The University of Manchester

MANCHESTER 1824

The Reality of Reproducibility of *in silico* Science

Prof Carole Goble FREng FBCS CITP

JCDL Washington DC, June 2012

X REUNIÓN CIENTÍFICA
DE LA SOCIEDAD ESPAÑOLA
DE ASTRONOMÍA

VALENCIA
9/13 JULIO

Digital Science

Reproducibility and Visibility in Astronomy

José Enrique Ruiz, Lourdes Verdes-Montenegro, Susana Sánchez,
Julian Garrido, Juan de Dios Santander and the Wf4Ever Team

SESIÓN INSTRUMENTACIÓN Y COMPUTACIÓN
VALENCIA, VIERNES 13 JULIO 2012

AMIGA
Analysis of the Interstellar Medium of Isolated GAlaxies

INSTITUTO de ASTROFISICA de ANDALUCIA
CSIC
I·A·A

CSIC

Workflow 4Ever

SEVENTH FRAMEWORK PROGRAMME

European Commission
Information Society and Media

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 rich infrastructure of data (VO) is not efficiently used

A normalized preservation of methodology is needed

Tools

The next generation of archives

Much wider FoV and spectral coverage

- Large volumes for an observed datacube
- Subproducts are **Virtual Data** generated on-the-fly

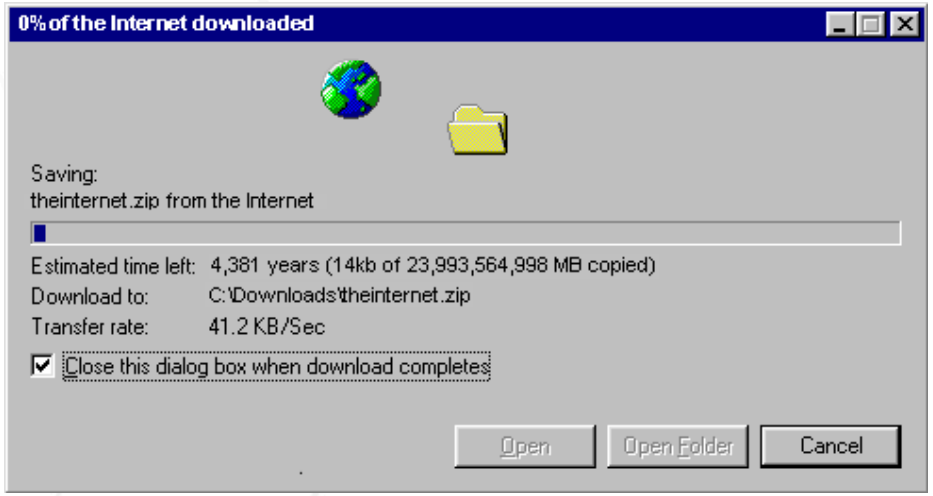
	Low Res		High Res		Extreme Res	
Number	4 Bytes	4B	4 Bytes	4B	4 Bytes	4B
Resolution	2,048 x 2,048	16MB	8,192 x 8,192	268MB	12,288 x 12,288	603MB
Channels	16,384	0.27TB	16,384	4.39TB	16,384	9.8TB
Stokes & Weighting	1	0.27TB	1	4.39TB	4 + 1	49.5TB

