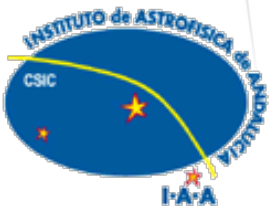


Research Objects and WS Characterization

Jose Enrique Ruiz
jer@iaa.es

November 29th 2012
Workflows Group ASOV France



Wf4Ever

Advanced Workflow Preservation Technologies for Enhanced Science





1. Intelligent Software Comp... (Spain)
2. University of Manchester
3. Universidad Politéc... (Spain)
4. Poznan Superco... Working
5. University of... TOXF, U
6. Institut... alucía
7. Leiden U... Medical

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

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



A normalized preservation of methodology is needed

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



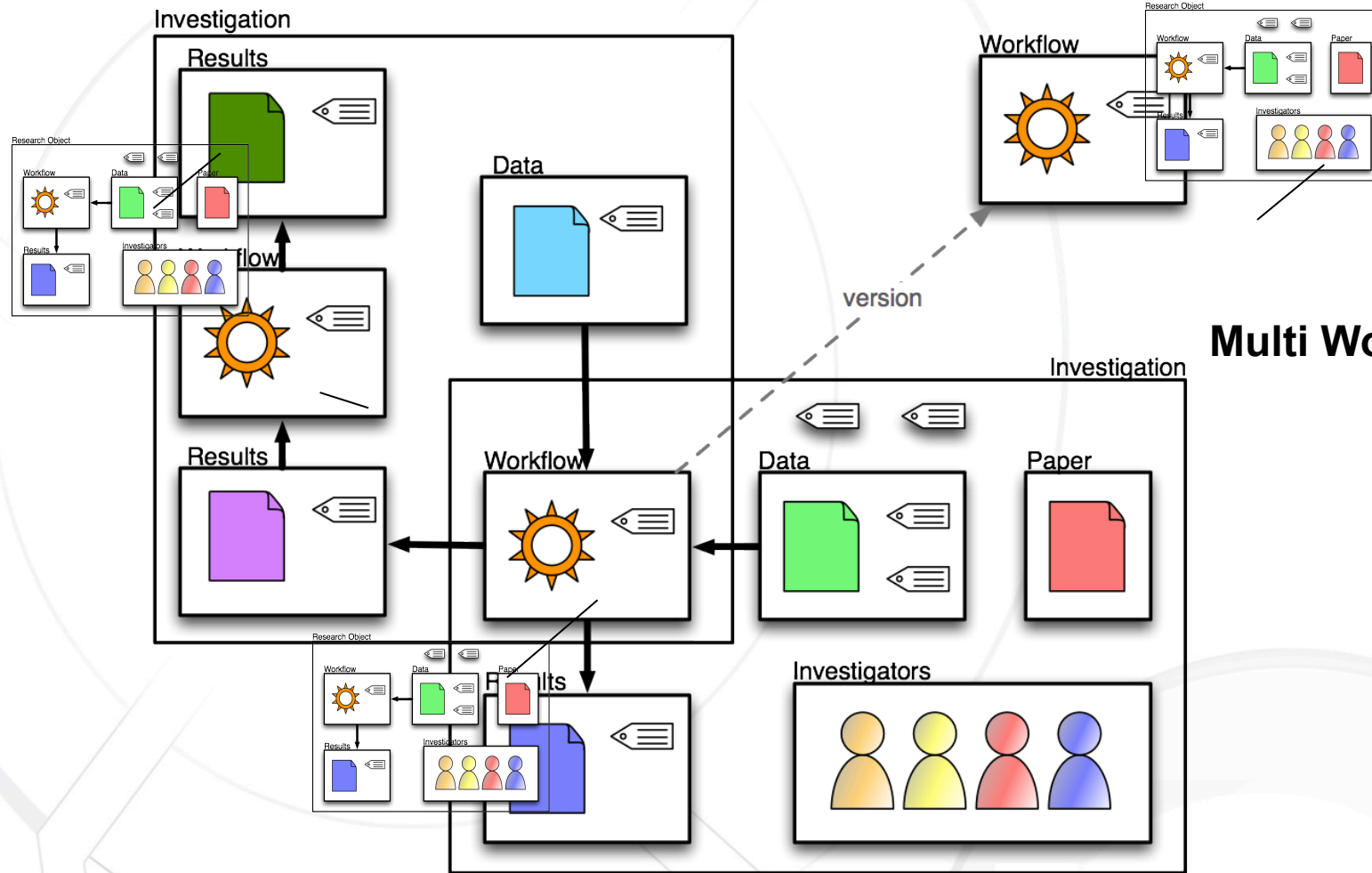
Barriers to Data and Code Sharing in Computational Science

