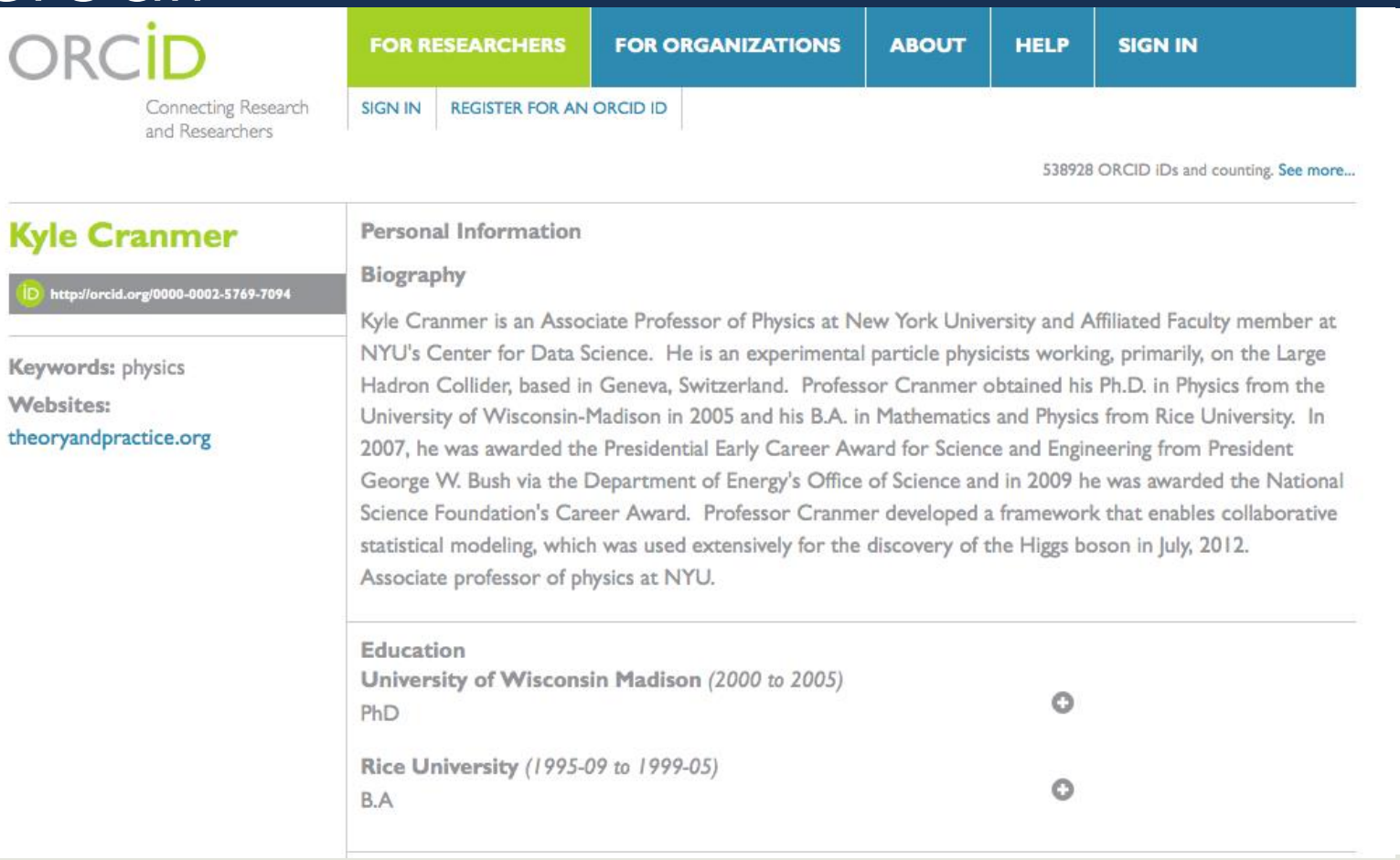
Kyle S. Cranmer (New York U.)
[Author Profile] [Google] [Students] [arXiv] [ADS]

PhD Advisor: Wu, Sau Lan
PhD Institution: Wisconsin U., Madison
Undergrad: Rice U.
Email: cranmer@cern.ch
URL: http://physics.as.nyu.edu/object/KyleCranmer.html
URL: http://twitter.com/KyleCranmer
Field: HEP-EX, HEP-PH, PHYSICS
Experiment: FNAL-E-0830, CERN-LHC-ATLAS, CERN-LEP-ALEPH, FNAL-TEV-CDF
Author Profile: K.S.Cranmer.1
Inspire ID: INSPIRE-00074922

Institutional History:

Institution	Rank	Start Date	End Date	UPDATE
New York U.	SENIOR	2007		

Using ORCID IDs for attributing credit



Using ORCID IDs for attributing credit

ORCID
Connecting Research and Researchers

Kyle Cranmer

http://orcid.org/0000-0002-5769-7094

Keywords: physics

Websites:
theoryandpractice.org

Works

Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC 2013-09

Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC 2013-09

Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC 2013-09

Replication data for: "Natural Priors, CMSSM Fits and LHC Weather Forecasts" 2013-07

Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC: Physics Letters B 2012

Challenge

- Adoption
- We can help with data curation and services, but we need collaboration to build services, operate them
- Services as an intrinsic process for researchers
  - Integrate in the publishing process
  - Provide incentives
  - Make them easy to use
  - Give it more visibility – make it discoverable

Thank you

artemis.lavasa@cern.ch