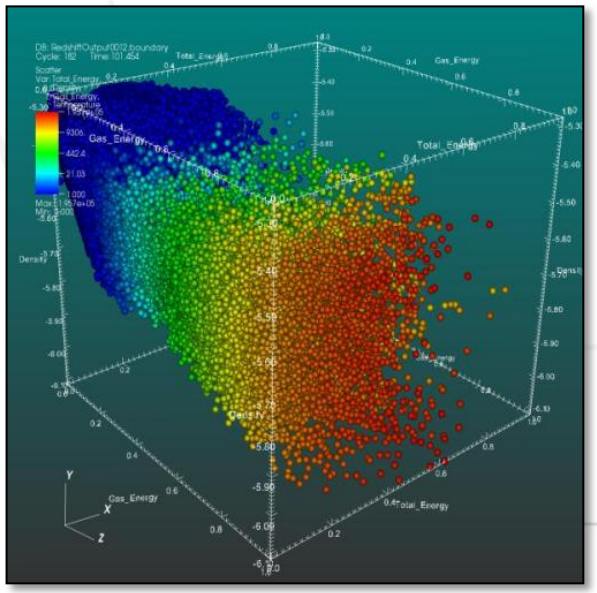
ASKAP Cubes
Prof. Kevin Vinsen

The next generation of archives

Automated surveys

- Huge amounts of tabular data
- Services for KDD





Extraction of scientifically relevant information from a multidimensional parameter space

- Exploration services
- Anomaly detection
- Cross-matching data
- Dimensionality reduction

The next generation of archives

- A cloud of Web Services

  Archives should evolve from data providers into
  - Virtual data providers
  - Software tasks providers

- Archives speaking Web Services

  Astronomy of multi archives/facilities/wavelength

  Interconnected and interoperable archives
  - Software Tasks
  - Data

Process should benefit of the same privileges acquired by data

Preserving the method ensures replication of final results at any moment

Digital Science - Reproducibility and Visibility in Astronomy

Efficiency and Reuse

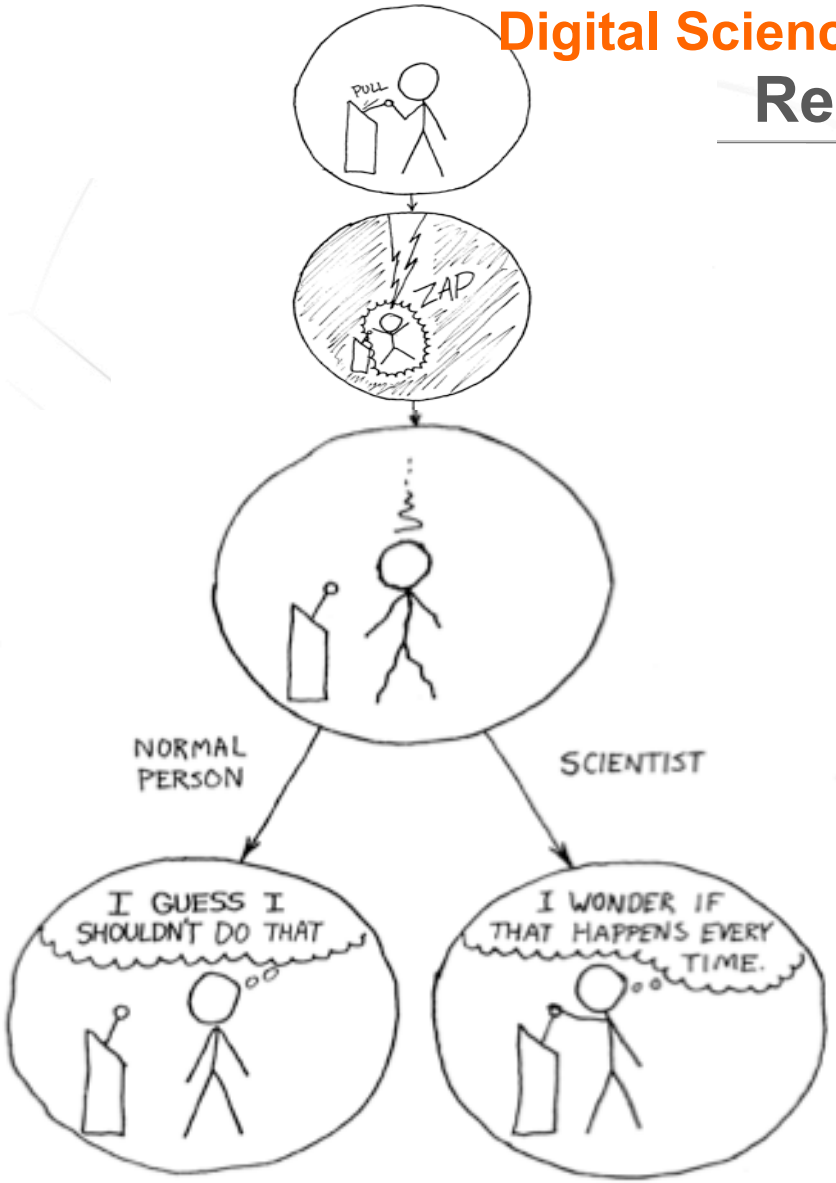
Optimize return on investments made on big facilities