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

Code	I don't know how	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%

Tools

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

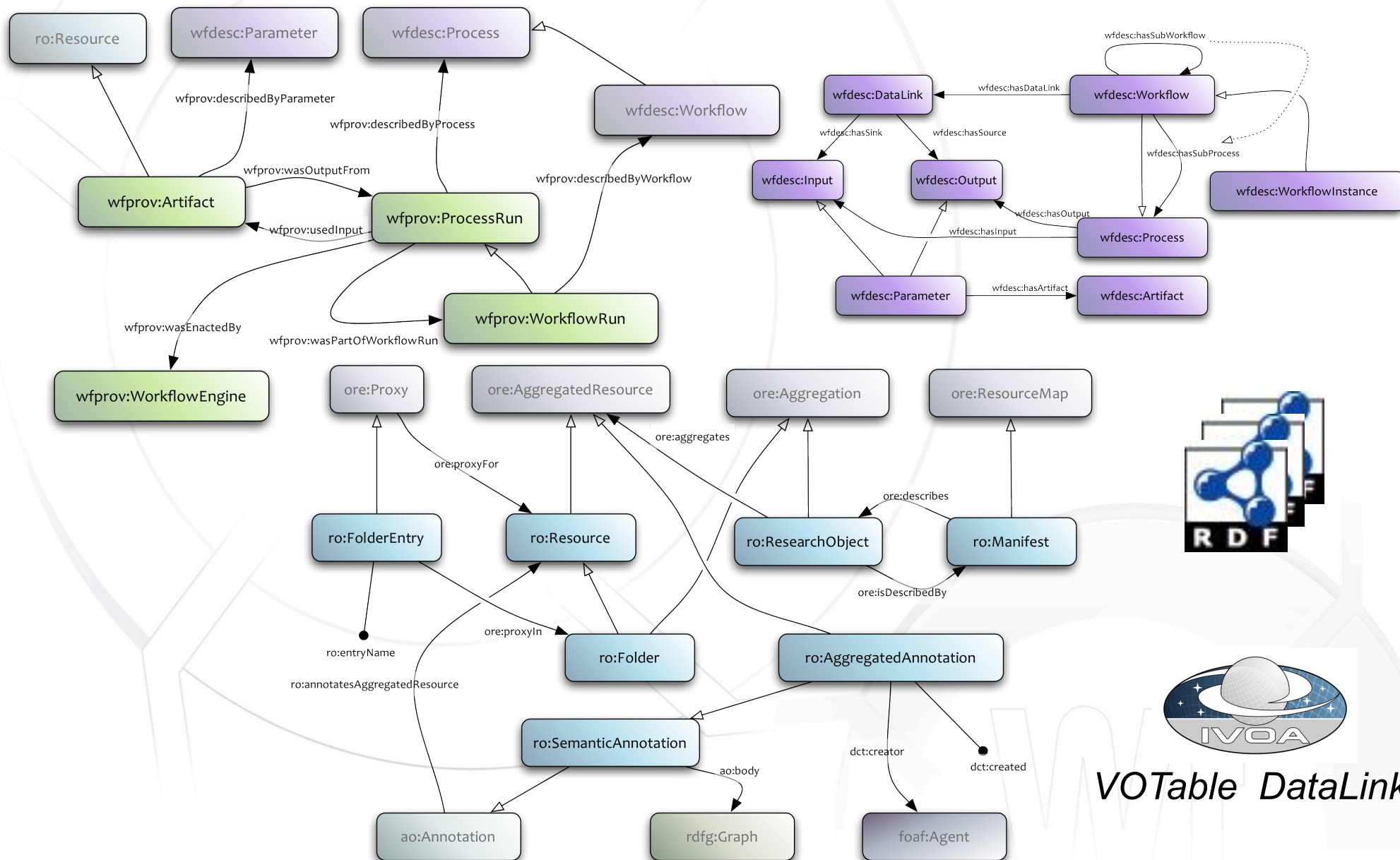
RO Template

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

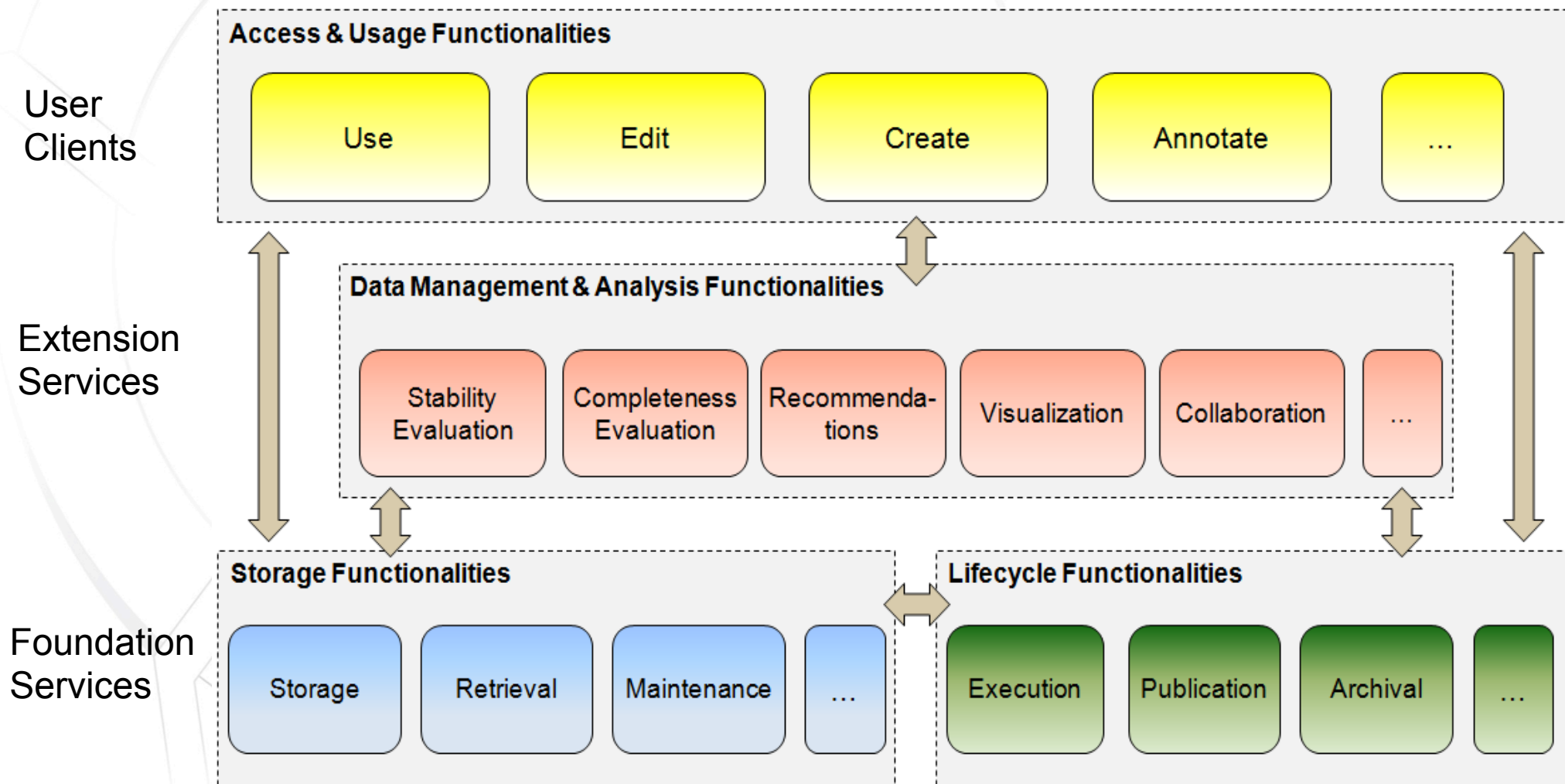
Semantic Annotations

- » 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

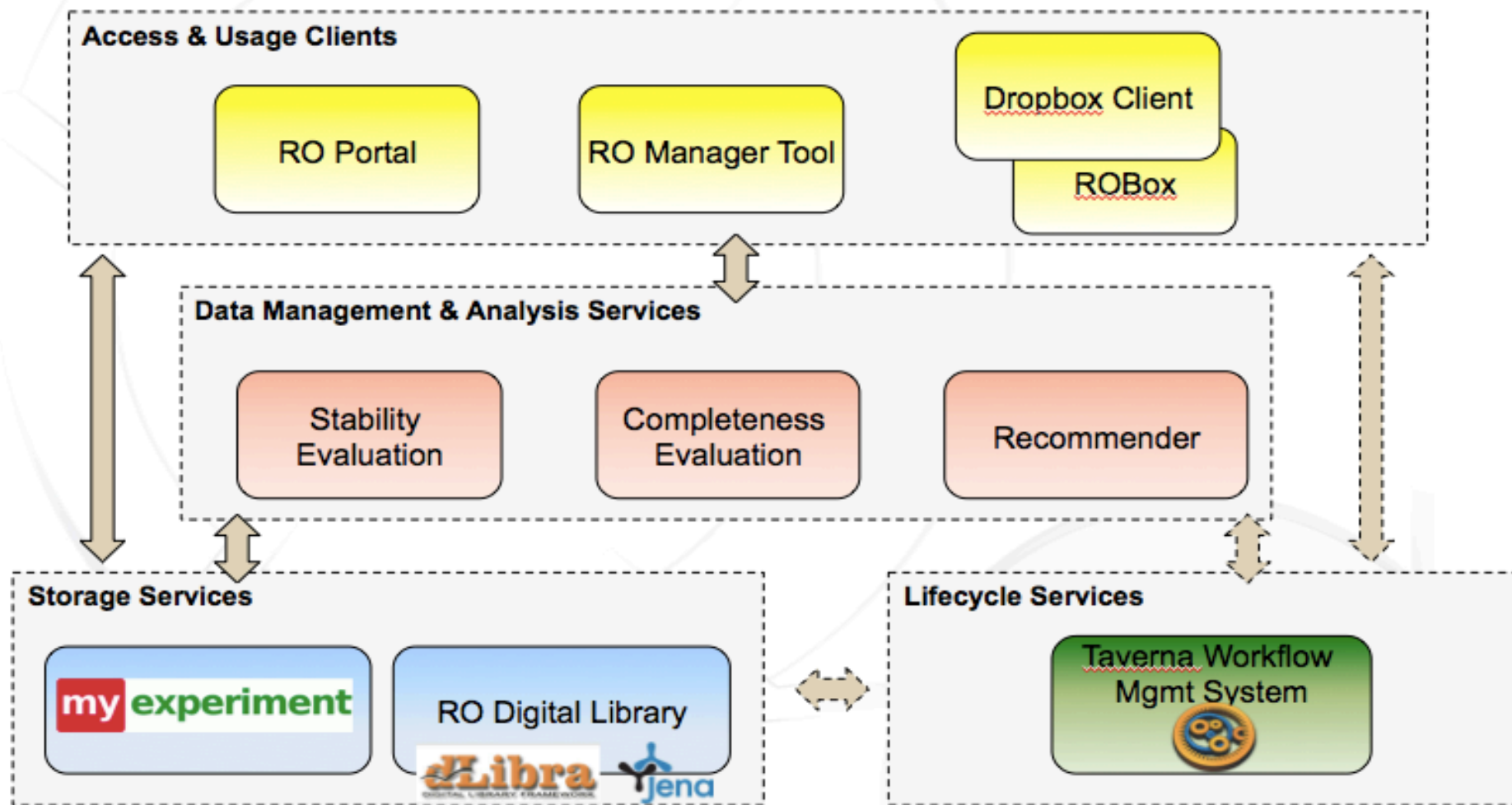
Research Object Wf4Ever Semantic Model



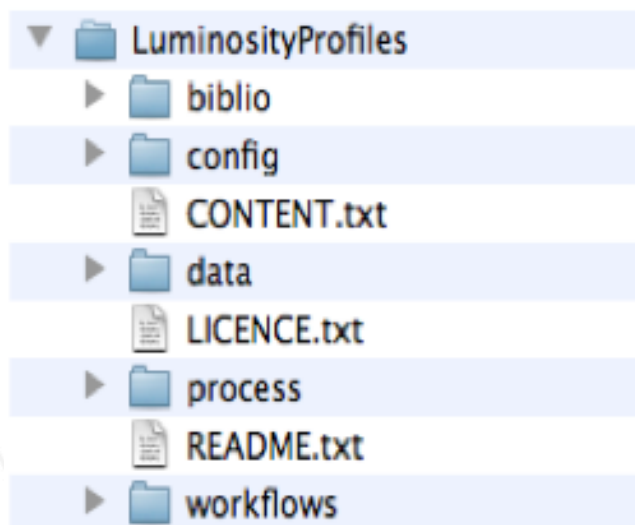
Research Object Digital Library Architecture



