Research Objects and WS Characterization

Jose Enrique Ruiz
jer@iaa.es

November 29th 2012
Workflows Group ASOV France

Wf4Ever

2011 - 2013

Advanced Workflow Preservation Technologies for Enhanced Science

1. Intelligent Software Compon Spain)
2. University of Manchester
3. Universidad Politécn , Spain)
4. Poznan Superco king C
5. University of OXF, U

6. Institu alucía (
7. Leiden U dical C

Reproducible Science

Web-services based Workflows

Astronomy Research Lifecycle

Astronomy research lifecycle is **entirely digital**

- Observation proposals
- Data reduction pipelines
- Analysis of science ready data
- Catalogs of objects and data
- Publish process
  - Final data results
  - Experiment in DL

  ADS/arXiv

Reproducible research is still not possible in a digital world

A normalized preservation of methodology is needed

Efficient use of rich data infrastructure (VO) may be improved

Tools

Efficiency and Reuse

Optimize return on investments made on big facilities

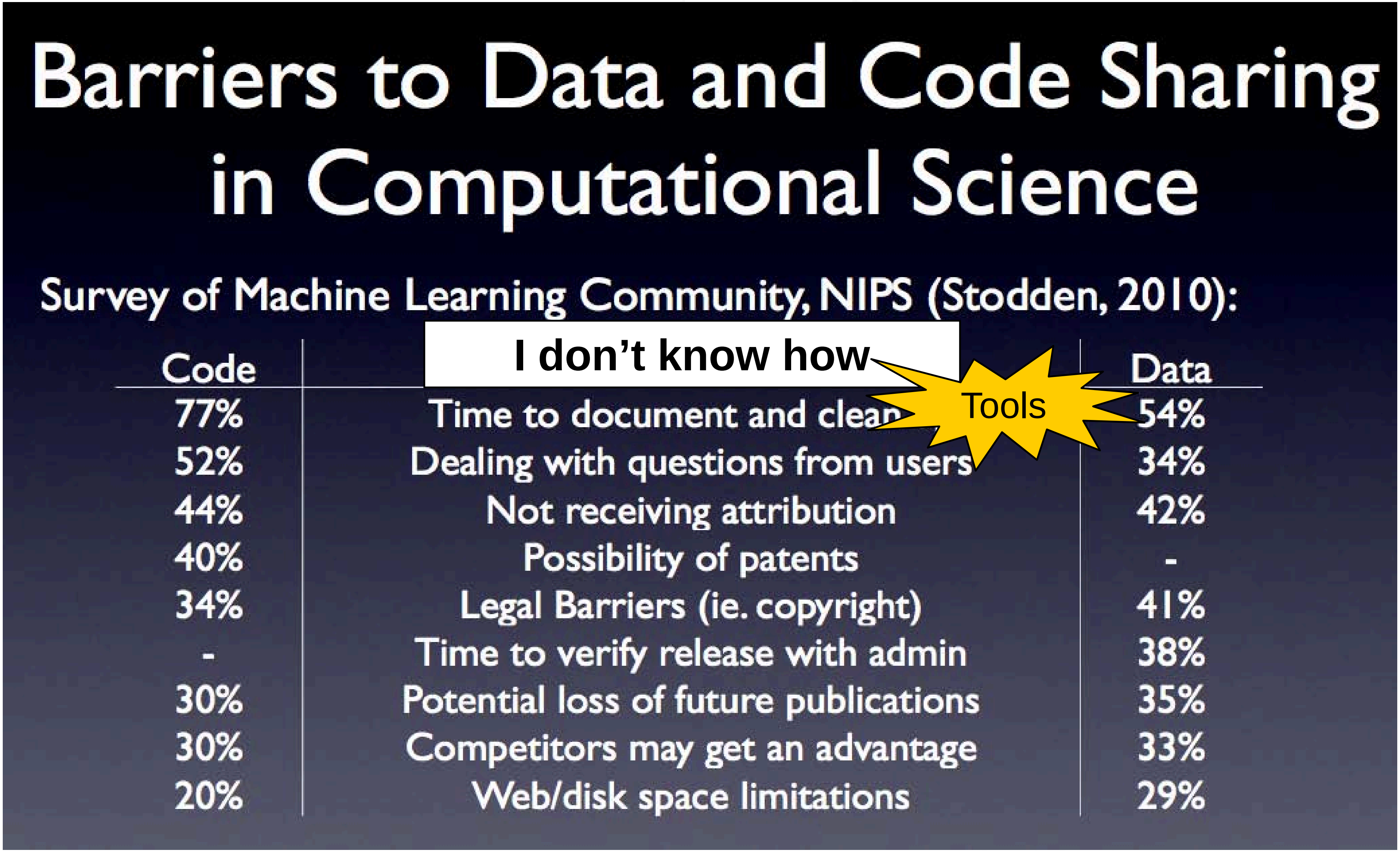
- Avoid duplication of efforts and reinvention
- How to discover and not duplicate ?
- How to re-use and not duplicate ?
- How to make use of best practices ?
- How to use the rich infrastructure of data ?
- **Intellectual contributions are encoded in soft**

More data in archives does not imply more knowledge

- Time has come to go beyond the PDF
- Expose complete scientific record, not the story
- Allow easy discovery of methods and tools

PDF
Adobe
IDL
FORTRAN
I R A F
Image Reduction and Analysis Facility
VOSPEC
esa
EXTRAGALACTIC DATABASE
NED
NASA IPAC
python™
IVOA

Reproducibility: documenting and sharing



Barriers to Data and Code Sharing in Computational Science

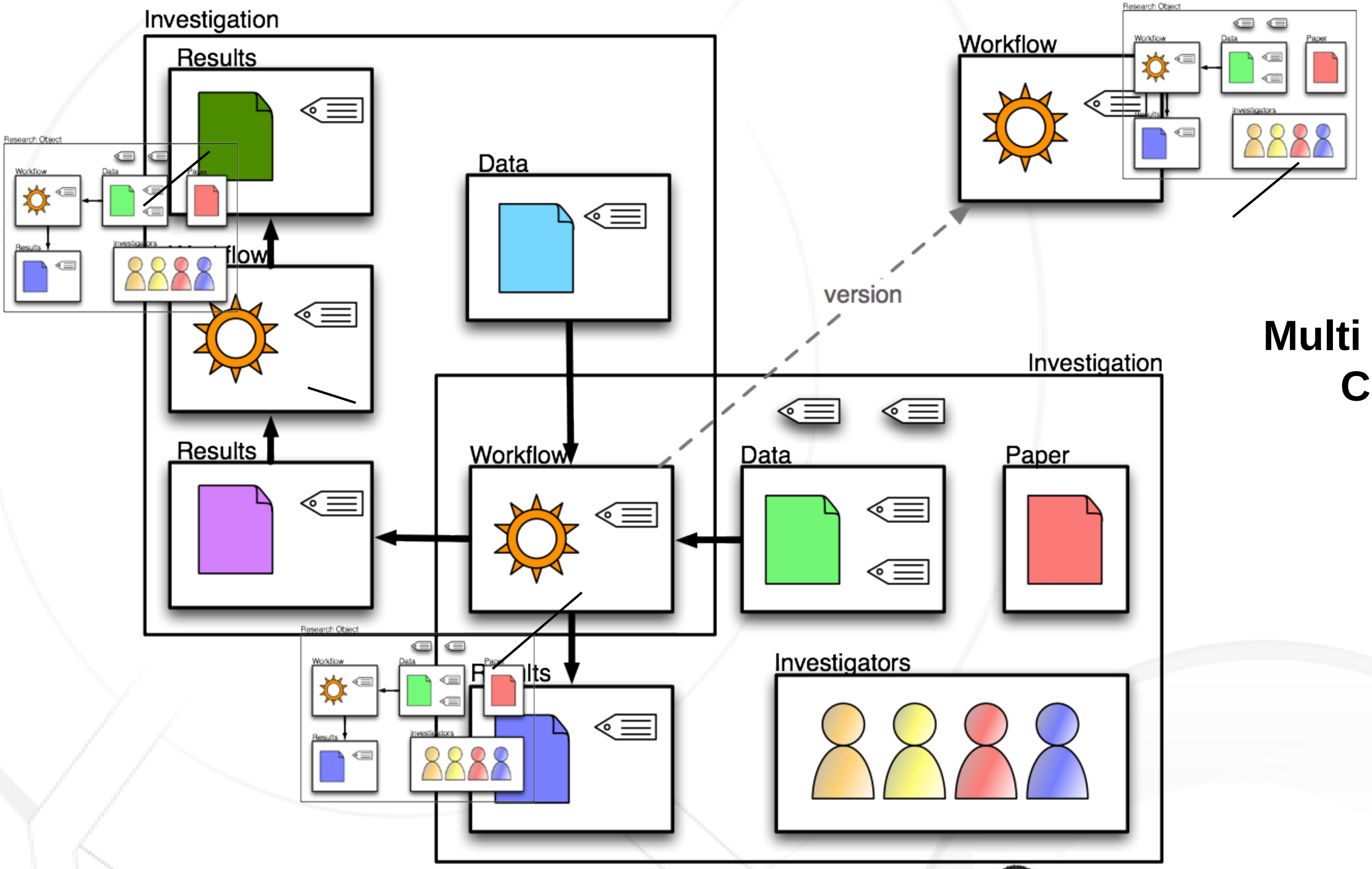
Survey of Machine Learning Community, NIPS (Stodden, 2010):

I don’t know how

Tools

Code		Data
77%	Time to document and clean	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

Research Objects in Wf4Ever



Multi Workflow Centric

Technical Objects
Distributed

Social Objects

RO Content

- Process (workflows), data, external resources and bibliography
- Execution environment set-up and local software dependencies
- Experimental protocol followed
- Roles, types and relationships among all digital components
- Provenance of intermediate and final results
- Decomposable attribution and authoring
- Fine-grained access control and permissions
- Example datasets for demonstration, reproducibility, monitoring, etc

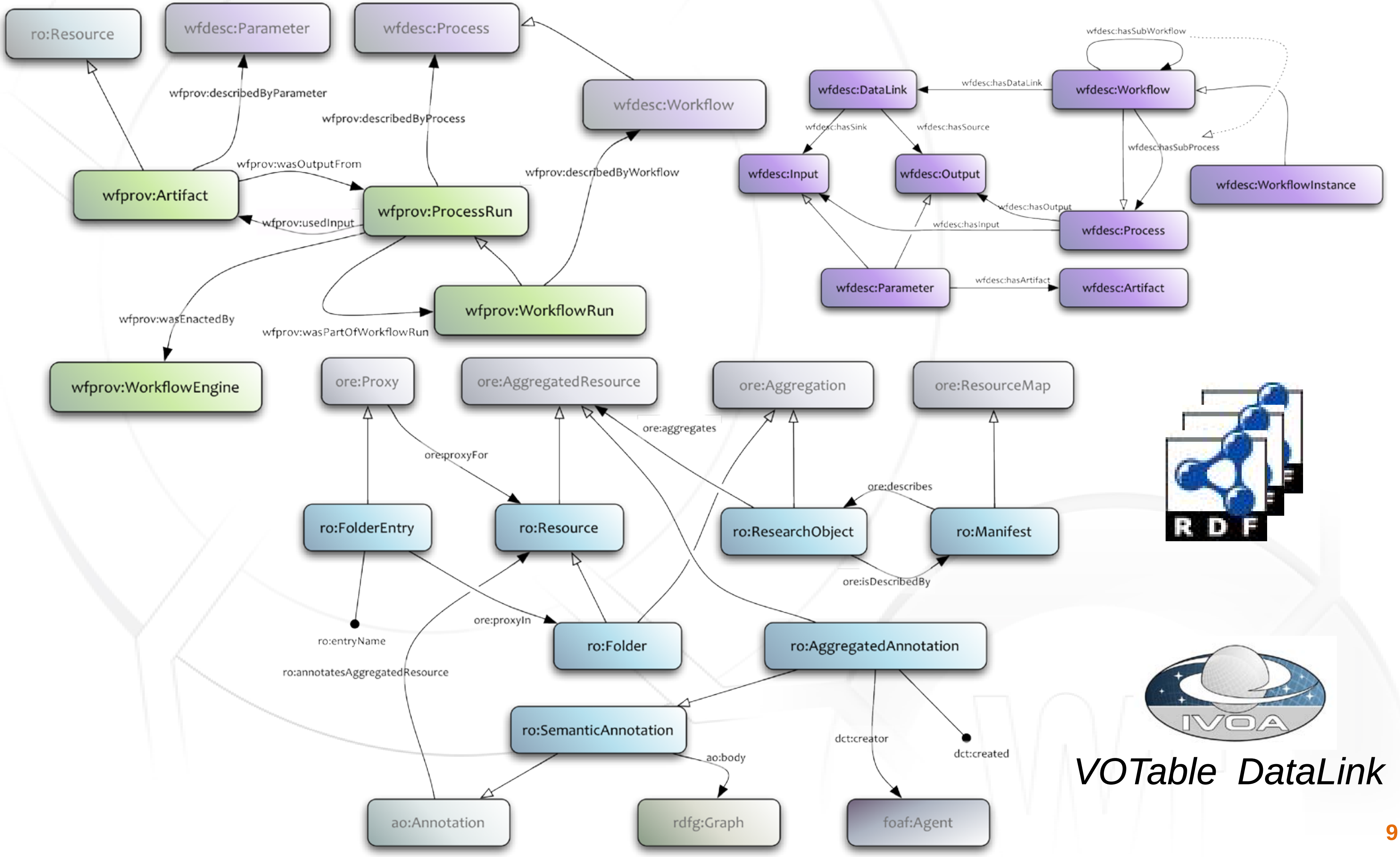
RO Template

- Placeholders to ease the aggregation process
- Completeness checking/quality assessment