- 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 softw**

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

Reproducibility and The Scientific Method



http://xkcd.com/242/

Benefits

- Publishing knowledge, not advertising
- The author, the referee and the re-user
- Reputation, prestige and respect
- Higher quality of publications
  - Authors will be more careful
  - Many eyes to check results

Challenges

- Hard and time consuming
- Need incentives – not rewarded now

Initiatives

- Elsevier Executable Papers Challenge
- Open Data / Open Science

Reproducibility and The Scientific Method

Barriers to Data and Code Sharing in Computational Science

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

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	[illegible]	-
34%	I don't know how	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%

Digital Science - Reproducibility and Visibility in Astronomy

Discovery, Visibility and Credit

nature International weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Specials & supplements archive > Science Metrics

SPECIALS

See all specials

How to improve the use of metrics

Nature **465**, 870–872 (17 June 2010) | doi:10.1038/465870a

... "Science is being killed by numerical ranking,"[...] Ranking systems lures scientists into pursuing high rankings first and good science second.

SCIENCE METRICS

The value of scientific output is often measured, to rank one nation against another, allocate funds between universities, or even grant or deny tenure. Scientometricians have devised a multitude of 'metrics' to help in these rankings. Do they work? Are they fair? Are they over-used? *Nature* investigates.

▾ Editorial ▾ Features ▾ Opinion ▾ From the archive

Journal home | Subscribe

Current issue | E-alert sign up

For authors | RSS feed

Top content

Emailed | Downloaded | Blogged

1. Public health: The toxic truth about sugar
   Nature | 01 February 2012
2. The case for open computer programs
   Nature | 22 February 2012
3. The great beyond
   Nature | 29 February 2012
4. The darker side of stem cells
   Nature | 29 February 2012
5. Cancer: Solving an age-old problem
   Nature | 29 February 2012

View all ▸

Discovery, Visibility and Credit

nature International weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Specials & supplements archive > Science Metrics

SPECIALS

See all specials

How to improve the use of metrics

Nature **465**, 870–872 (17 June 2010) | doi:10.1038/465870a

Research reverts to a kind of 'academic prostitution', in which work is done to please editors and referees rather than to further knowledge.

SCIENCE METRICS

The value of scientific output is often measured, to rank one nation against another, allocate funds between universities, or even grant or deny tenure. Scientometricians have devised a multitude of 'metrics' to help in these rankings. Do they work? Are they fair? Are they over-used? *Nature* investigates.

▾ Editorial ▾ Features ▾ Opinion ▾ From the archive

Journal home | Subscribe
Current issue | E-alert sign up
For authors | RSS feed

Top content

Emailed | Downloaded | Blogged

1. Public health: The toxic truth about sugar
   Nature | 01 February 2012
2. The case for open computer programs
   Nature | 22 February 2012
3. The great beyond
   Nature | 29 February 2012
4. The darker side of stem cells
   Nature | 29 February 2012
5. Cancer: Solving an age-old problem
   Nature | 29 February 2012

View all ▸

Discovery, Visibility and Credit

nature International weekly journal of science

Search Go

Advanced search

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video | For Authors

Archive > Specials & supplements archive > Science Metrics

SPECIALS

See all specials

How to improve the use of metrics

Nature **465**, 870–872 (17 June 2010) | doi:10.1038/465870a

... an author's h-index can reflect longevity as much as quality — and can never go down with age, even if a researcher drops out of science altogether.