Research Object Digital Library Architecture



Luminosity Profiles RO



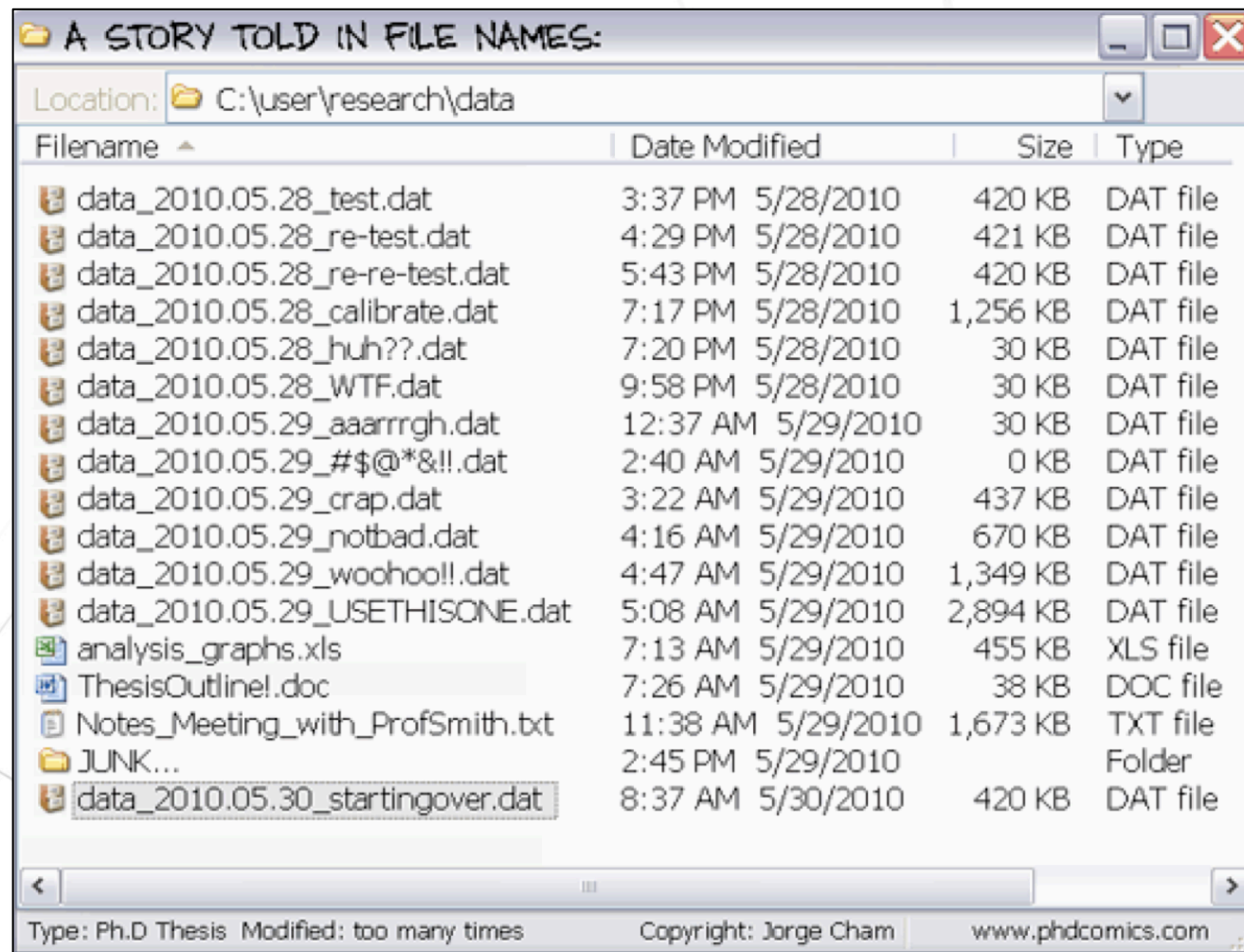
1010 Files, 200 MB
External Sources ~ 8 GB

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

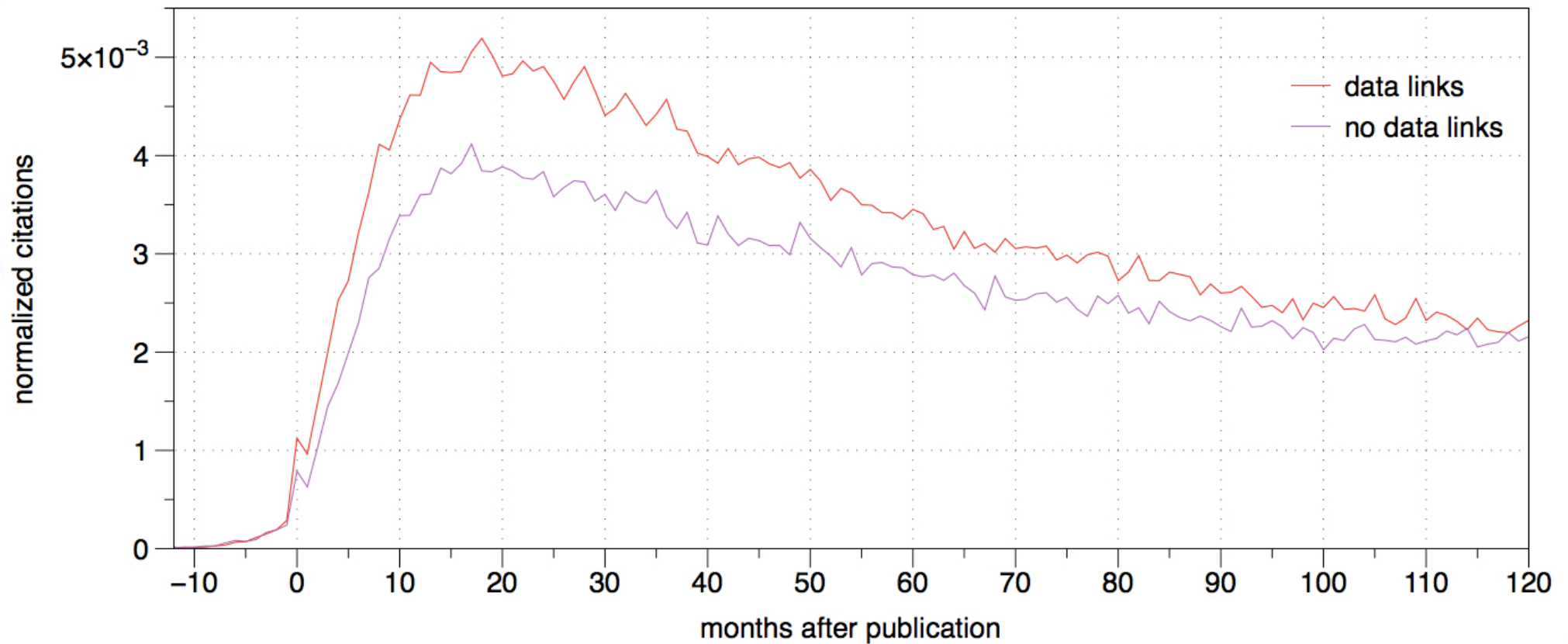
When organization is better than automation



Credit and attribution

Papers with data links are cited more than those without

1995 - 2000

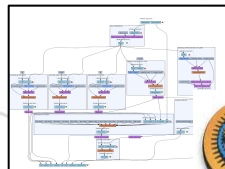
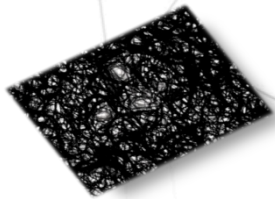


Effect of E-printing on Citation Rates in Astronomy and Physics
2006. Edwin A. Henneken et al.

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 ADS Labs Fulltext Search

Home Labs Home ADS Classic Help

Limit your search

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

Welcome to ADS Labs 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.

Search

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](#) [f](#) Me gusta 1

Site powered by Blacklight

ads labs ADS Labs Streamlined Search

Home Labs Home ADS Classic Help Sign on

Search Astronomy SEARCH

Author First author Title Object Examples

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](#) [f](#) 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