Semantic Annotations

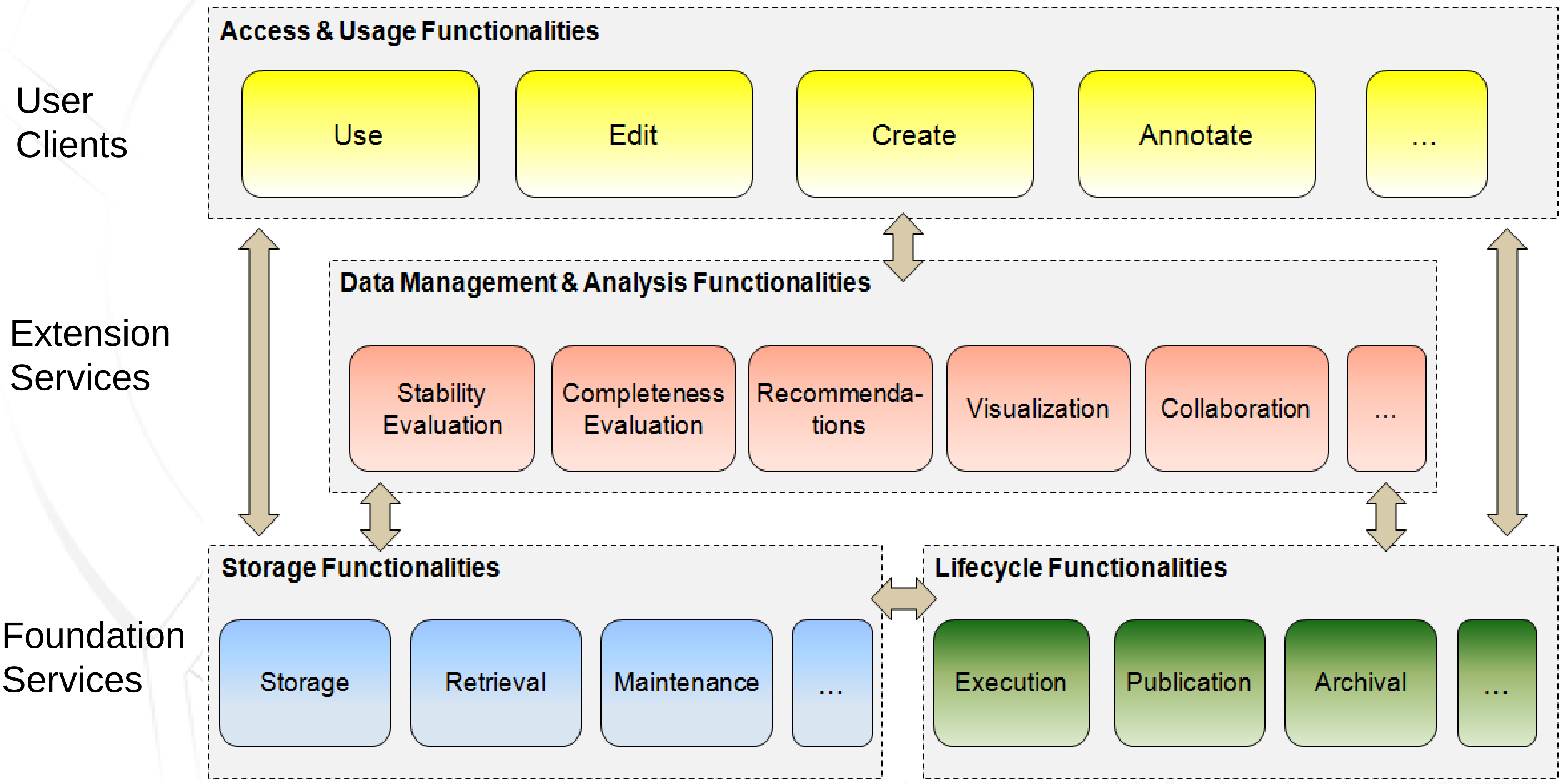
- Author of an **annotation**
- **Author and co-authors** of a **workflow**; reference link to a re-used workflow and its author
- Who has performed the **execution** of a workflow leading to the results provided in the RO
- Computing execution environment of the RO and local software **dependencies**
- Special **access requirements** to web services
- Datasets **provider**: person, webpage, survey, data release, etc.
- How much **time** does it take to run a workflow using the full data and the provided subsample
- The number of **elements** of the sample dataset where one workflow and/or RO iterates
- Previous and subsequent workflows to be executed, as in the experimental **protocol**
- Research institution, country, and scientific domain of the RO
- The actual **size** of the RO and/or a folder
- The **version** of a workflow

Research Object Wf4Ever Semantic Model

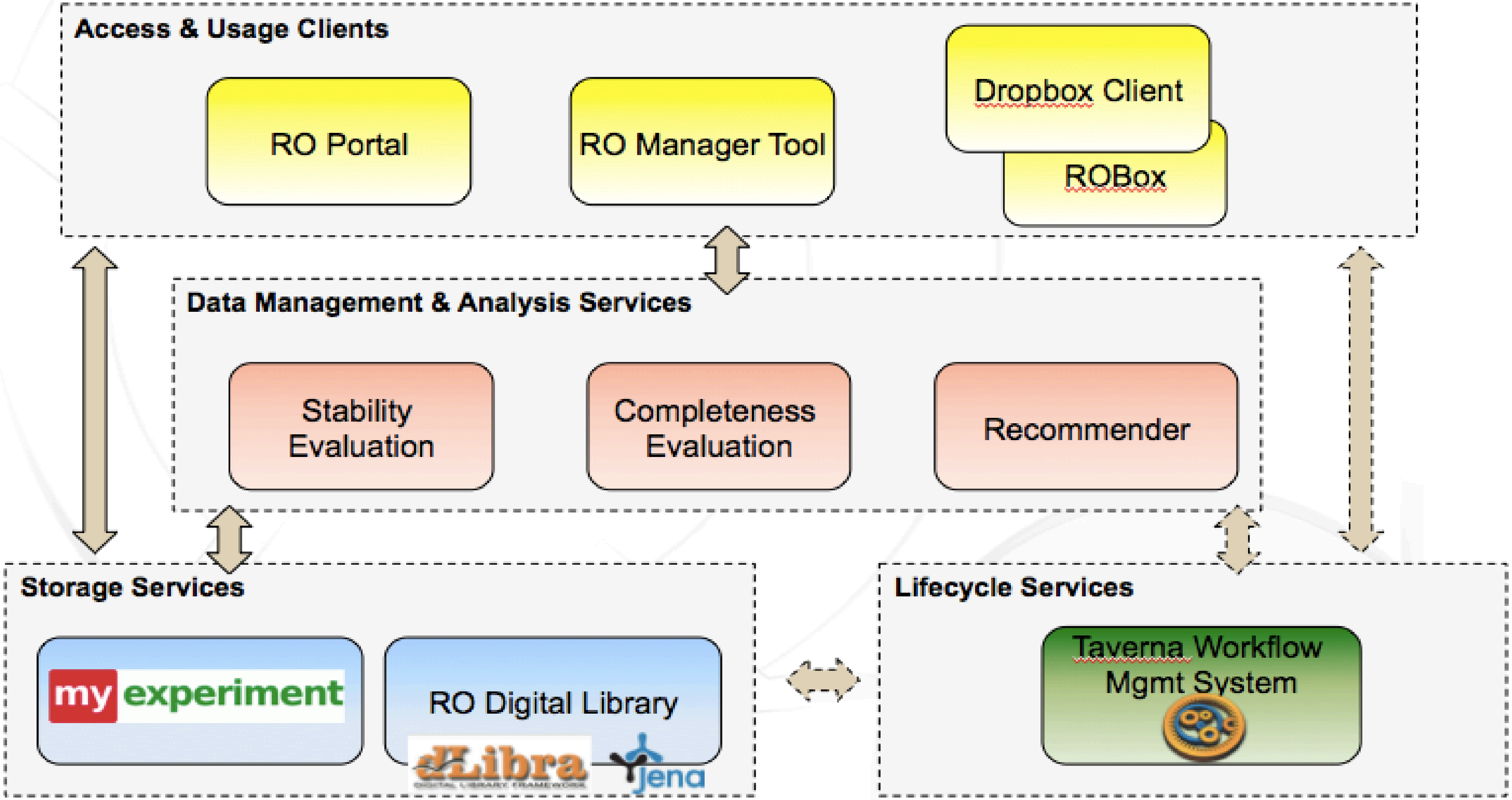


VOTable DataLink

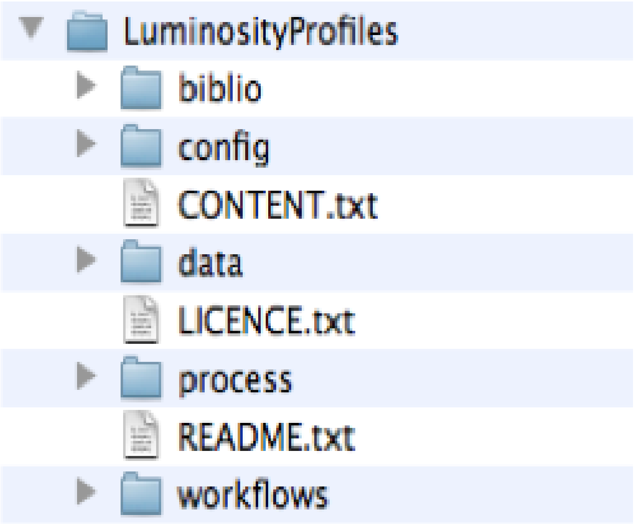
Research Object Digital Library Architecture



Research Object Digital Library Architecture



Luminosity Profiles RO



- LuminosityProfiles
  - biblio
  - config
  - CONTENT.txt
  - data
  - LICENCE.txt
  - process
  - README.txt
  - workflows