SCIENCE METRICS

The value of scientific output is often measured, to rank one nation against another, allocate funds between universities, or even grant or deny tenure. Scientometricians have devised a multitude of 'metrics' to help in these rankings. Do they work? Are they fair? Are they over-used? *Nature* investigates.

▾ Editorial ▾ Features ▾ Opinion ▾ From the archive

Journal home | Subscribe
Current issue | E-alert sign up
For authors | RSS feed

Top content

Emailed | Downloaded | Blogged

1. Public health: The toxic truth about sugar
   Nature | 01 February 2012
2. The case for open computer programs
   Nature | 22 February 2012
3. The great beyond
   Nature | 29 February 2012
4. The darker side of stem cells
   Nature | 29 February 2012
5. Cancer: Solving an age-old problem
   Nature | 29 February 2012

View all ▸

Exploring and understanding scientific metrics in citation

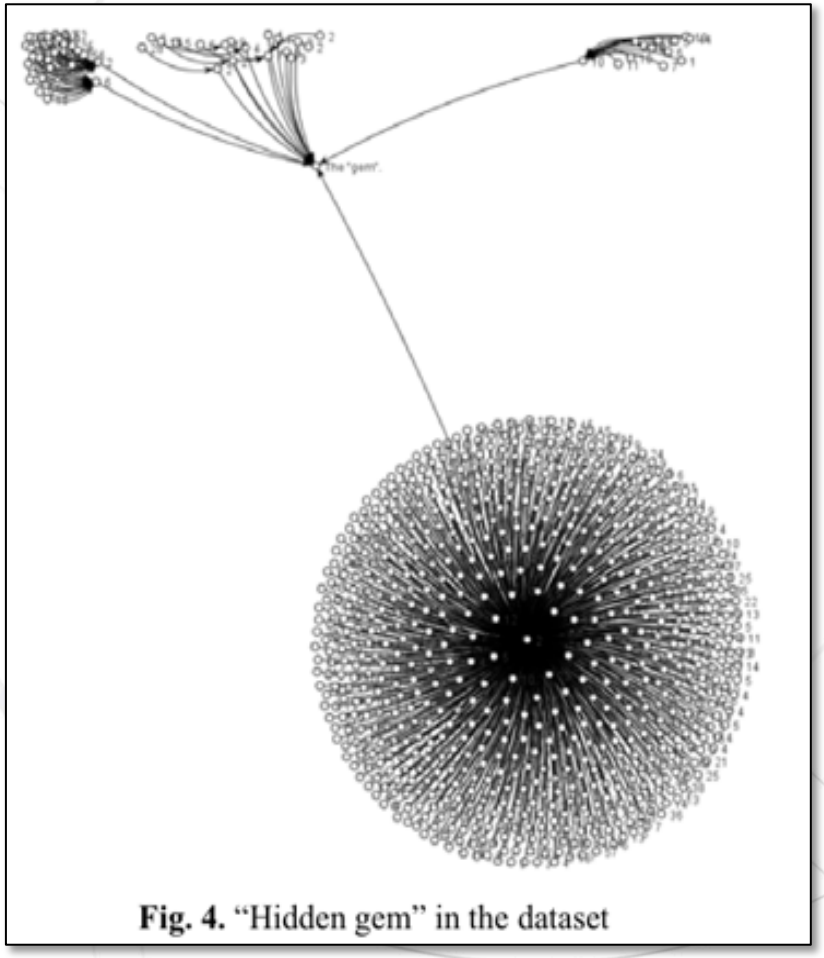


Fig. 4. "Hidden gem" in the dataset

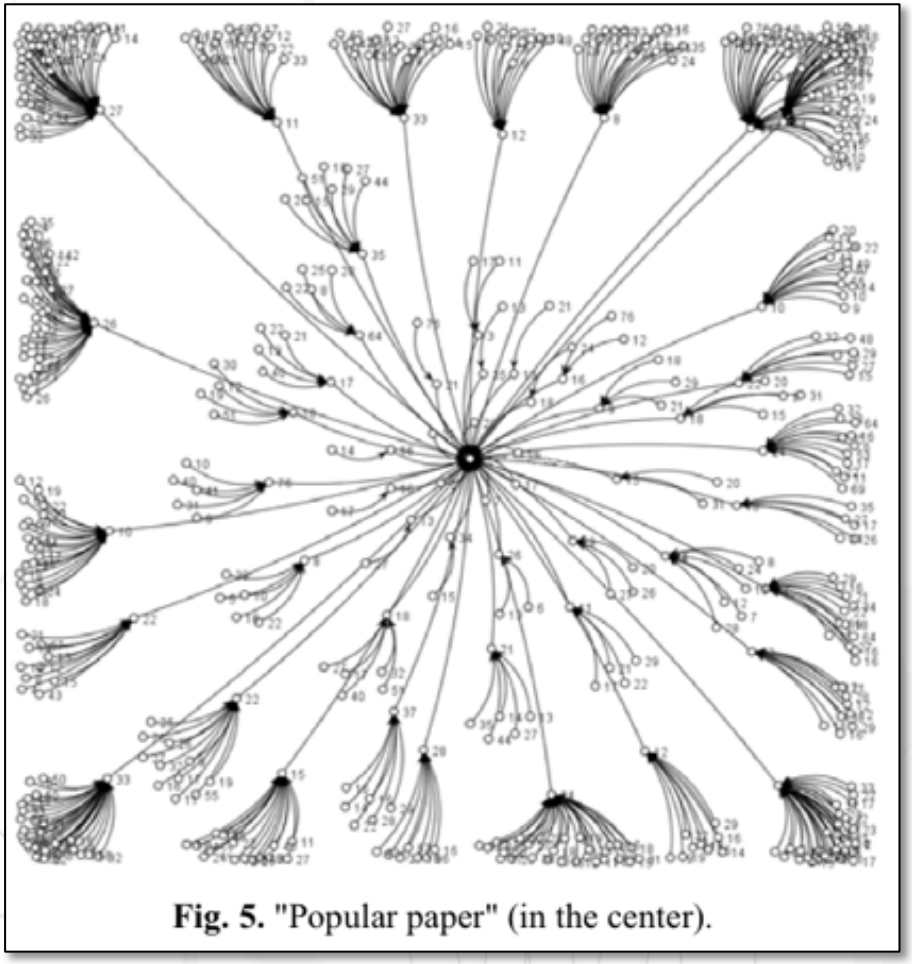


Fig. 5. "Popular paper" (in the center).