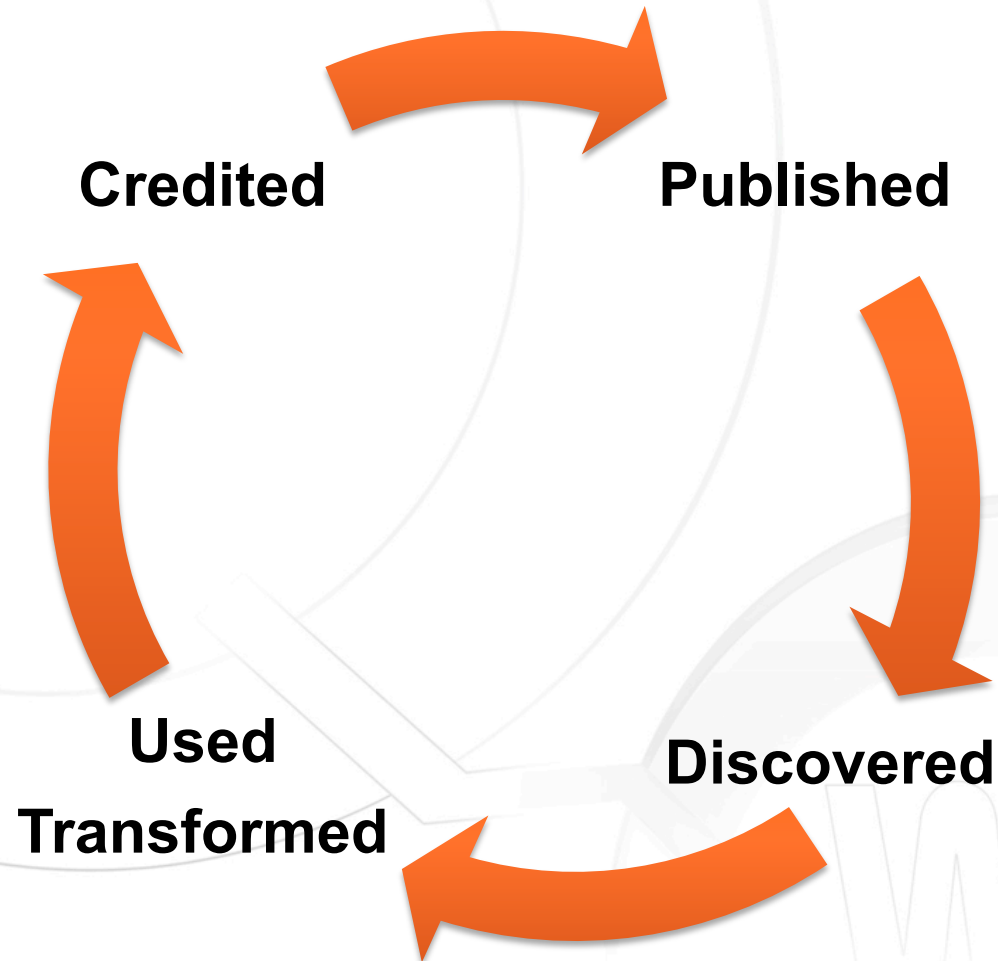
Web Services Discovery

Web Services Access

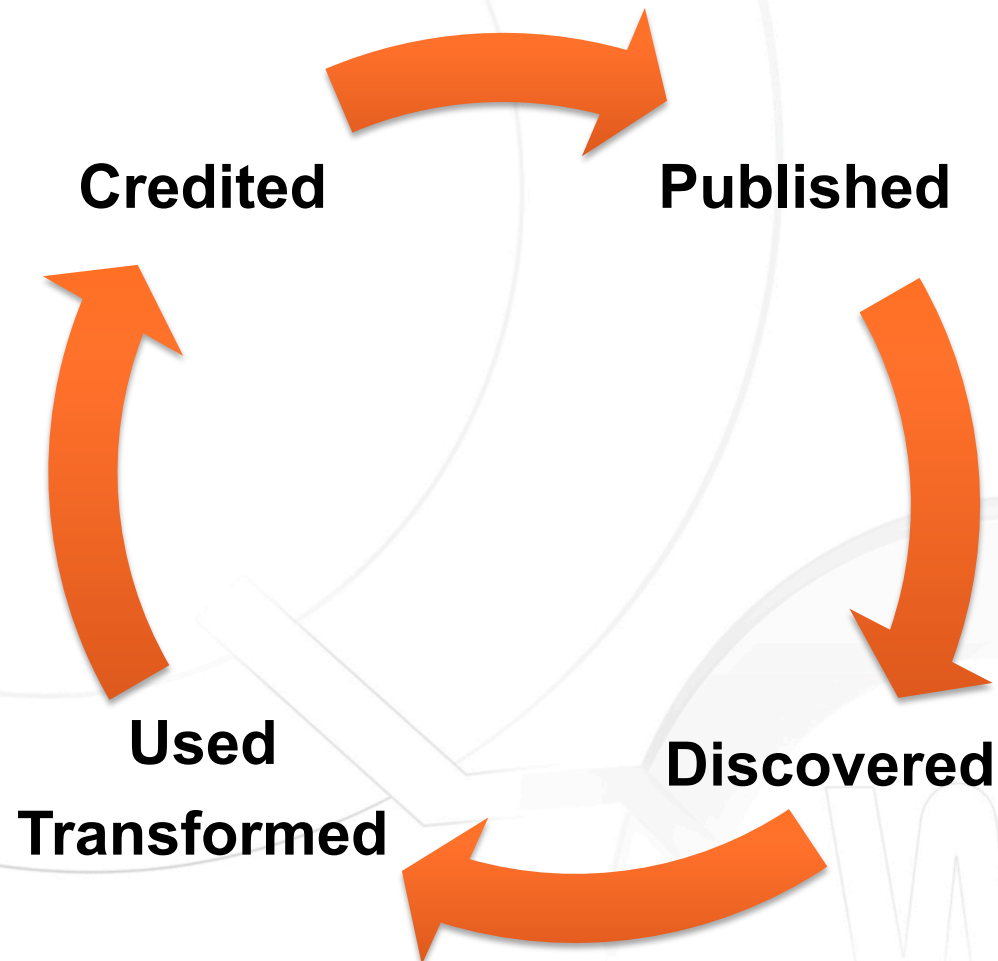
Web Services Management

The Web-Services
Deluge

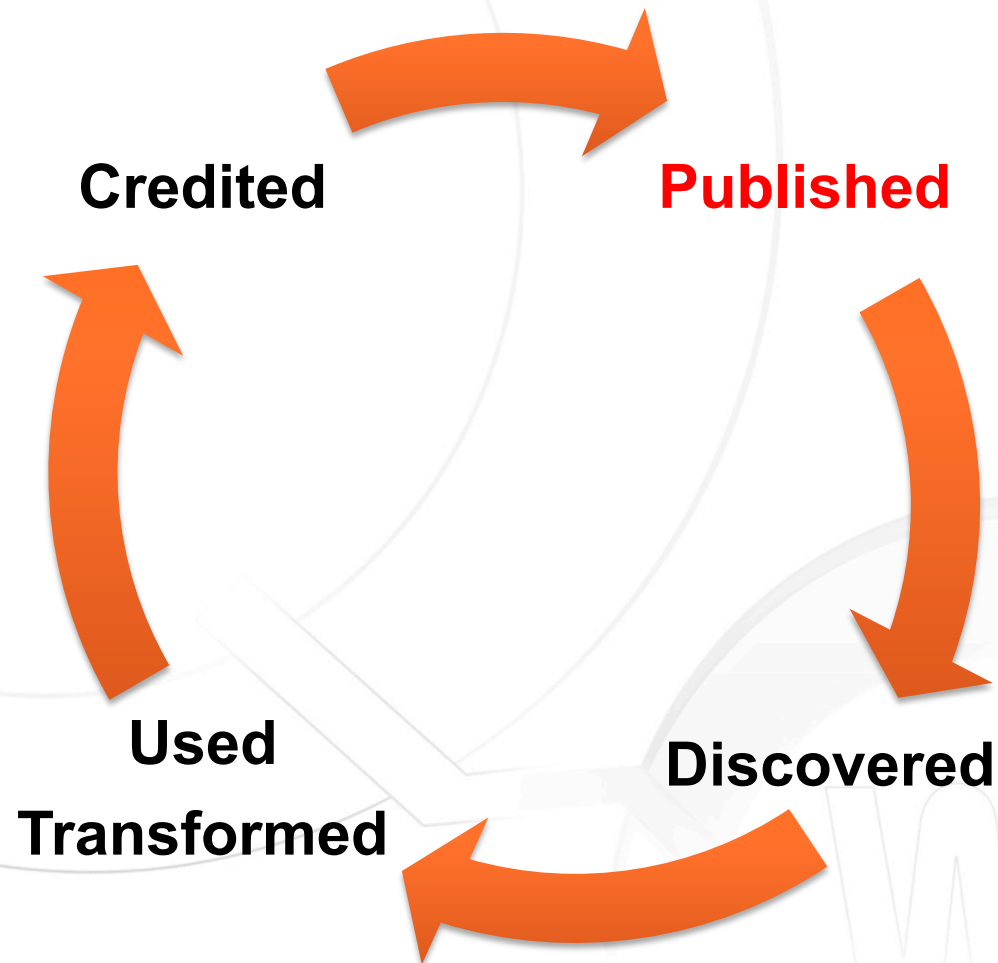
The Data Lifecycle



The **Service** Lifecycle



The **Service** Lifecycle



Published

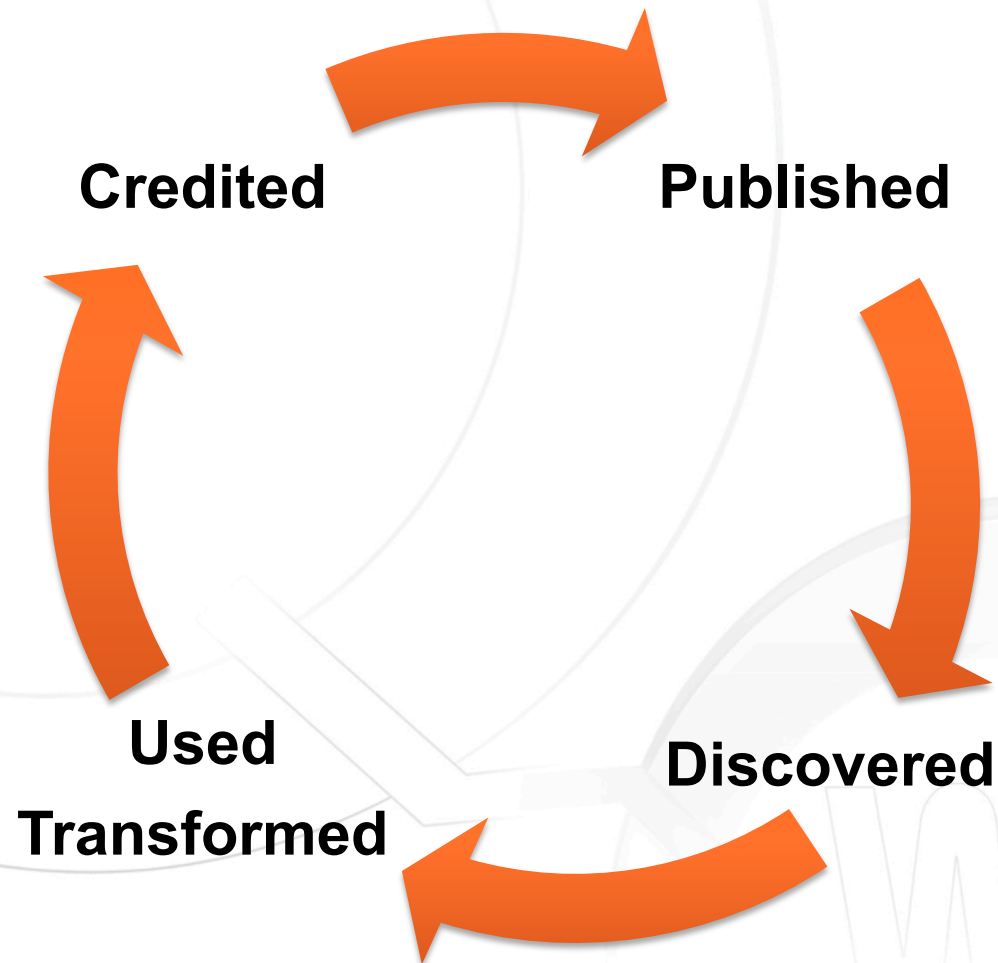
- The VO Registry
- Easier to publish services than datasets in the VO ?
- WS are not exclusive property of big data archives
- Publication is not Preservation
- Backup strategies
- Replication/Mirrors
- Versioning
- Software Publishing Platforms