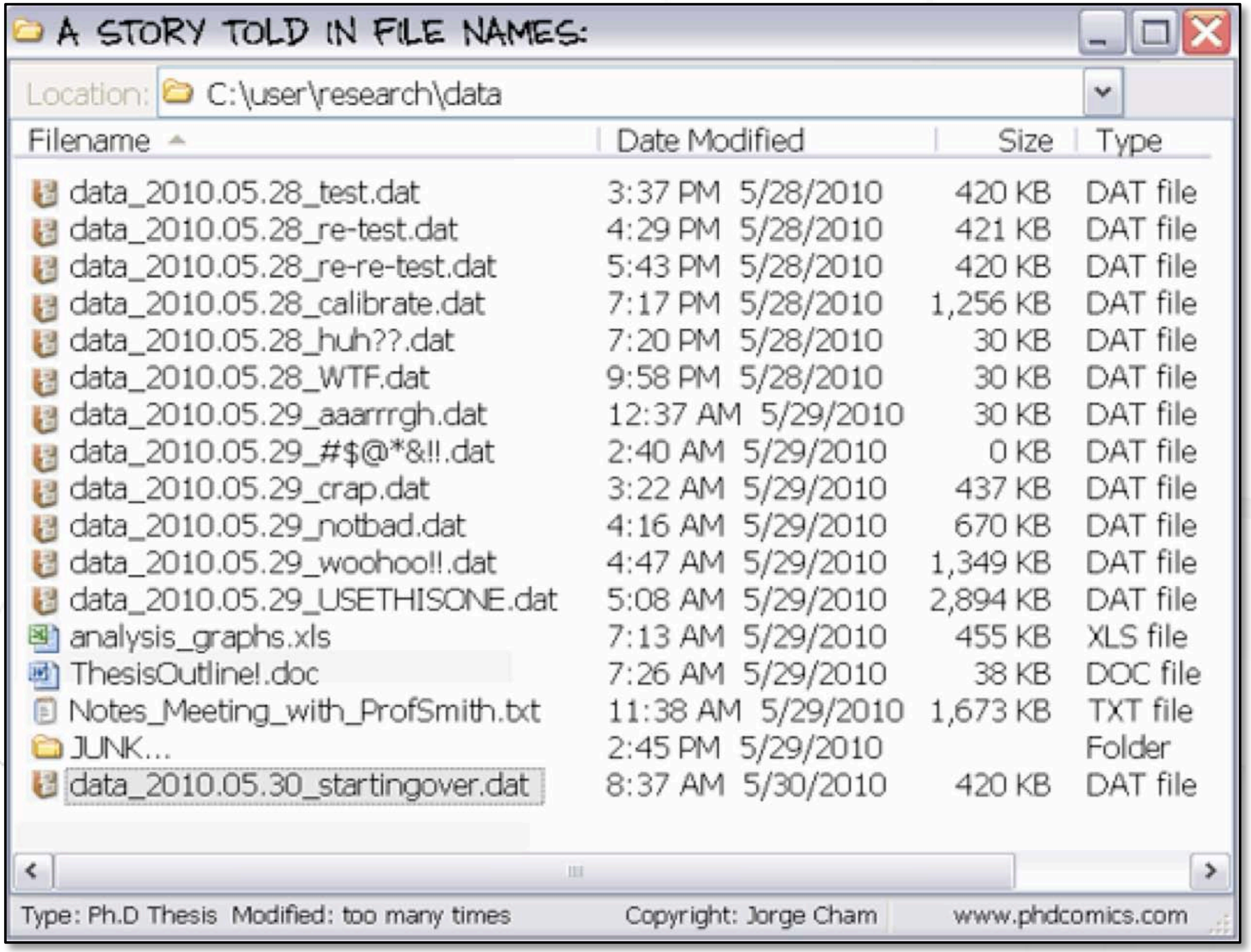
1010 Files, 200 MB
External Sources ~ 8 GB

5 Main Workflows, 14 Nested Workflows, 25 Scripts, 11 Configuration files
10 Software dependencies, 1 Web Service

Dataset: 90 galaxies observed in 3 bands

Reproducibility

When organization is better than automation



A STORY TOLD IN FILE NAMES:

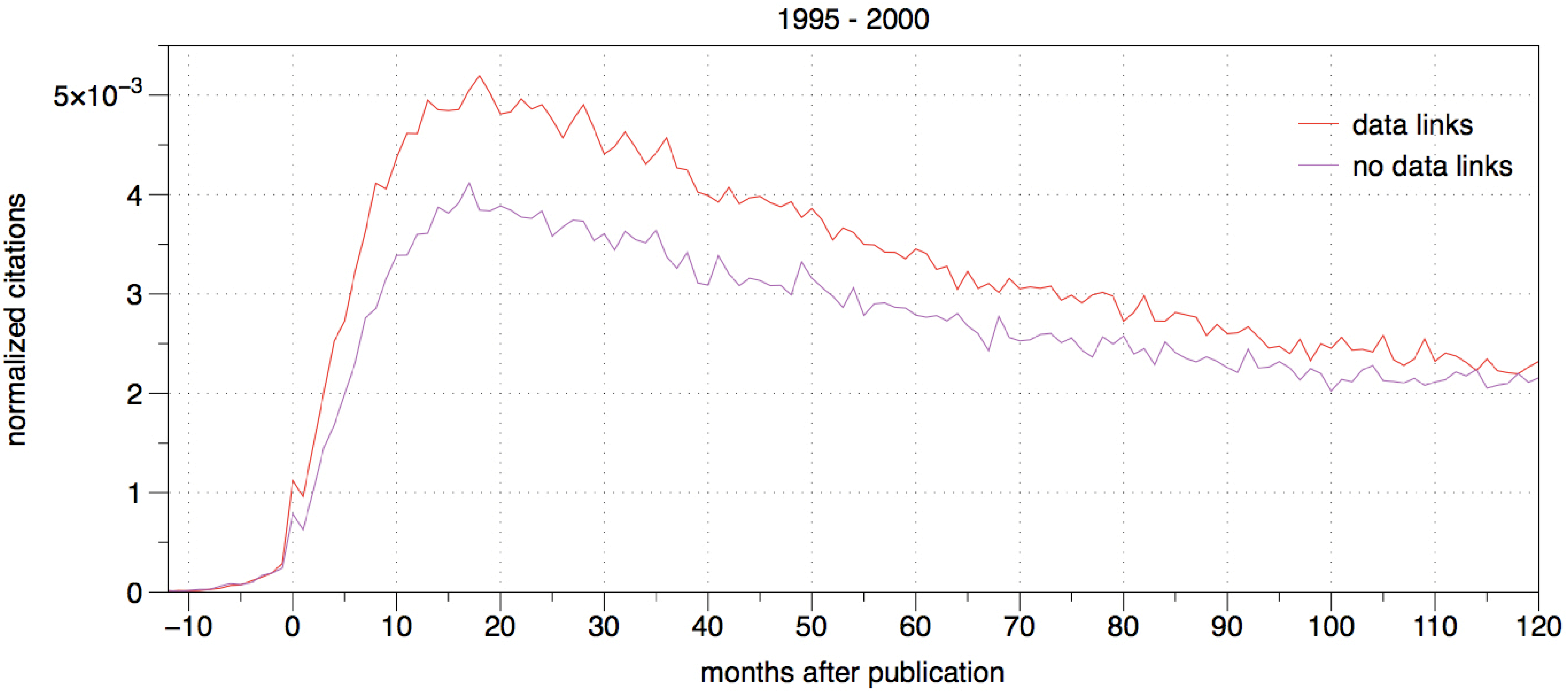
Location: C:\user\research\data

Filename	Date Modified	Size	Type
data_2010.05.28_test.dat	3:37 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_re-test.dat	4:29 PM 5/28/2010	421 KB	DAT file
data_2010.05.28_re-re-test.dat	5:43 PM 5/28/2010	420 KB	DAT file
data_2010.05.28_calibrate.dat	7:17 PM 5/28/2010	1,256 KB	DAT file
data_2010.05.28_huh??.dat	7:20 PM 5/28/2010	30 KB	DAT file
data_2010.05.28_WTF.dat	9:58 PM 5/28/2010	30 KB	DAT file
data_2010.05.29_aaarrrgh.dat	12:37 AM 5/29/2010	30 KB	DAT file
data_2010.05.29_#$@*&!!.dat	2:40 AM 5/29/2010	0 KB	DAT file
data_2010.05.29_crap.dat	3:22 AM 5/29/2010	437 KB	DAT file
data_2010.05.29_notbad.dat	4:16 AM 5/29/2010	670 KB	DAT file
data_2010.05.29_woohoo!!.dat	4:47 AM 5/29/2010	1,349 KB	DAT file
data_2010.05.29_USETHISONE.dat	5:08 AM 5/29/2010	2,894 KB	DAT file
analysis_graphs.xls	7:13 AM 5/29/2010	455 KB	XLS file
ThesisOutline!.doc	7:26 AM 5/29/2010	38 KB	DOC file
Notes_Meeting_with_ProfSmith.txt	11:38 AM 5/29/2010	1,673 KB	TXT file
JUNK...	2:45 PM 5/29/2010		Folder
data_2010.05.30_startingover.dat	8:37 AM 5/30/2010	420 KB	DAT file

Type: Ph.D Thesis Modified: too many times Copyright: Jorge Cham www.phdcomics.com

Credit and attribution

Papers with data links are cited more than those without

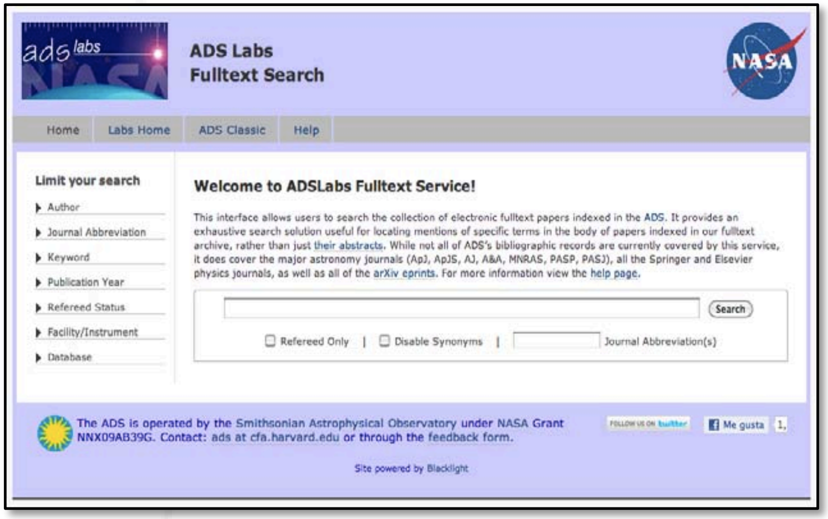


Effect of E-printing on Citation Rates in Astronomy and Physics
2006. Edwin A. Henneken et al.

ADSLabs Research Objects

ADO Linked Components

- » Authors
- » Publications
- » Journals
- » Objects SIMBAD
- » Tabular data behind the plots CDS
- » ASCL reference of used software
- » Observing time Proposals
- » Used facilities, surveys or missions



ADS Labs
Fulltext Search

Limit your search

- Author
- Journal Abbreviation
- Keyword
- Publication Year
- Refereed Status
- Facility/Instrument
- Database

Welcome to ADSLabs Fulltext Service!

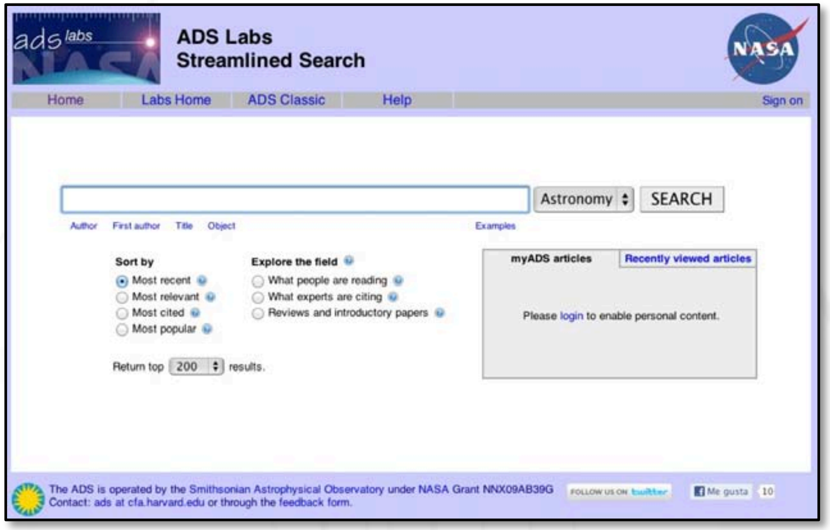
This interface allows users to search the collection of electronic fulltext papers indexed in the ADS. It provides an exhaustive search solution useful for locating mentions of specific terms in the body of papers indexed in our fulltext archive, rather than just their abstracts. While not all of ADS's bibliographic records are currently covered by this service, it does cover the major astronomy journals (ApJ, ApJS, AJ, A&A, MNRAS, PASP, PASJ), all the Springer and Elsevier physics journals, as well as all of the arXiv eprints. For more information view the help page.