LiquidPub

2010 Krapivin et al.

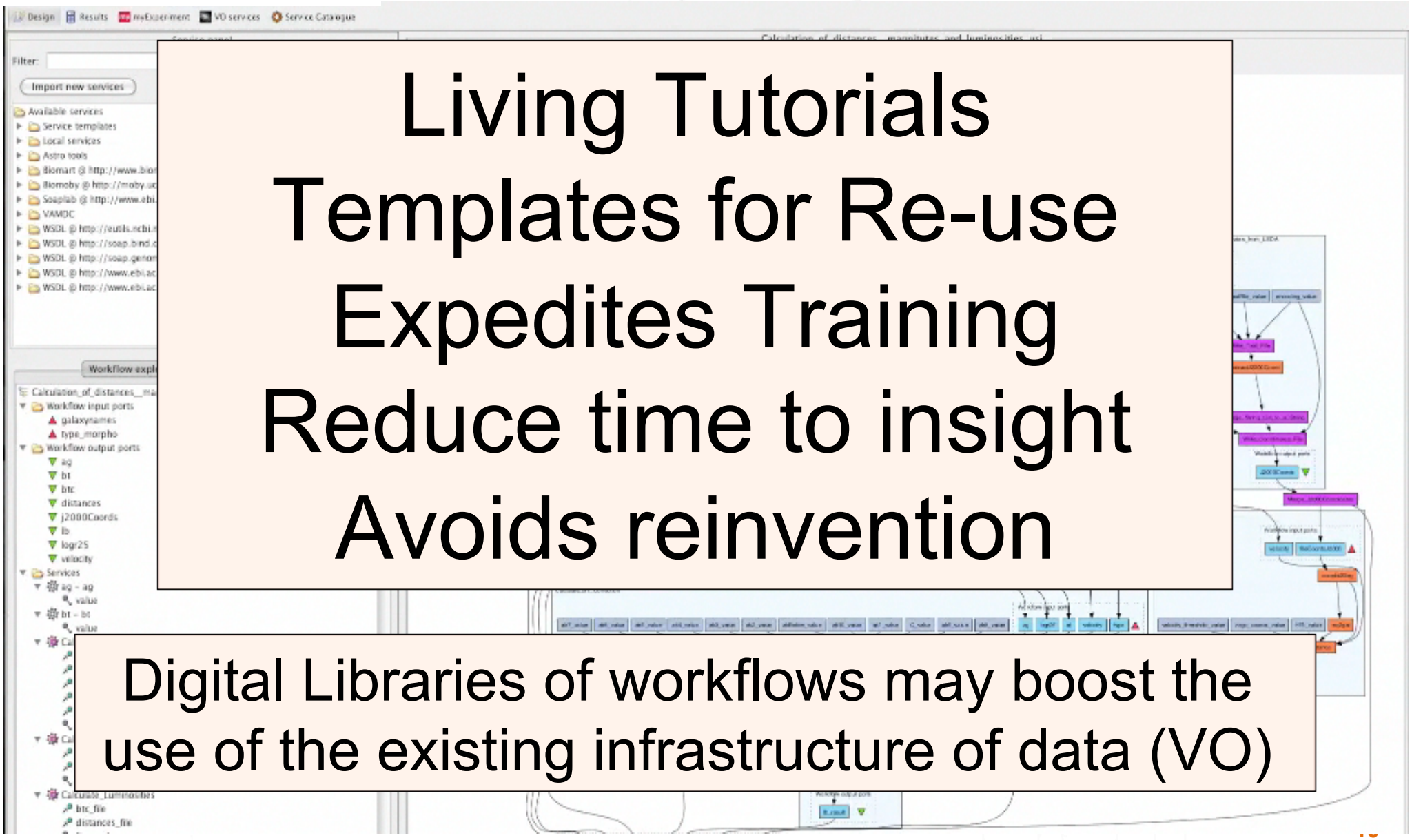
Paper discovery: the social dimension

EU funded FP7 STREP Project

December 2010 – December 2013

1. Intelligent Software Components (**ISOCO**, Spain)
2. University of Manchester (**UNIMAN**, UK)
3. Universidad Politécnica de Madrid (**UPM**, Spain)
4. Poznan Supercomputing and Networking Centre (**PSNC**, Poland)
5. University of Oxford (**OXF**, UK)
6. Instituto de Astrofísica de Andalucía (**IAA**, Spain)
7. Leiden University Medical Centre (**LUMC**, NL)

Scientific Workflows



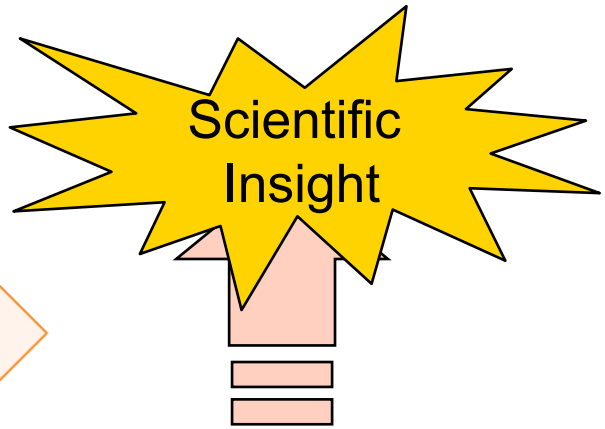
Living Tutorials
Templates for Re-use
Expedites Training
Reduce time to insight
Avoids reinvention

Digital Libraries of workflows may boost the use of the existing infrastructure of data (VO)

Scientific Workflows

Survey in the domain of astrophysical workflows

- Personal script-based recipes
  - Python, IDL, Software..
- Multi-archive VO recipes
  - Euro-VO, IVOA..
- Internal group developments
  - GRID, Clusters, Specific knowledge
- Processing pipelines
  - Facilities provide data, comp… …ture, tools..