github

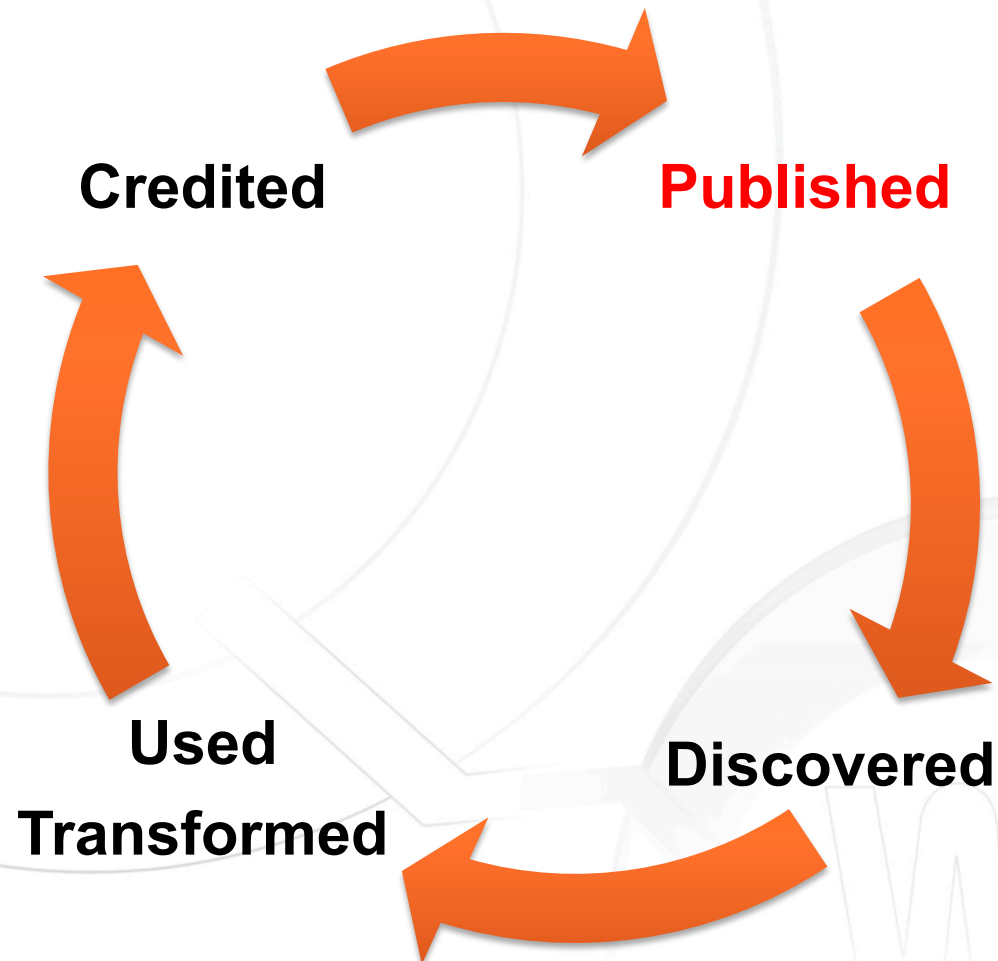
runmycode

The image shows two overlapping screenshots from the Publishing Registry Portal. The top screenshot is the 'Register a New Publisher' form, which includes instructions and a legend for required (red) and recommended (green) fields. The bottom screenshot shows the main interface of the EURO-VO Registry, featuring a sidebar menu with categories like 'Search Resources', 'Insert Resources', and 'Resource Curation'. The main content area is titled 'Insert a New Resource' and contains a form for creating a new service resource, with fields for Title, Identifier, Short Name, Status, Password, Publisher, and Creator.