□ Refereed Only | □ Disable Synonyms | Journal Abbreviation(s)

The ADS is operated by the Smithsonian Astrophysical Observatory under NASA Grant NNX09AB39G. Contact: ads at cfa.harvard.edu or through the feedback form.

FOLLOW US ON twitter Me gusta 1,

Site powered by Blacklight



ADS Labs
Streamlined Search

Astronomy

Sort by
- ● Most recent
- ○ Most relevant
- ○ Most cited
- ○ Most popular

Explore the field
- ○ What people are reading
- ○ What experts are citing
- ○ Reviews and introductory papers

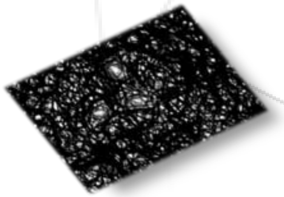
Return top 200 results.

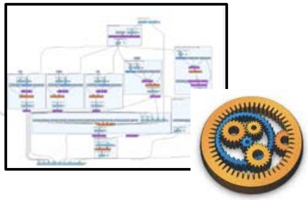
myADS articles | Recently viewed articles

Please login to enable personal content.

The ADS is operated by the Smithsonian Astrophysical Observatory under NASA Grant NNX09AB39G
Contact: ads at cfa.harvard.edu or through the feedback form.

FOLLOW US ON twitter Me gusta 10





The next generation of archives

Much wider FoV and spectral coverage

- Huge sized datasets (~ tens TB)
- Big Data science highly dependent on I/O data rates
- Subproducts as virtual data generated on-the-fly

We are moving into a world where

- **computing and storage are cheap**
- **data movement is death**

The next generation of archives

Much wider FoV and spectral coverage

- Huge sized datasets (~ tens TB)
- Big Data science highly dependent on I/O data rates
- Subproducts as virtual data generated on-the-fly

The *move computing to data* paradigm

Archives should evolve from data providers into **services providers**, where web services may help to solve bandwidth issues.

The next generation of archives

Much wider FoV and spectral coverage

- Huge sized datasets (~ tens TB)
- Big Data science highly dependent on I/O data rates
- Subproducts as virtual data generated on-the-fly

Data Discovery

Data Access

Data Management

The next generation of archives

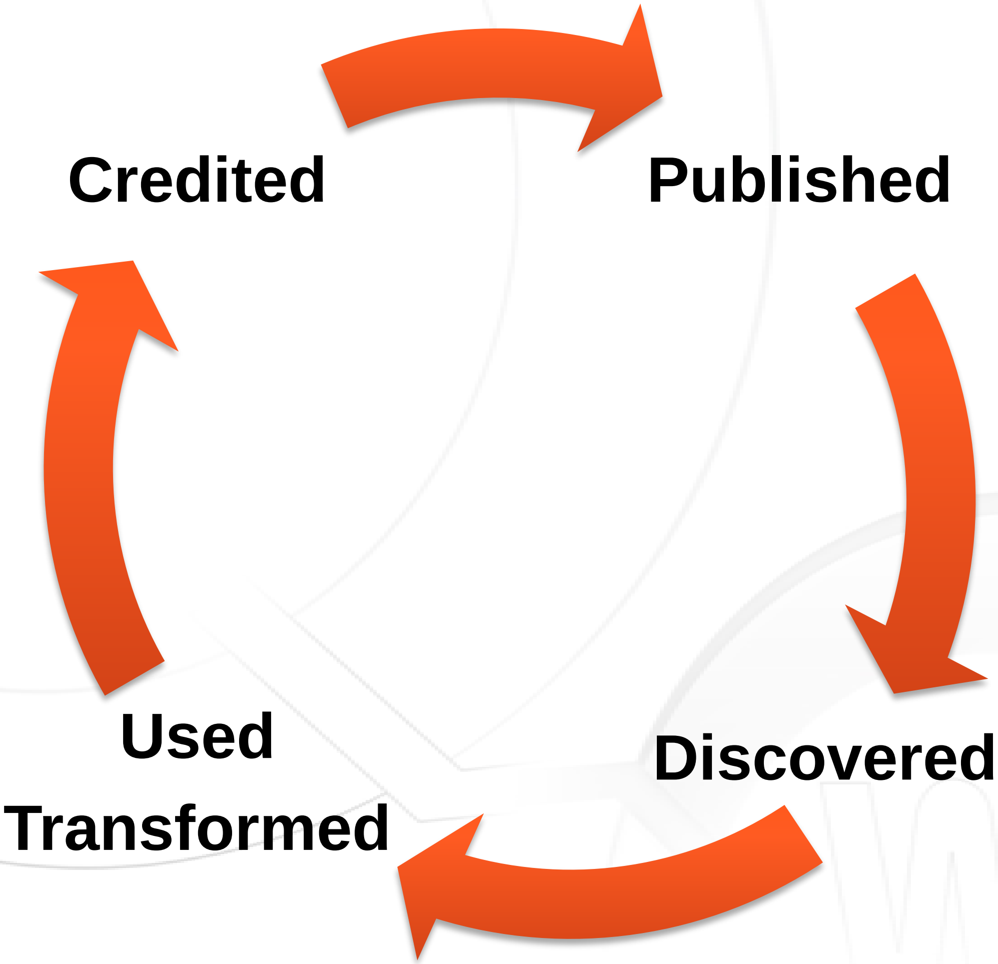
Much wider FoV and spectral coverage

- Huge sized datasets (~ tens TB)
- Big Data science highly dependent on I/O data rates
- Subproducts as virtual data generated on-the-fly