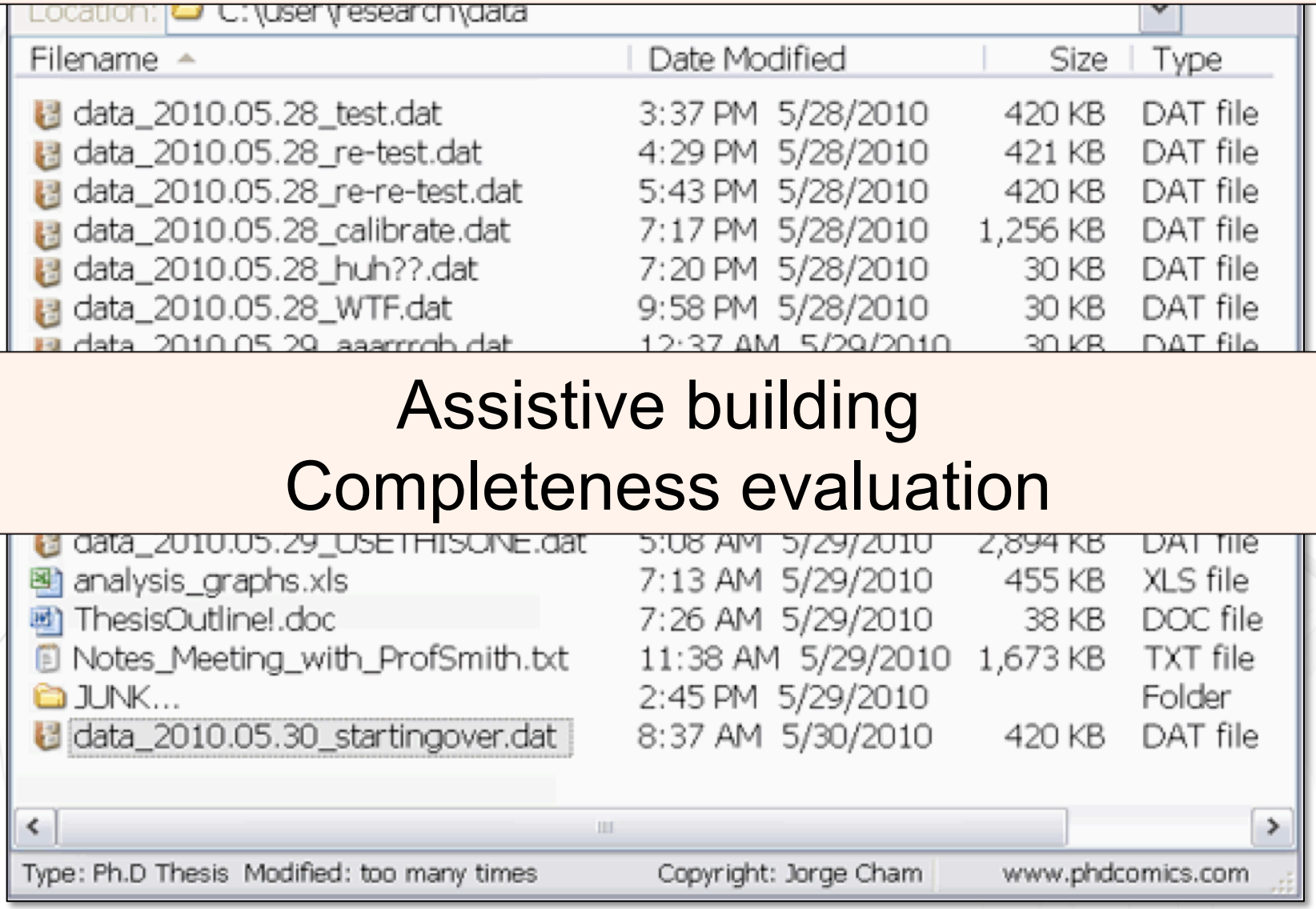


Hand-crafted workflows

Accessible
Shareable
Reusable
Adaptable
Understandable

- **Clarity** (workflows) for re-use and re-porpuse **vs. automation** (pipelines)