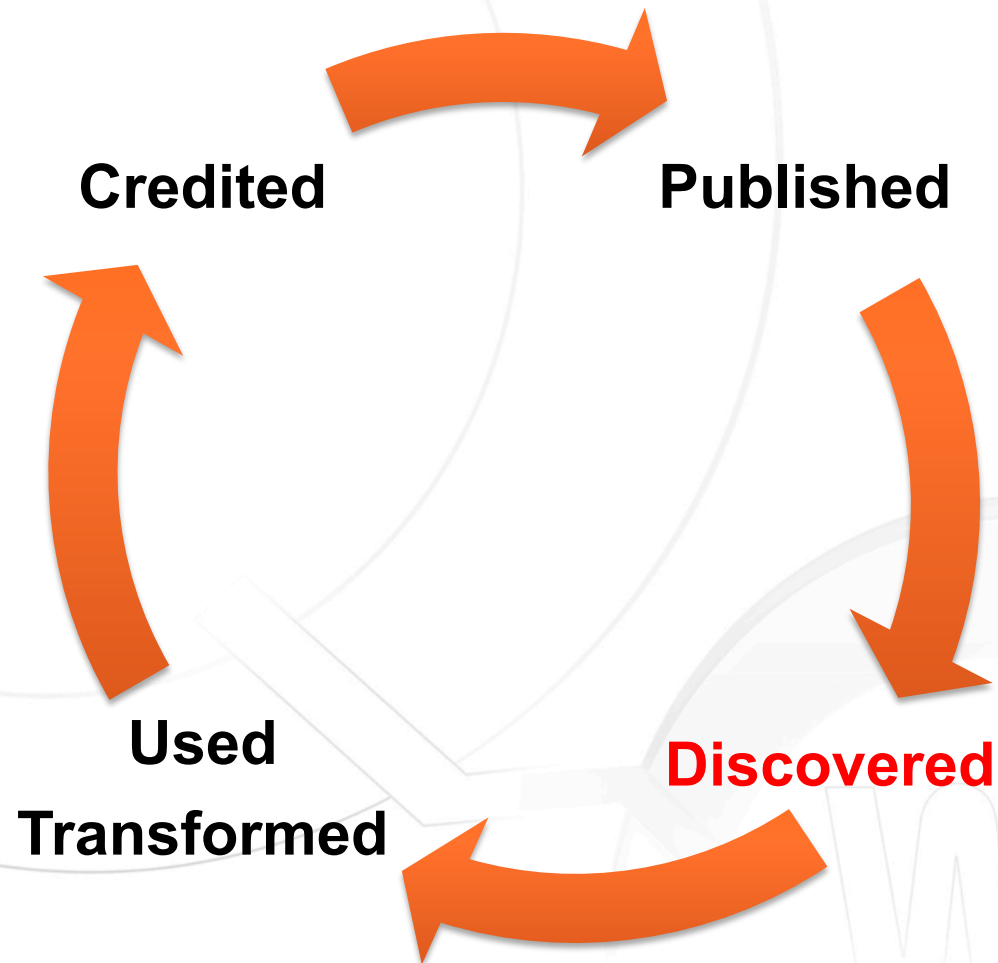
The **Service** Lifecycle



The **Service** Lifecycle



The **Service** Lifecycle



Discovered

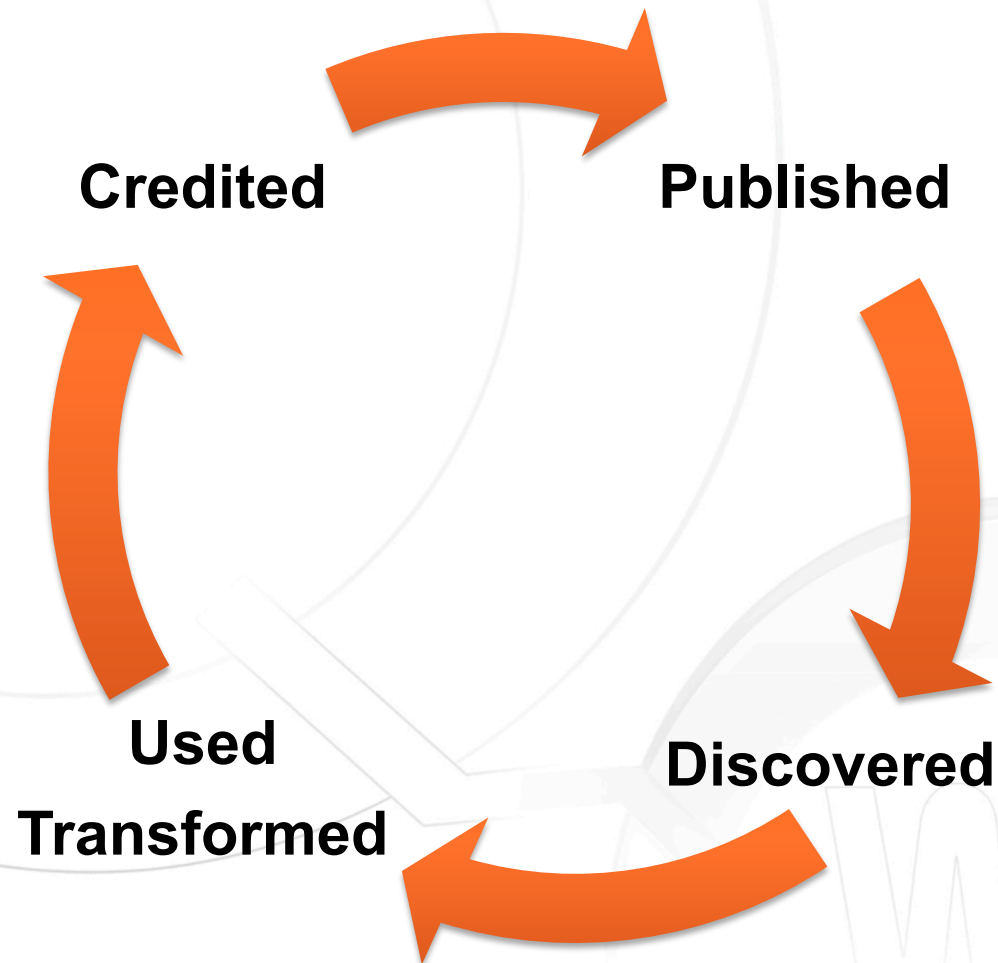
- **Search Criteria**

- 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

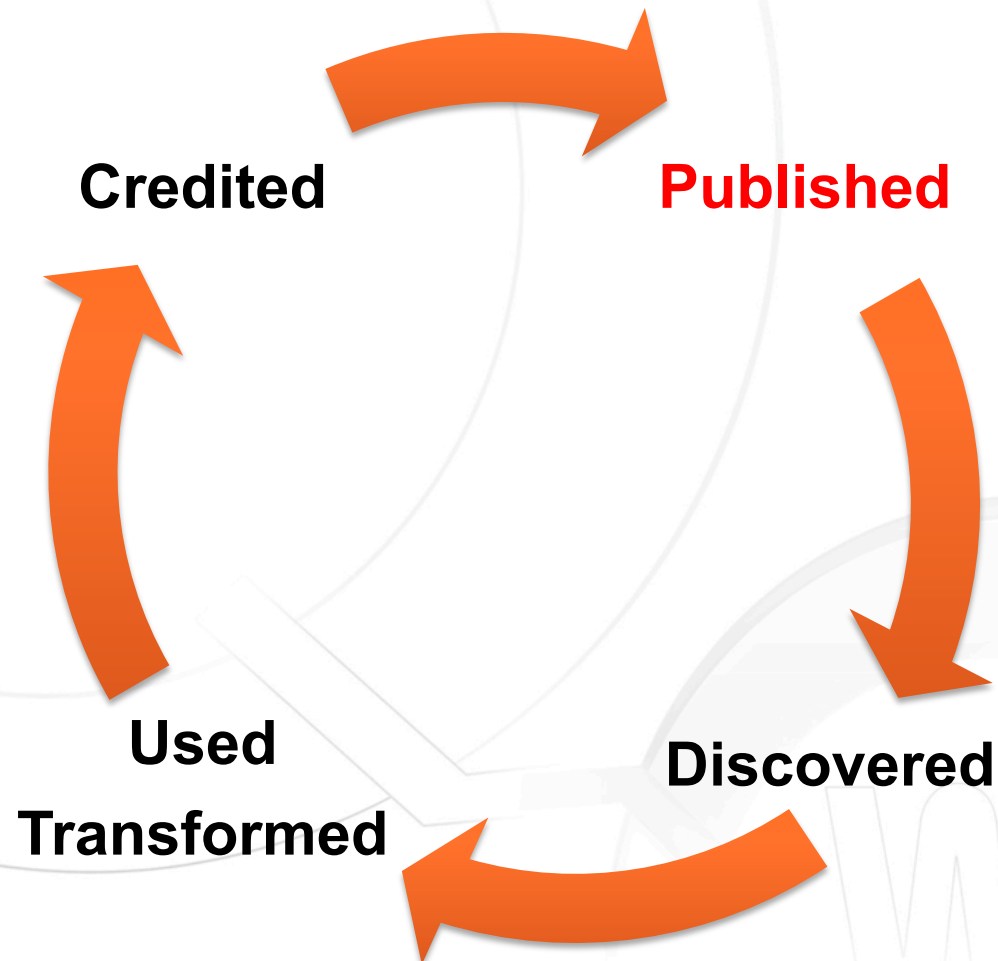


- **Proposition of alternatives and similars**

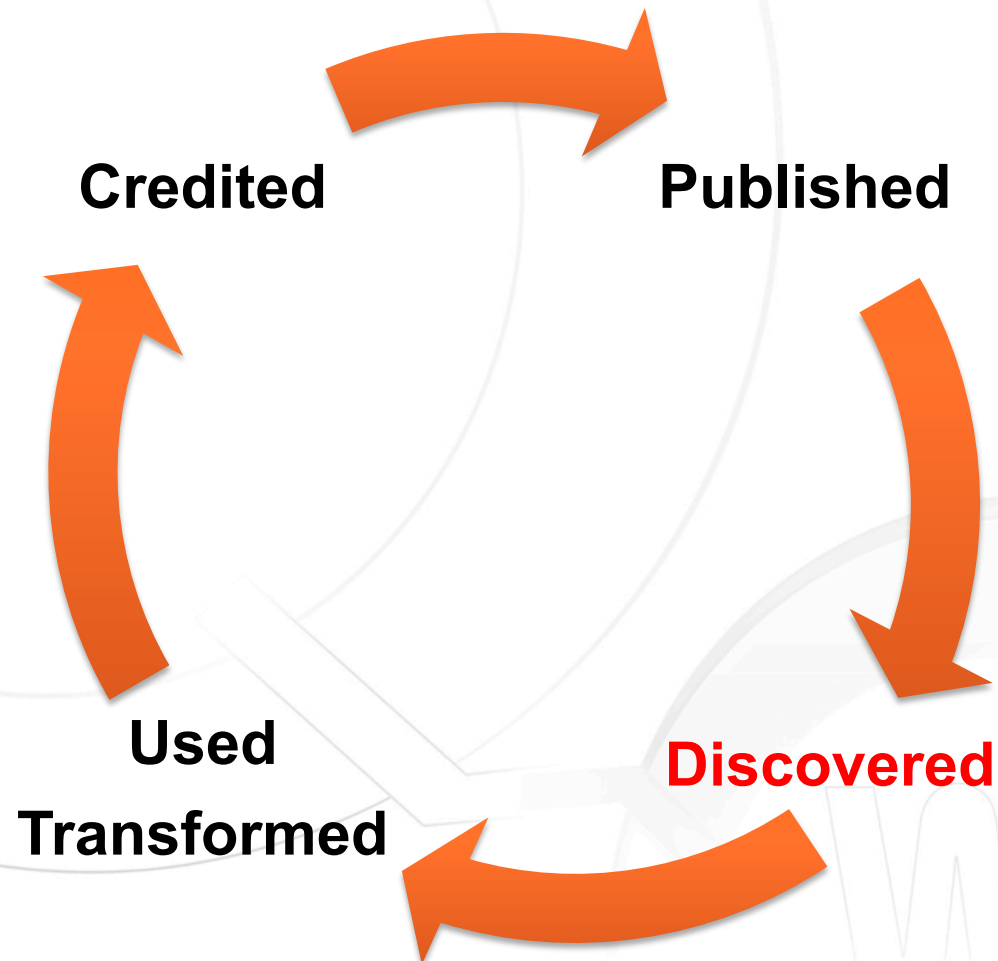
The **Service** Lifecycle



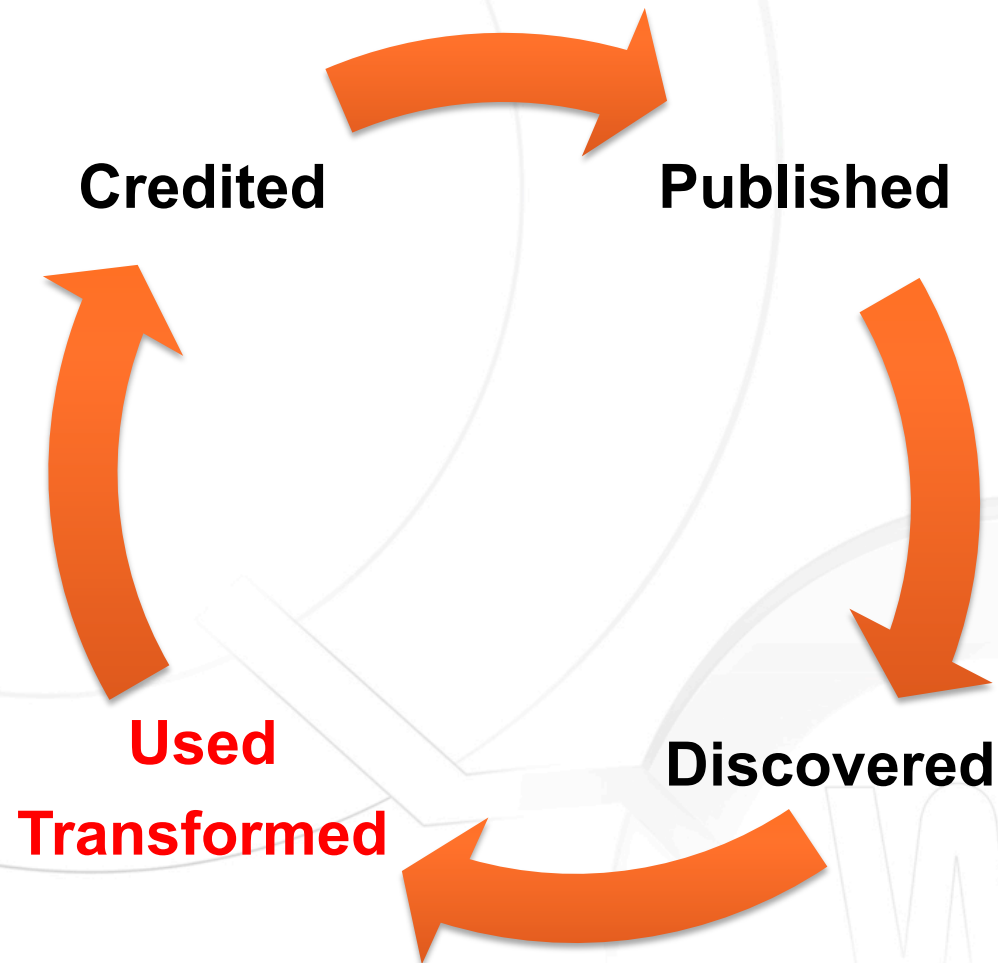
The **Service** Lifecycle



The **Service** Lifecycle