Web Services Discovery
Web Services Access
Web Services Management

The Web-Services Deluge

The Data Lifecycle



Credited

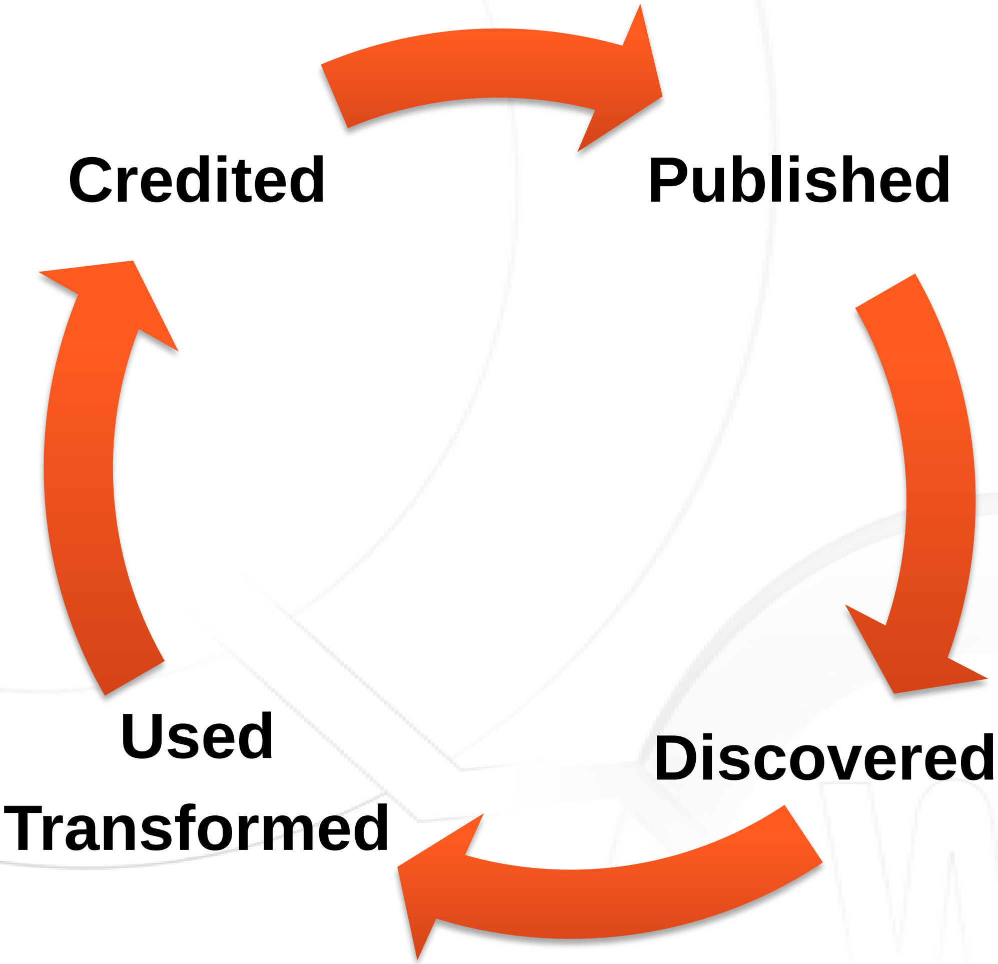
Published

Discovered

Used

Transformed

The Service Lifecycle



Credited

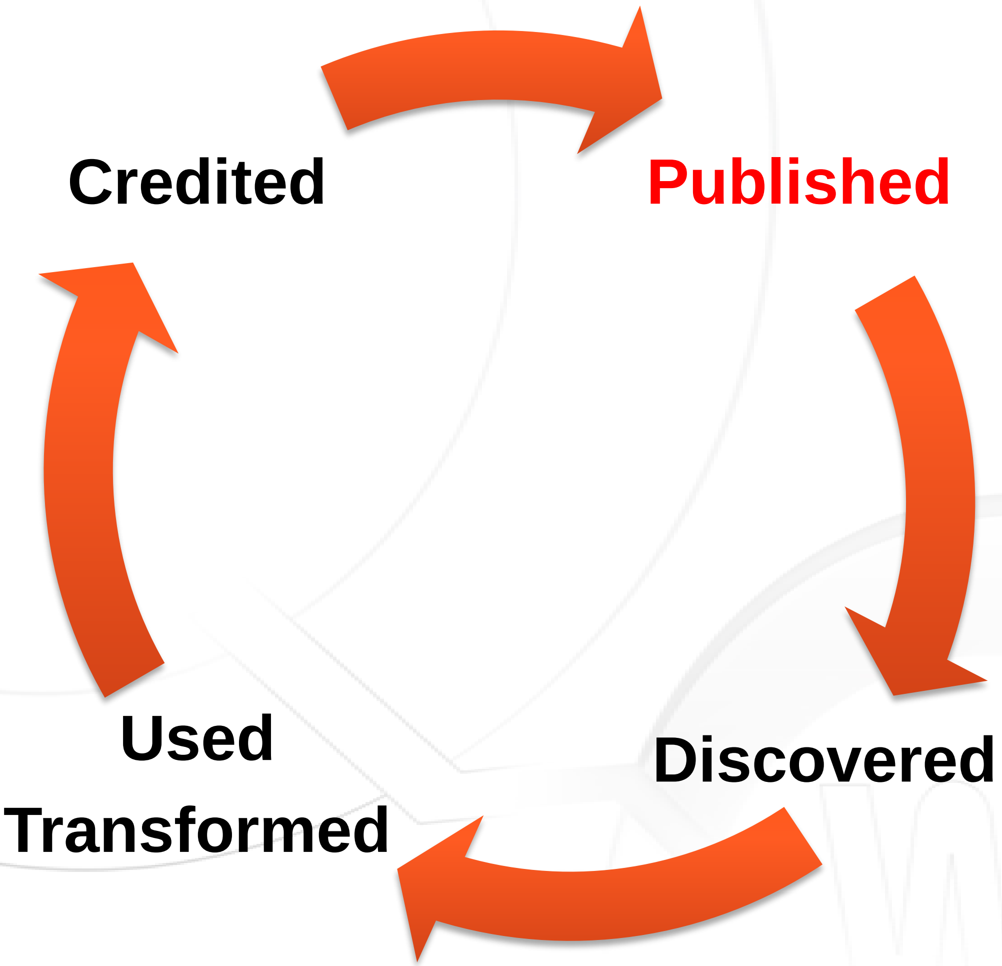
Published

Discovered

Used

Transformed

The Service Lifecycle



Published

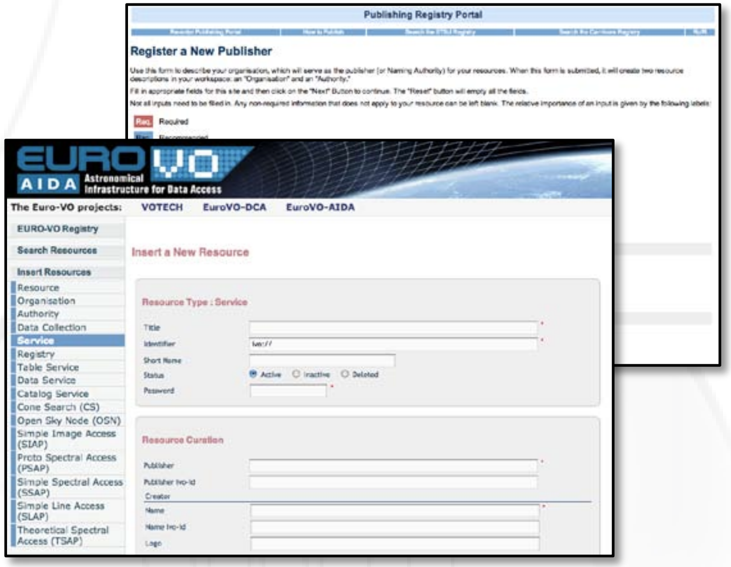
- The VO Registry
- Easier to publish services than datasets in the VO ?
- WS are not exclusive property of big data archives

- Publication is not Preservation
- Backup strategies
- Replication/Mirrors
- Versioning

- Software Publishing Platforms

github

runmycode


Publishing Registry Portal

Register a New Publisher

Use this form to describe your organisation, which will serve as the publisher (or Naming Authority) for your resources. When this form is submitted, it will create two resource descriptions in your workspace: an "Organisation" and an "Authority."

Fill in appropriate fields for this site and then click on the "Next" Button to continue. The "Reset" button will empty all the fields.

Not all inputs need to be filled in. Any non-required information that does not apply to your resource can be left blank. The relative importance of an input is given by the following labels:

Req. Required

Insert a New Resource

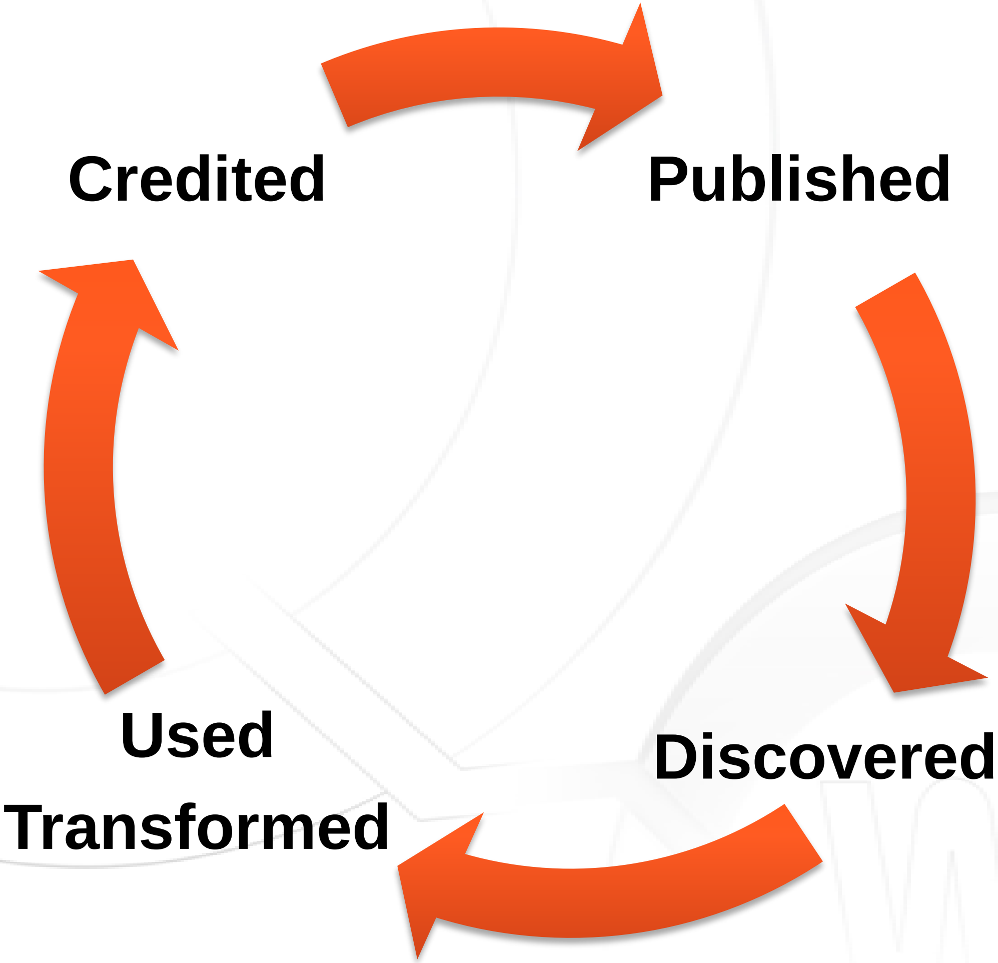
Resource Type : Service

Title
Identifier ivo://
Short Name
Status ● Active ○ Inactive ○ Deleted
Password

Resource Curation

Publisher
Publisher Ivo-Id
Creator
Name
Name Ivo-Id
Logo

The Service Lifecycle



Credited

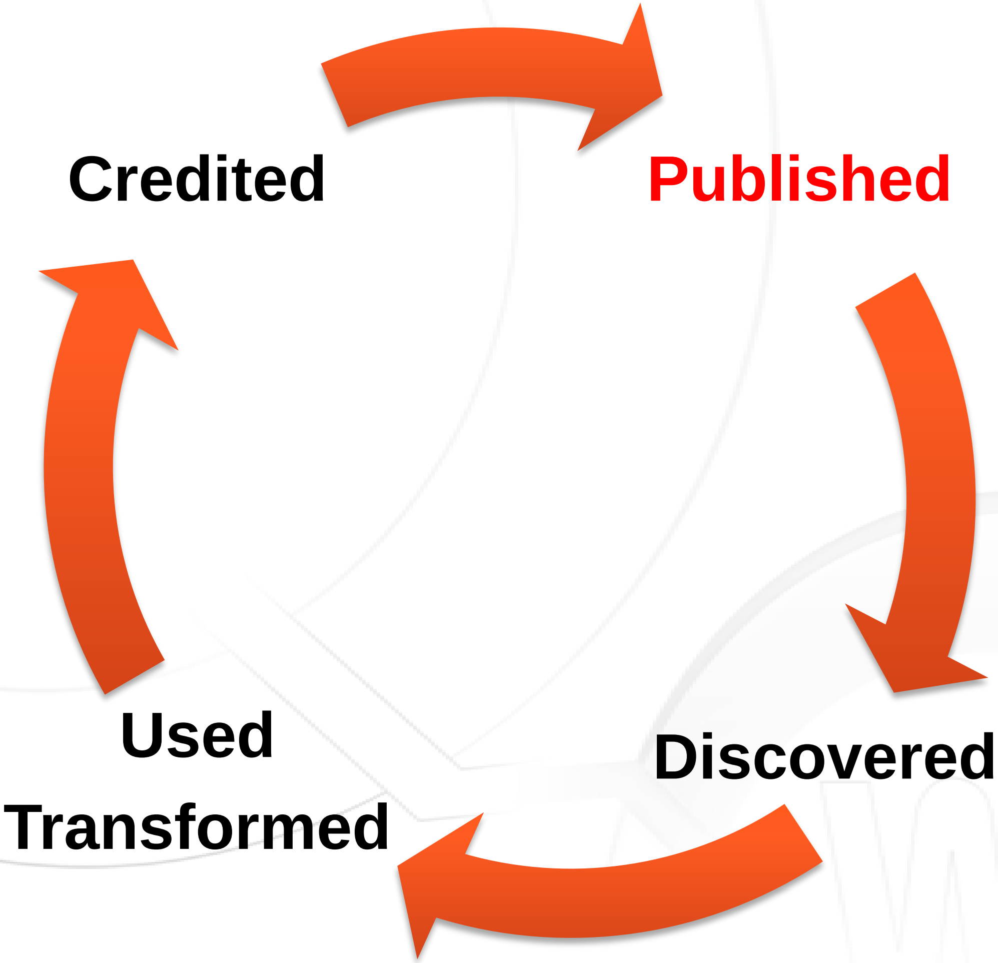
Published

Discovered

Used

Transformed

The Service Lifecycle



Credited

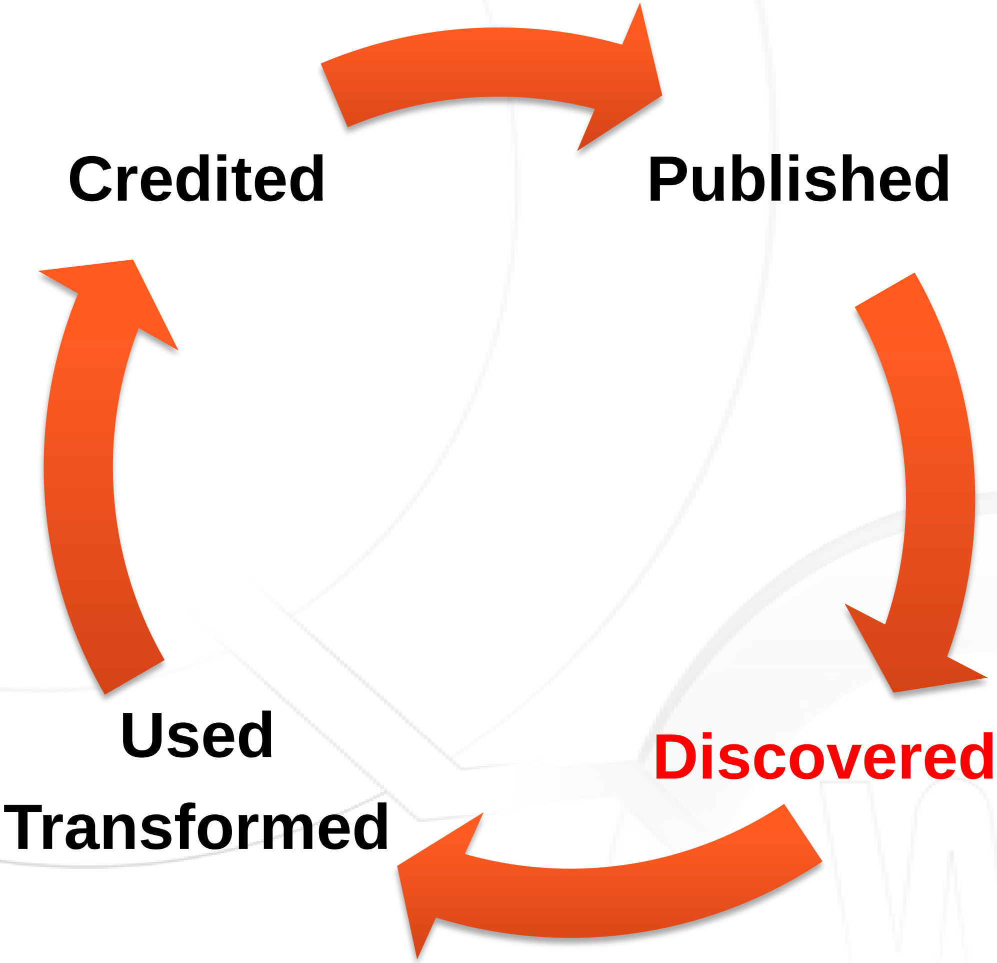
Published

Discovered

Used

Transformed

The Service Lifecycle



Credited

Published

Discovered

Used

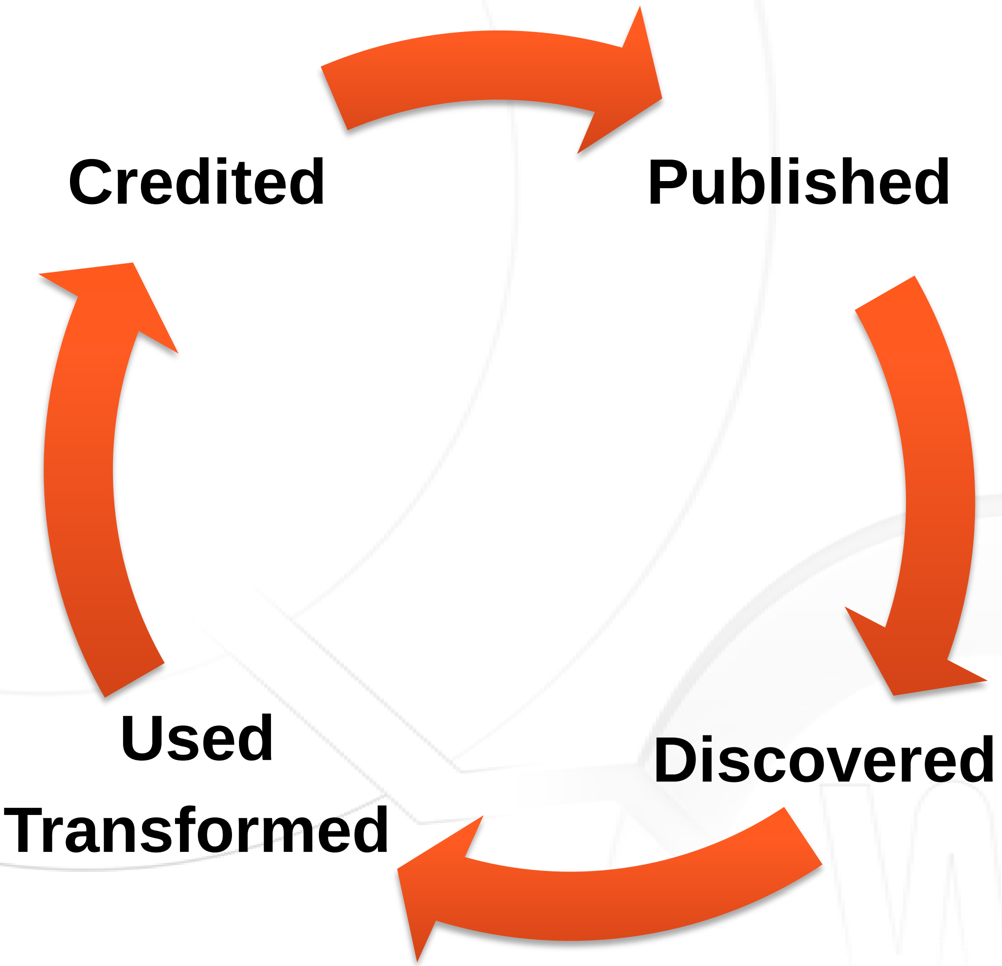
Transformed

Discovered

- **Search Criteria**
  - Relevant Keywords (Semantics)
  - Authoring - Institution, Archive
  - Waveband, Science
  - Function-based
    - VO Services mainly focused on Data Discovery and Access (DAL)
    - Wrapped Legacy Apps and Data Processing (SIAv2, Theory IG)
    - KDD IG
  - Input/Output Data (TAP, UTypes, VOSI #tables)
  - Access Policy (Authentication – SSO, OAuth)
  - A-Synchrony (SOAP, REST) and Stage Data (VOSpace)
  - Allocation of CPU/Storage, Estimated Computing Time