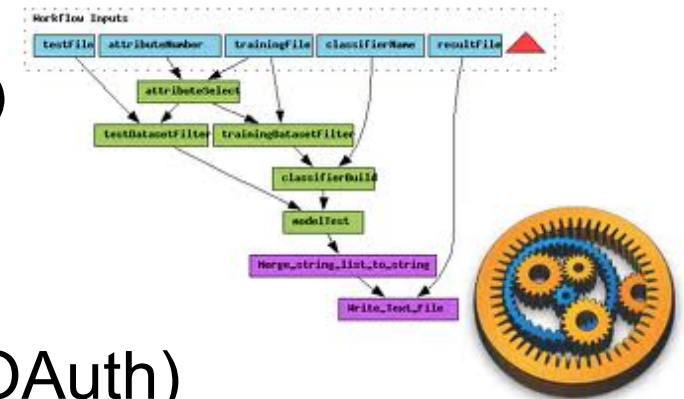


The **Service** Lifecycle

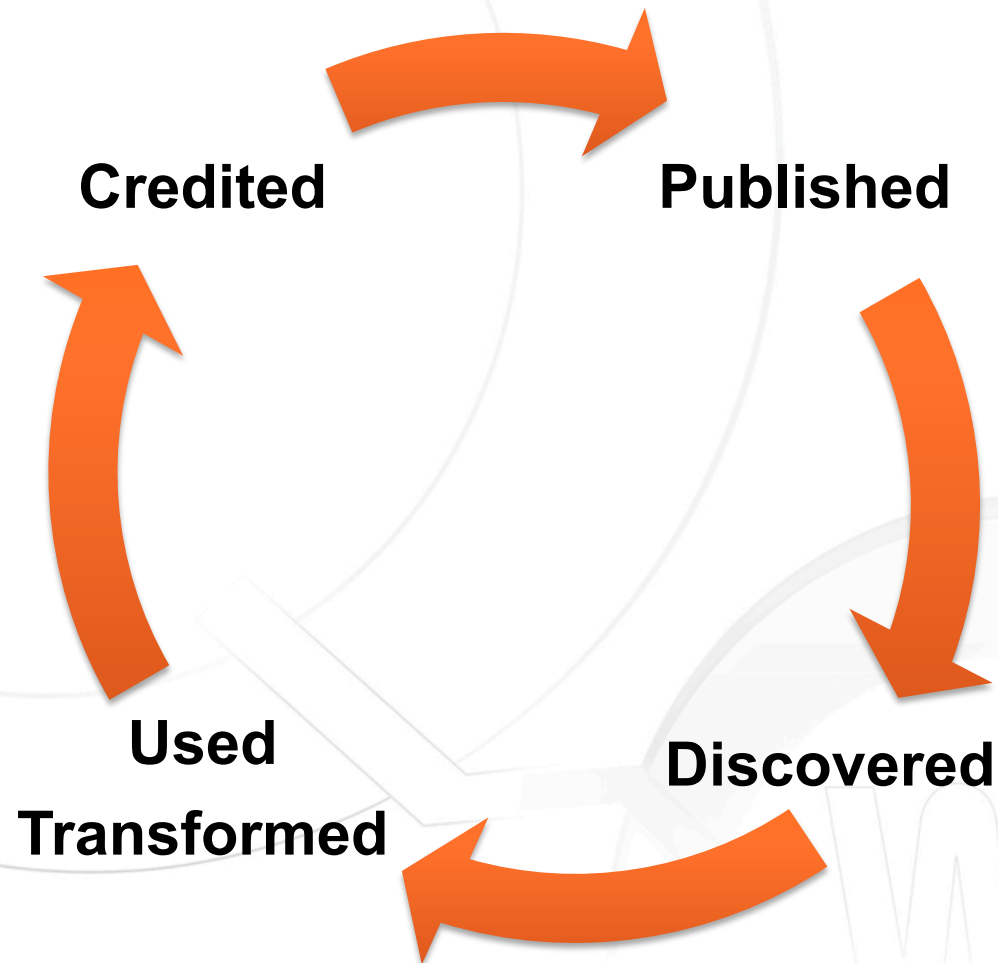


Used and Transformed

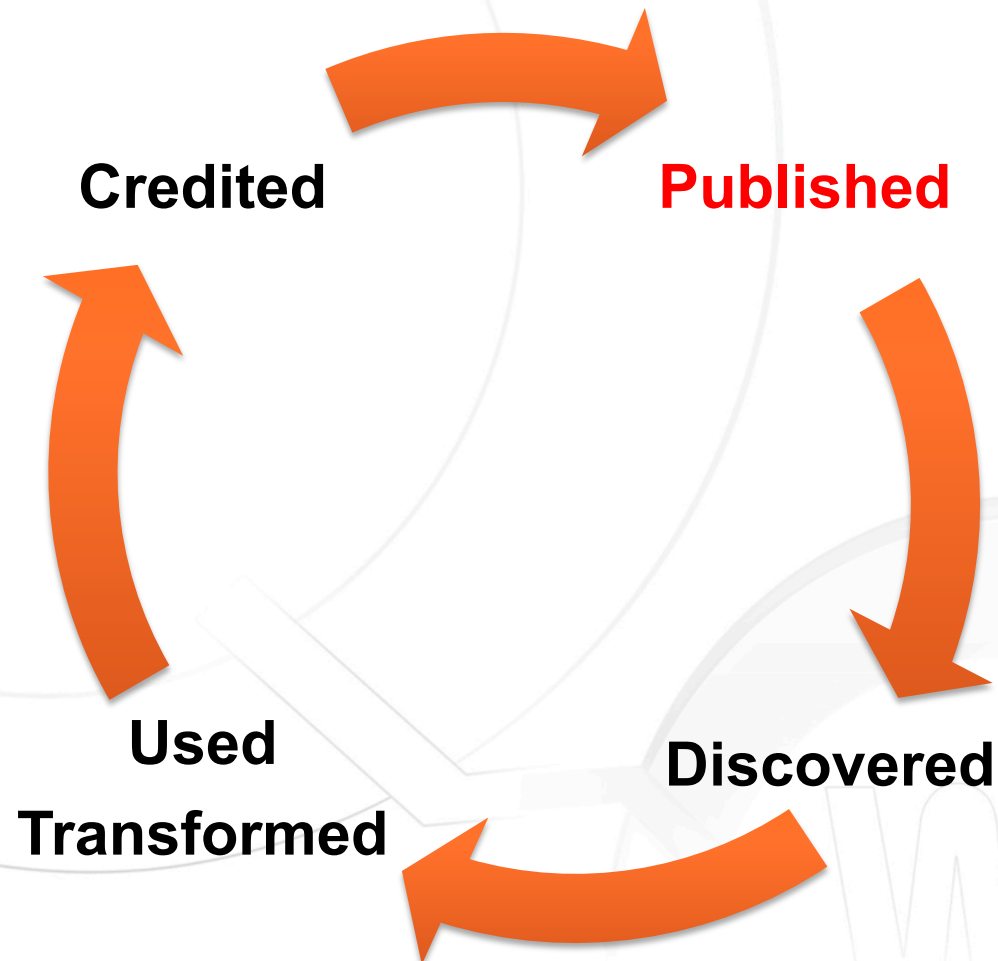
- **How to use them ?** (WADL, WSDL – VOSI #capabilities)
 - Input Data -> Parameters needed and formats
 - Self-described WS (PDL, S3, SimDAL, SimDB)
 - Output Data -> Response format - TAP
 - Example Data, Self-Consistency Checking
- Access Policy (**Authentication** – SSO, OAuth)
- **WS orchestration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



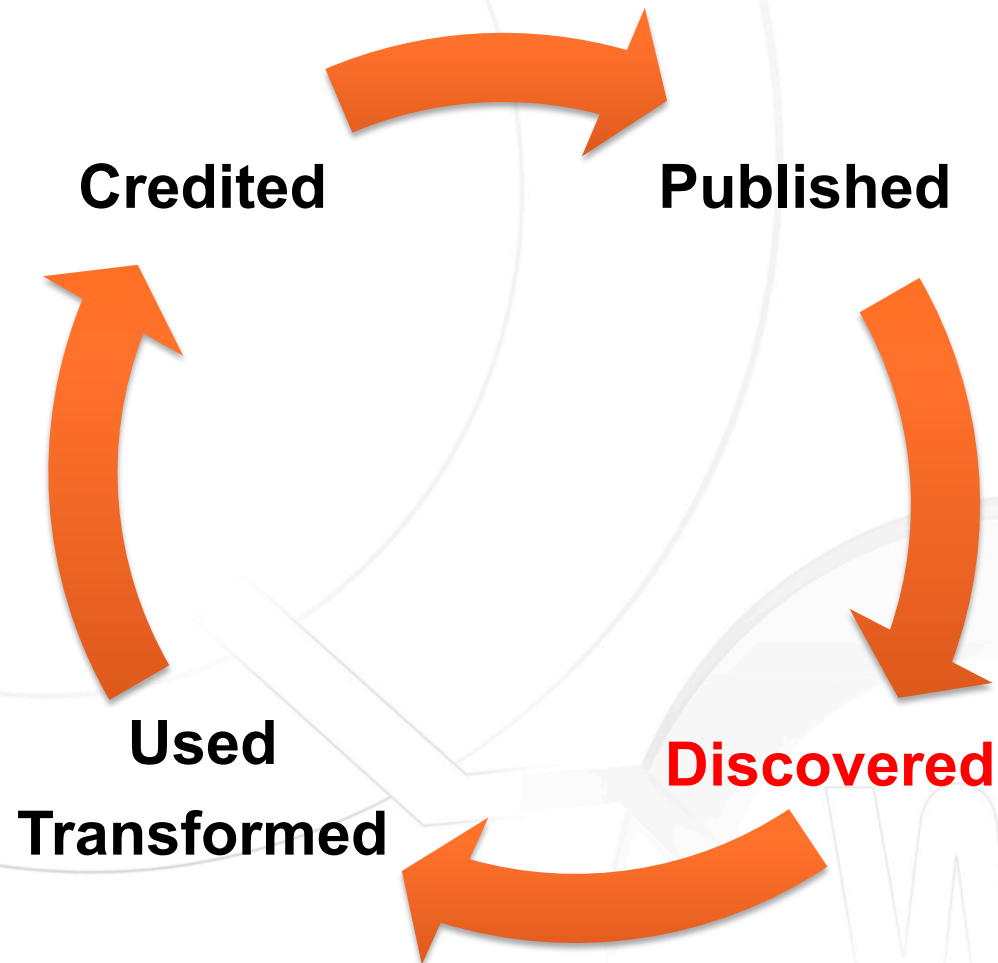
The **Service** Lifecycle



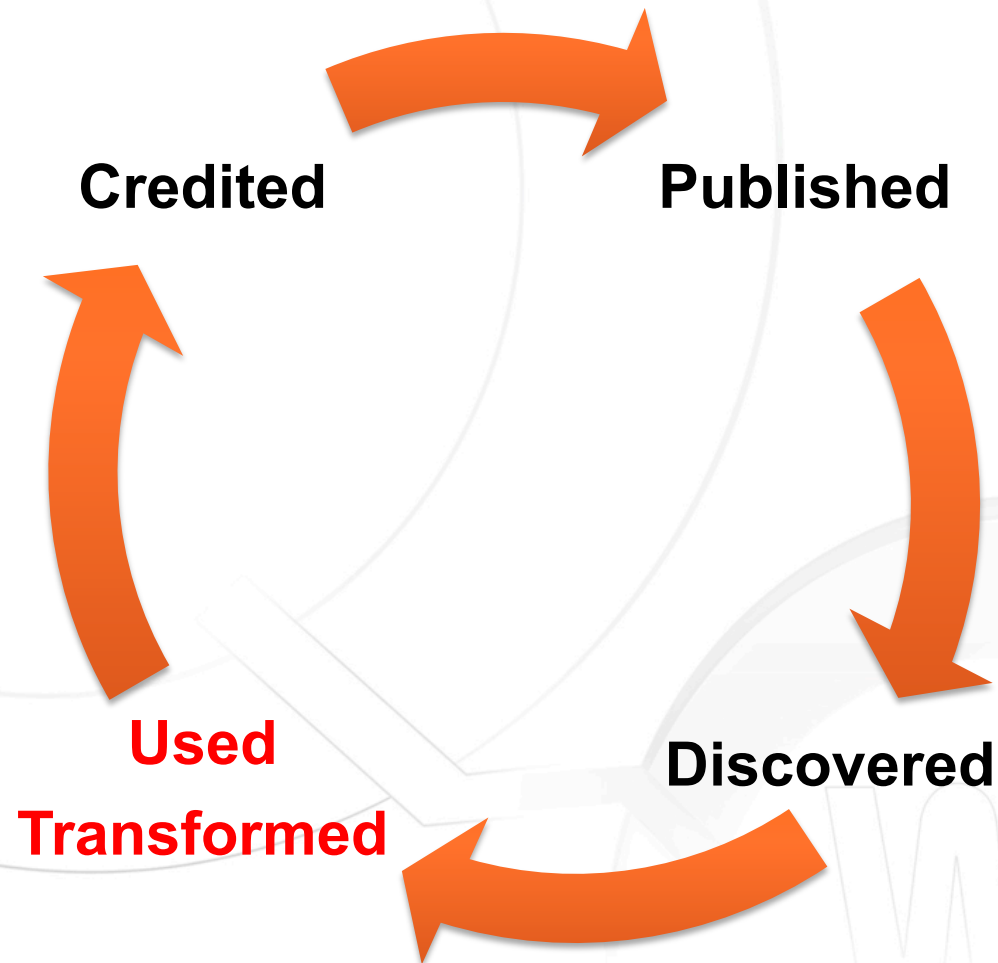
The **Service** Lifecycle



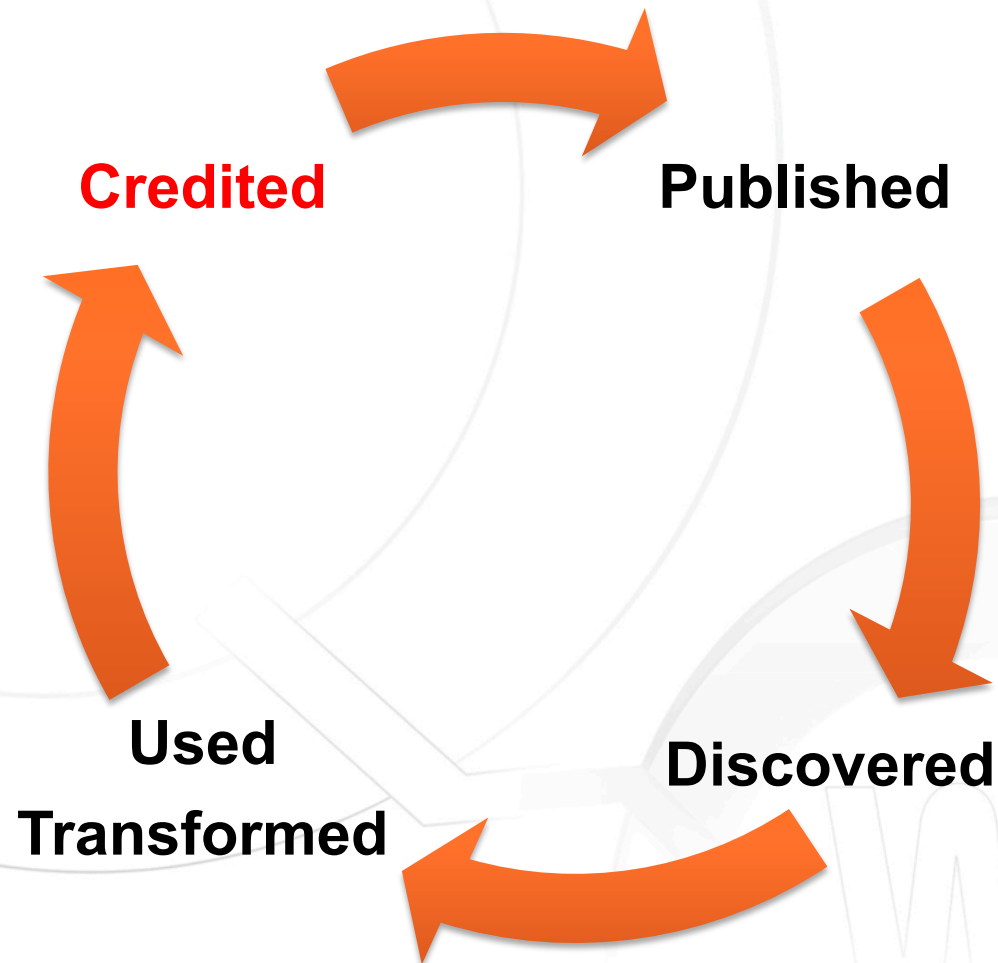
The **Service** Lifecycle



The **Service** Lifecycle

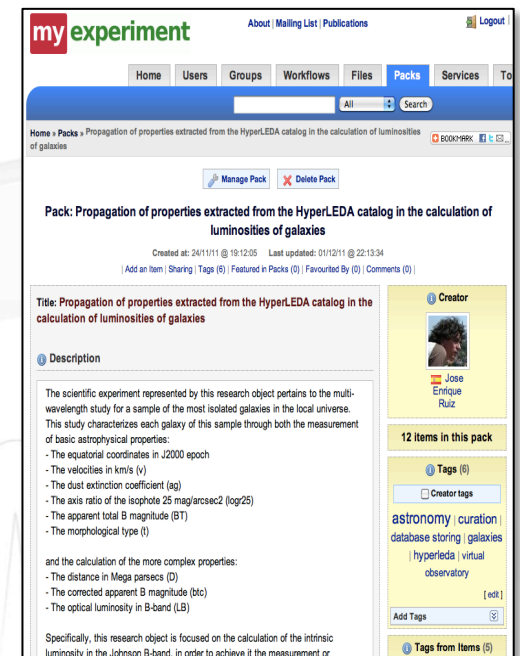
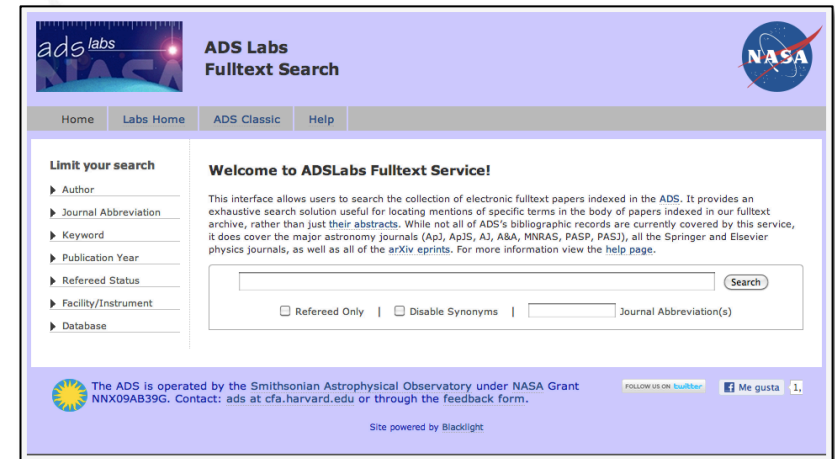


The **Service** Lifecycle



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



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**