- **Proposition of alternatives and similars**

The Service Lifecycle



Credited

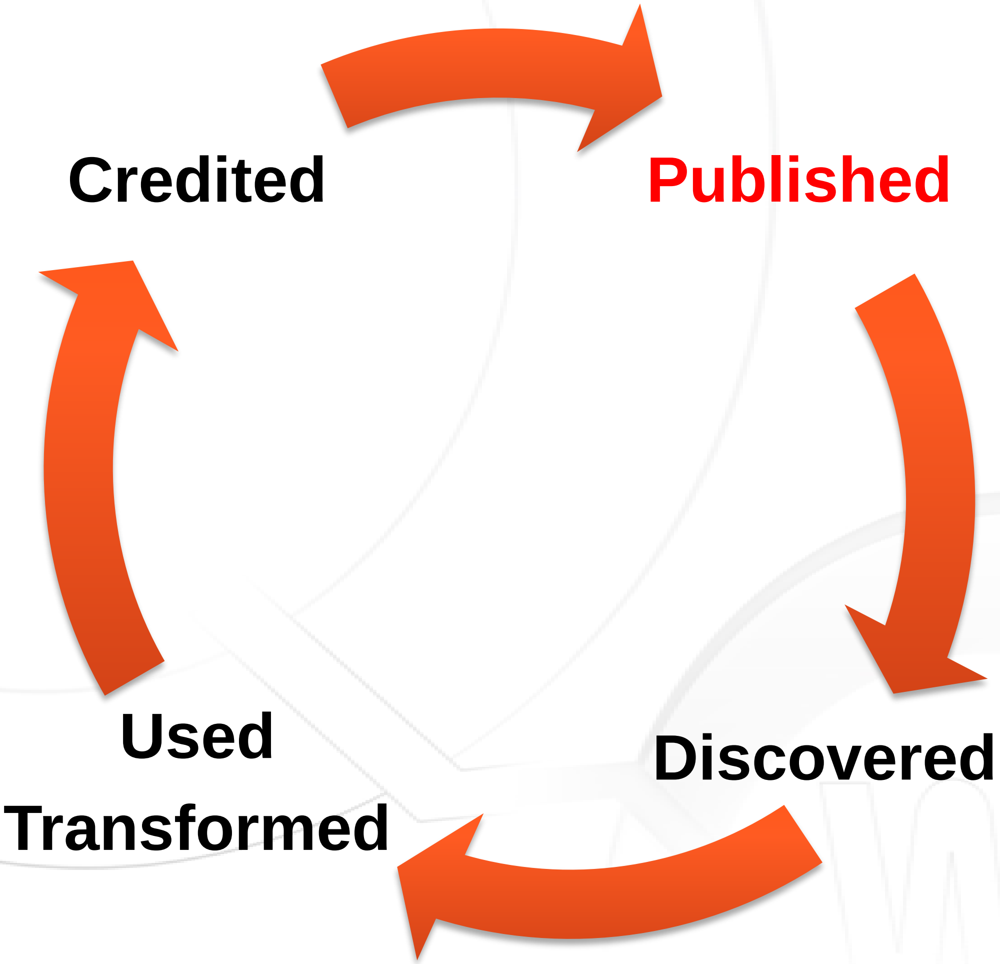
Published

Discovered

Used

Transformed

The Service Lifecycle



Credited

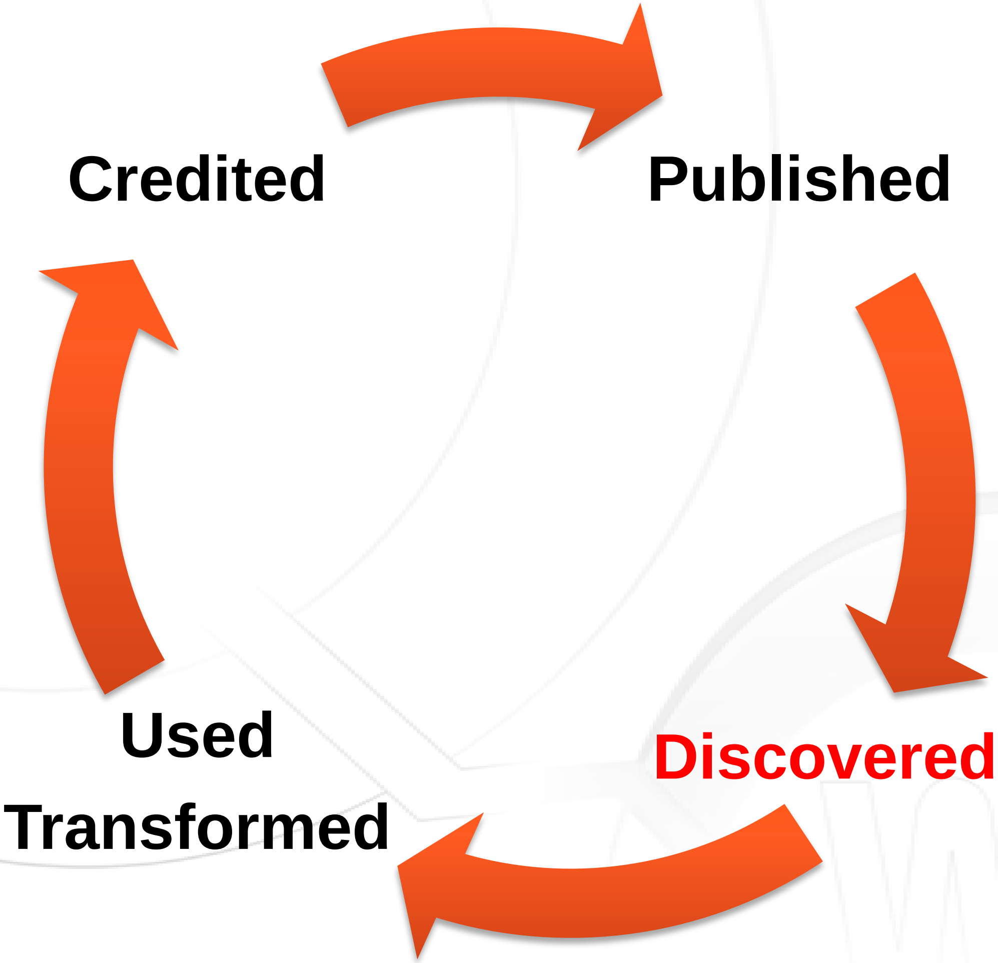
Published

Discovered

Used

Transformed

The Service Lifecycle



Credited

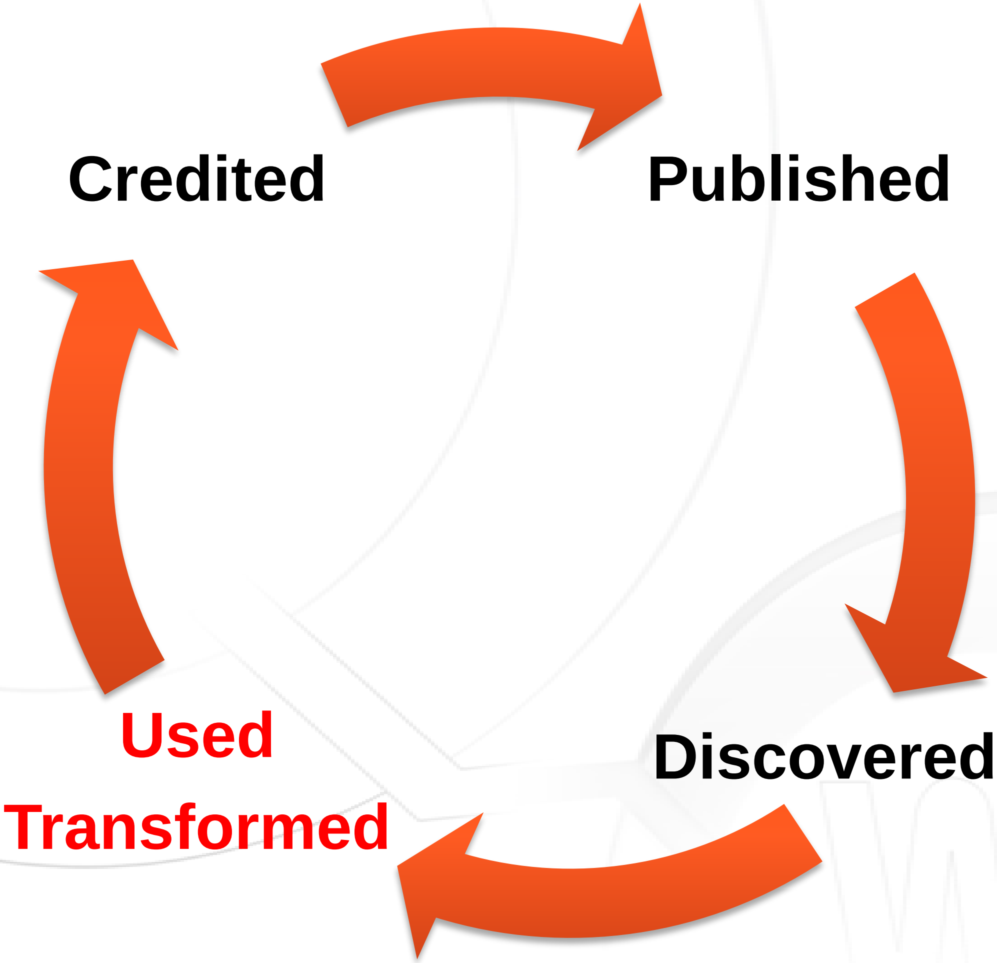
Published

Discovered

Used

Transformed

The Service Lifecycle



Credited

Published

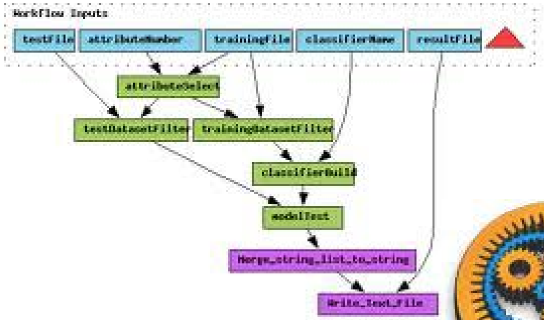
Discovered

Used

Transformed

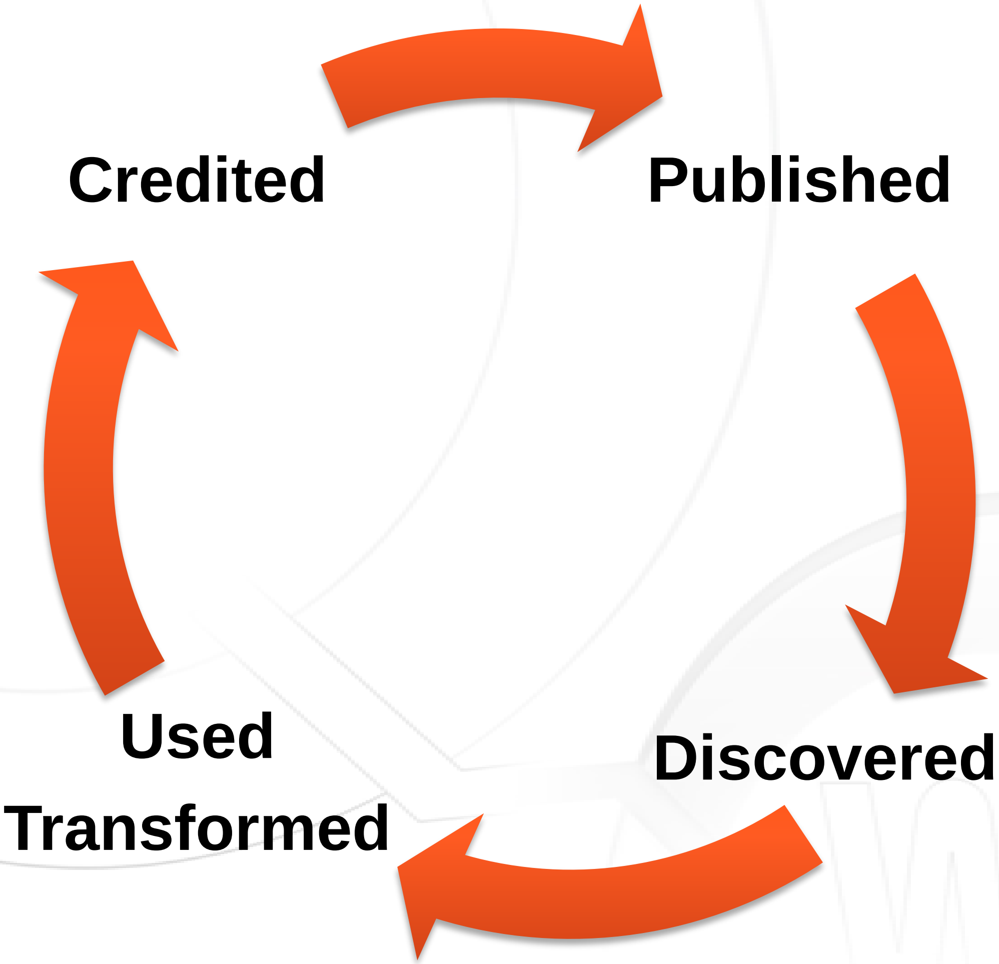
Used and Transformed

- **How to use them ?** (WADL, WSDL – VOSI #capabilities)
  - Input Data -> Parameters needed and formats
  - Self-described WS (PDL, S3, SimDAL, SimDB)
  - Output Data -> Response format - TAP
  - Example Data, Self-Consistency Checking



- Access Policy (**Authentication** – SSO, OAuth)
- **WS orchestration in Workflows** (Data-flow vs. Control-flow)
- How the **community** uses WS ?
- Propositions based on patterns of statistical use or popularity
- **Provenance** of the methods is Wf-evolution by re-use
- Consumed by Humans and Machines - **Interoperable** (WS-I)

The Service Lifecycle



Credited

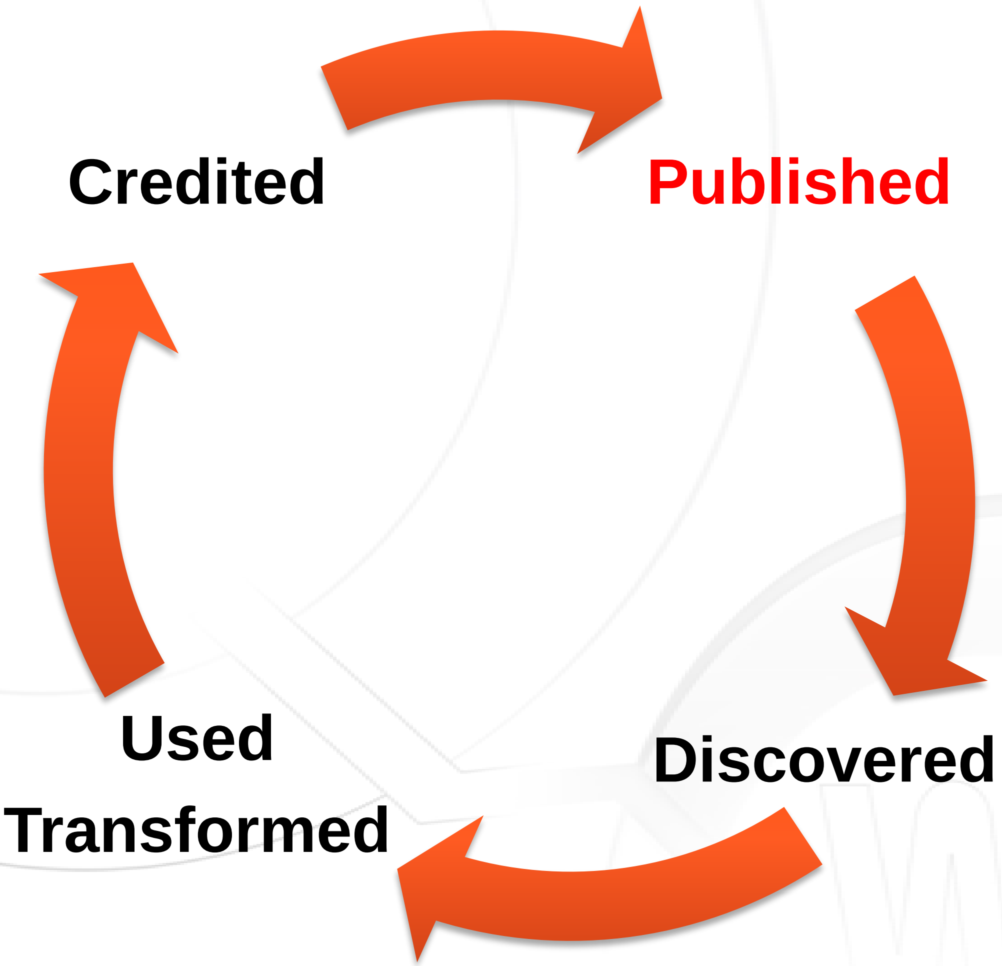
Published

Discovered

Used

Transformed

The Service Lifecycle



Credited

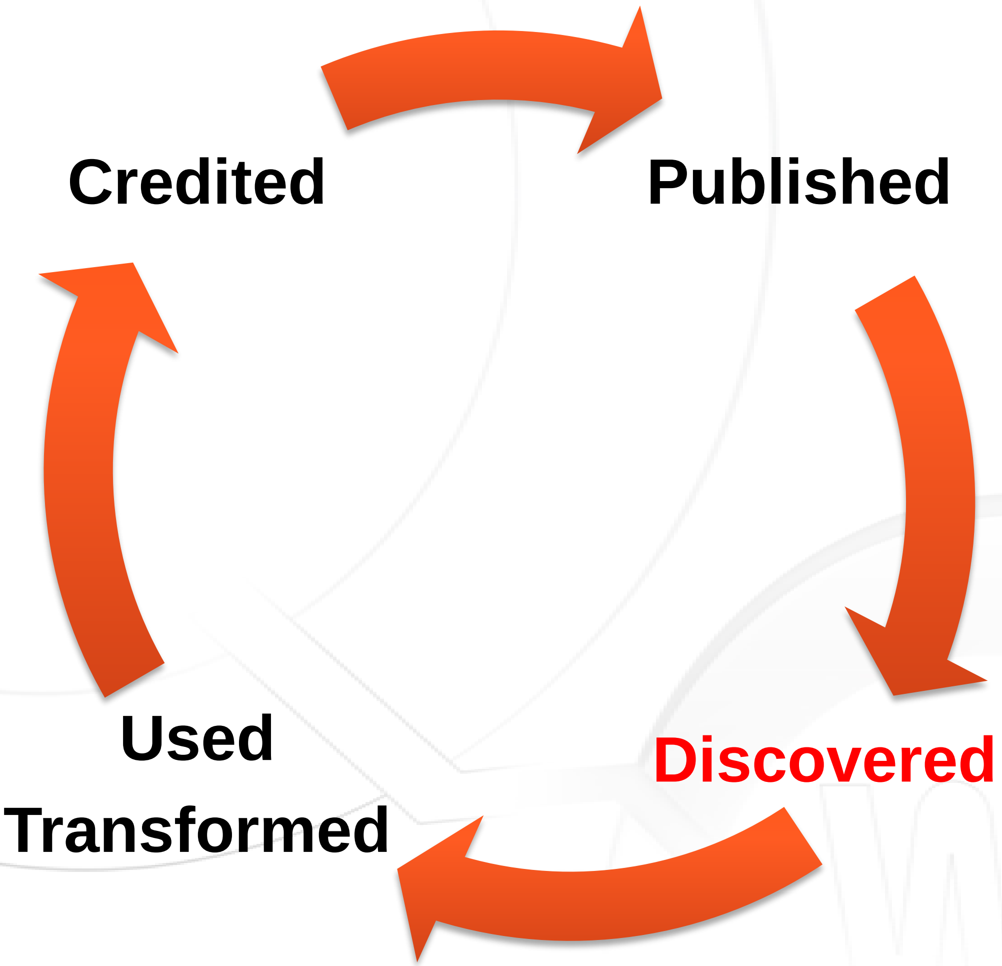
Published

Discovered

Used

Transformed

The Service Lifecycle



Credited

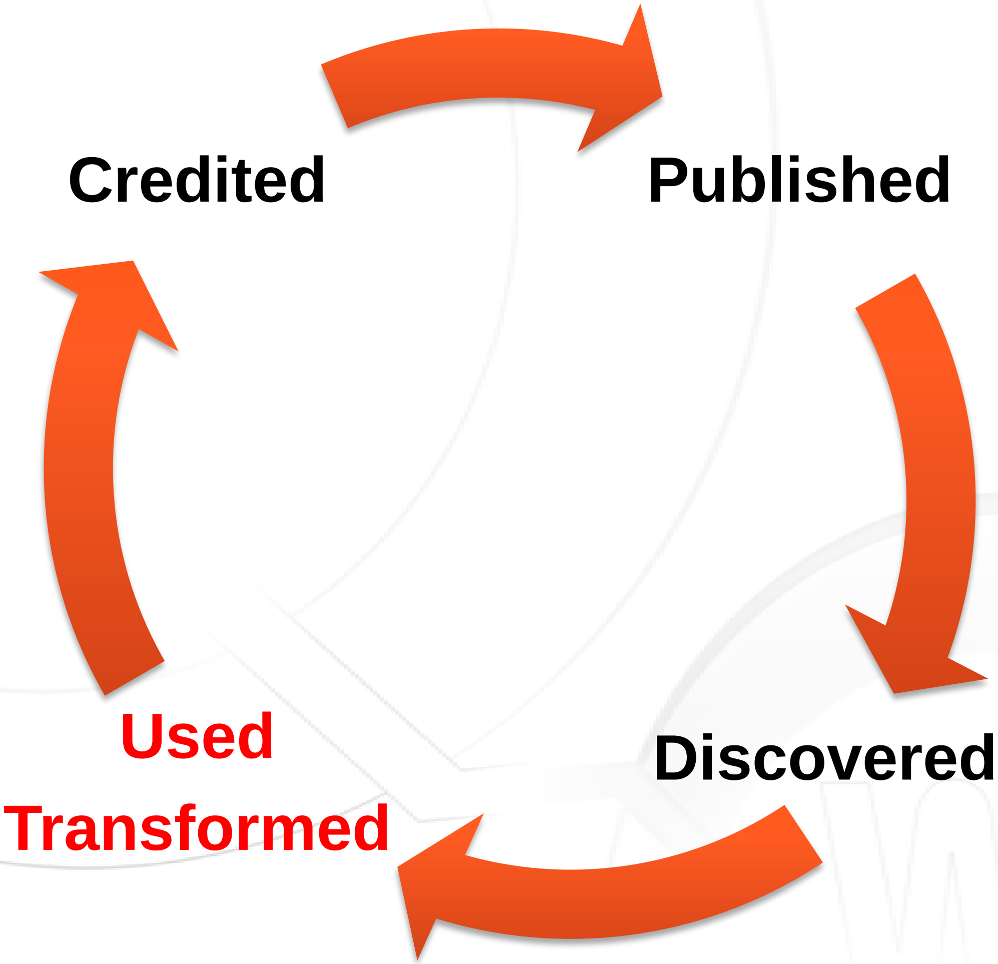
Published

Discovered

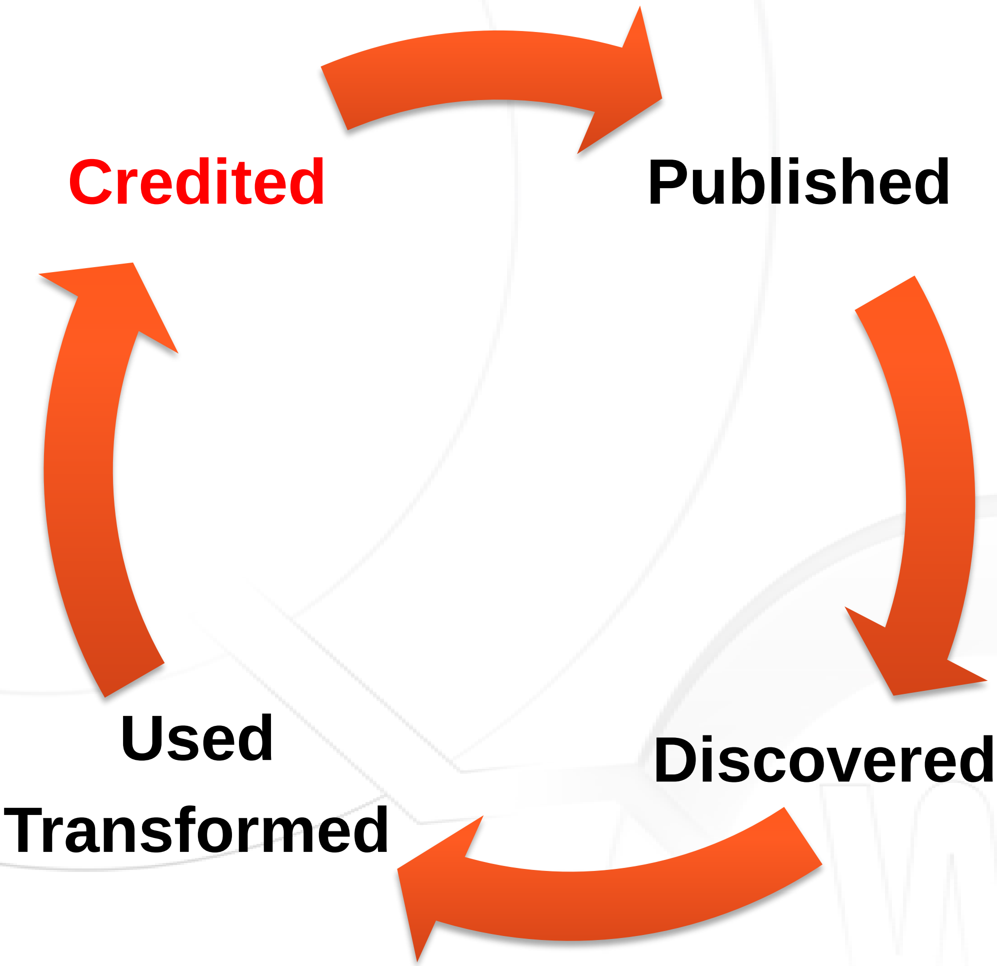
Used

Transformed

The Service Lifecycle

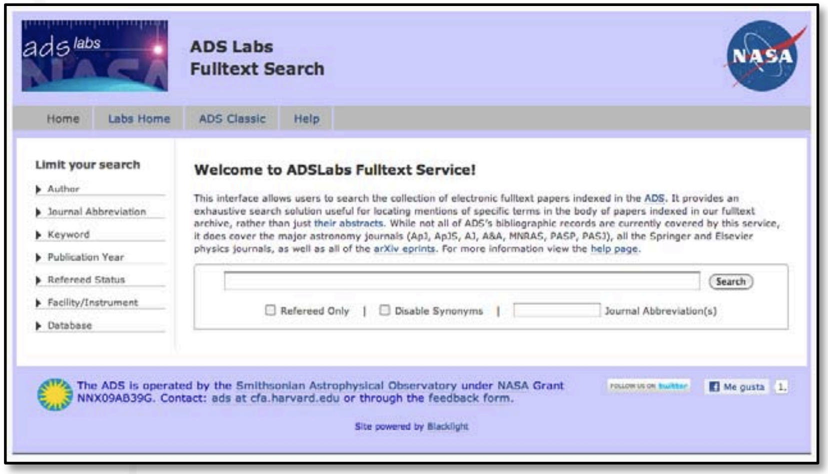


The Service Lifecycle



Credited

- **Linked to related Artefacts**
  - Data Facilities and Archives
  - Authors, ASCL Software, Wfs
- **Quality Assessment**
  - Technical and scientific
  - Penalize abandoned and award the maintained
- **Automate Monitoring (VOSI #availability)**
  - Decay
  - Performance, WS Analytics
  - Modifs. on interfaces, permissions, etc.
- **Community Curation**
  - Blogging
  - Recommendation
  - Folksonomy



ADS Labs
Fulltext Search

Limit your search
- Author
- Journal Abbreviation
- Keyword
- Publication Year
- Refereed Status
- Facility/Instrument
- Database

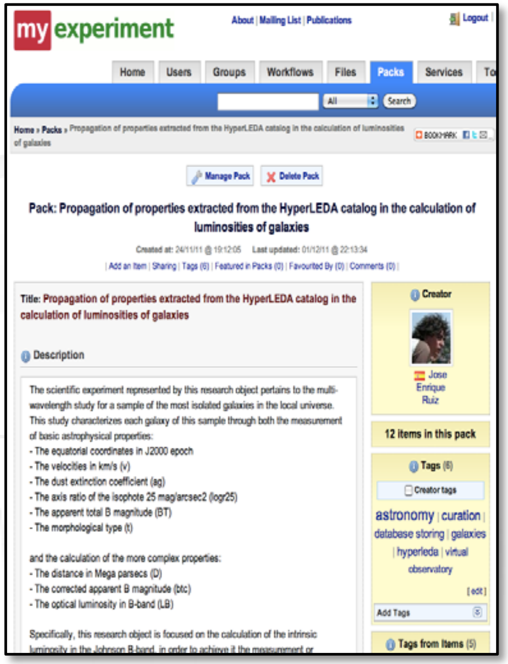
Welcome to ADSLabs Fulltext Service!

This interface allows users to search the collection of electronic fulltext papers indexed in the ADS. It provides an exhaustive search solution useful for locating mentions of specific terms in the body of papers indexed in our fulltext archive, rather than just their abstracts. While not all of ADS's bibliographic records are currently covered by this service, it does cover the major astronomy journals (ApJ, ApJS, AJ, A&A, MNRAS, PASP, PASJ), all the Springer and Elsevier physics journals, as well as all of the arXiv eprints. For more information view the help page.

□ Refereed Only | □ Disable Synonyms | Journal Abbreviation(s)

The ADS is operated by the Smithsonian Astrophysical Observatory under NASA Grant NNX09AB39G. Contact: ads at cfa.harvard.edu or through the feedback form.

Site powered by Blacklight



Home » Packs » Propagation of properties extracted from the HyperLEDA catalog in the calculation of luminosities of galaxies

Pack: Propagation of properties extracted from the HyperLEDA catalog in the calculation of luminosities of galaxies

Created at: 24/11/11 @ 19:12:05 Last updated: 01/12/11 @ 22:13:34

| Add an Item | Sharing | Tags (6) | Featured in Packs (0) | Favourited By (0) | Comments (0) |

Title: Propagation of properties extracted from the HyperLEDA catalog in the calculation of luminosities of galaxies

Description

The scientific experiment represented by this research object pertains to the multi-wavelength study for a sample of the most isolated galaxies in the local universe. This study characterizes each galaxy of this sample through both the measurement of basic astrophysical properties:
- The equatorial coordinates in J2000 epoch
- The velocities in km/s (v)
- The dust extinction coefficient (ag)
- The axis ratio of the isophote 25 mag/arcsec2 (logr25)
- The apparent total B magnitude (BT)
- The morphological type (t)

and the calculation of the more complex properties:
- The distance in Mega parsecs (D)
- The corrected apparent B magnitude (btc)
- The optical luminosity in B-band (LB)

Specifically, this research object is focused on the calculation of the intrinsic luminosity in the Johnson B-band, in order to achieve it the measurement or

Creator

Jose Enrique Ruiz

12 items in this pack

Tags (6)

□ Creator tags

astronomy | curation | database storing | galaxies | hyperleda | virtual observatory

[edit]

Tags from Items (5)

In a cloud of web services and data..

Web Services should benefit of the *same privileges* acquired by Data until now.

Start thinking on how to provide

- **Detailed curation**
- **Thorough characterization**