- A black box is not **re-usable**, cannot be broken into parts
- **Reproducibility vs. industrial** paper publishing

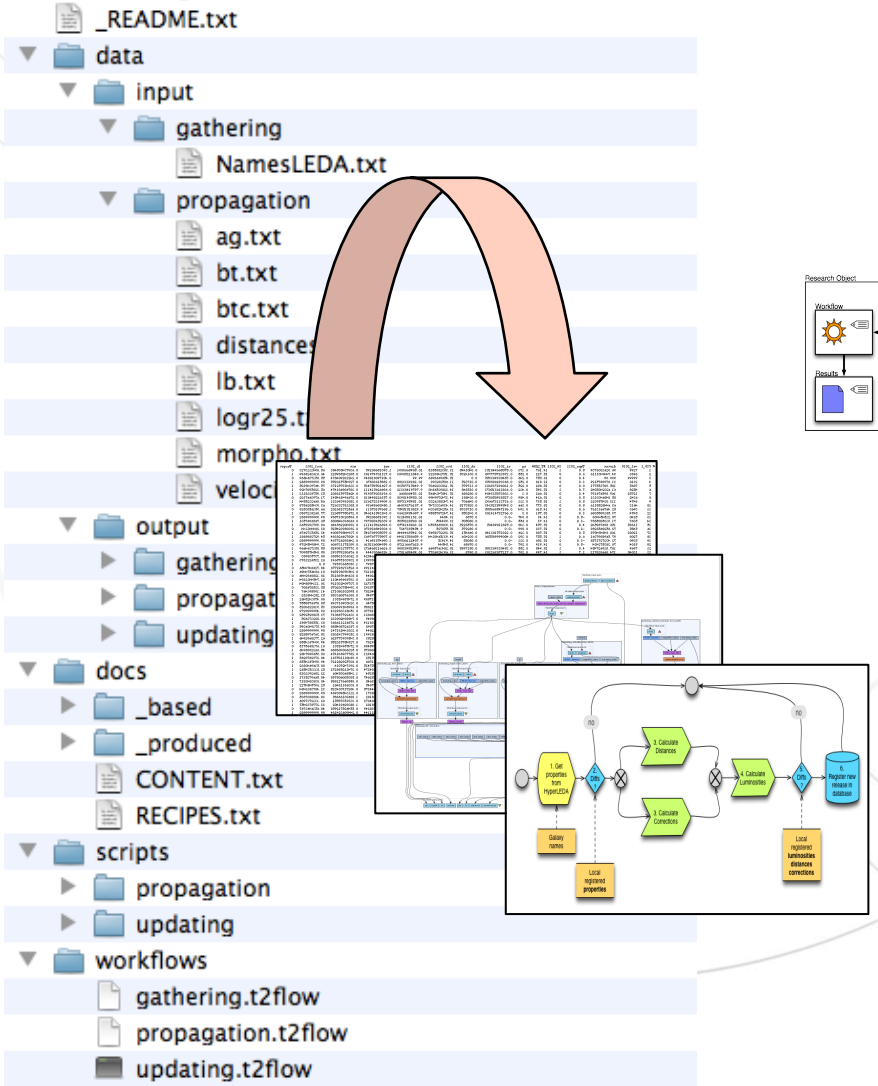
Organization is more sexy than automation

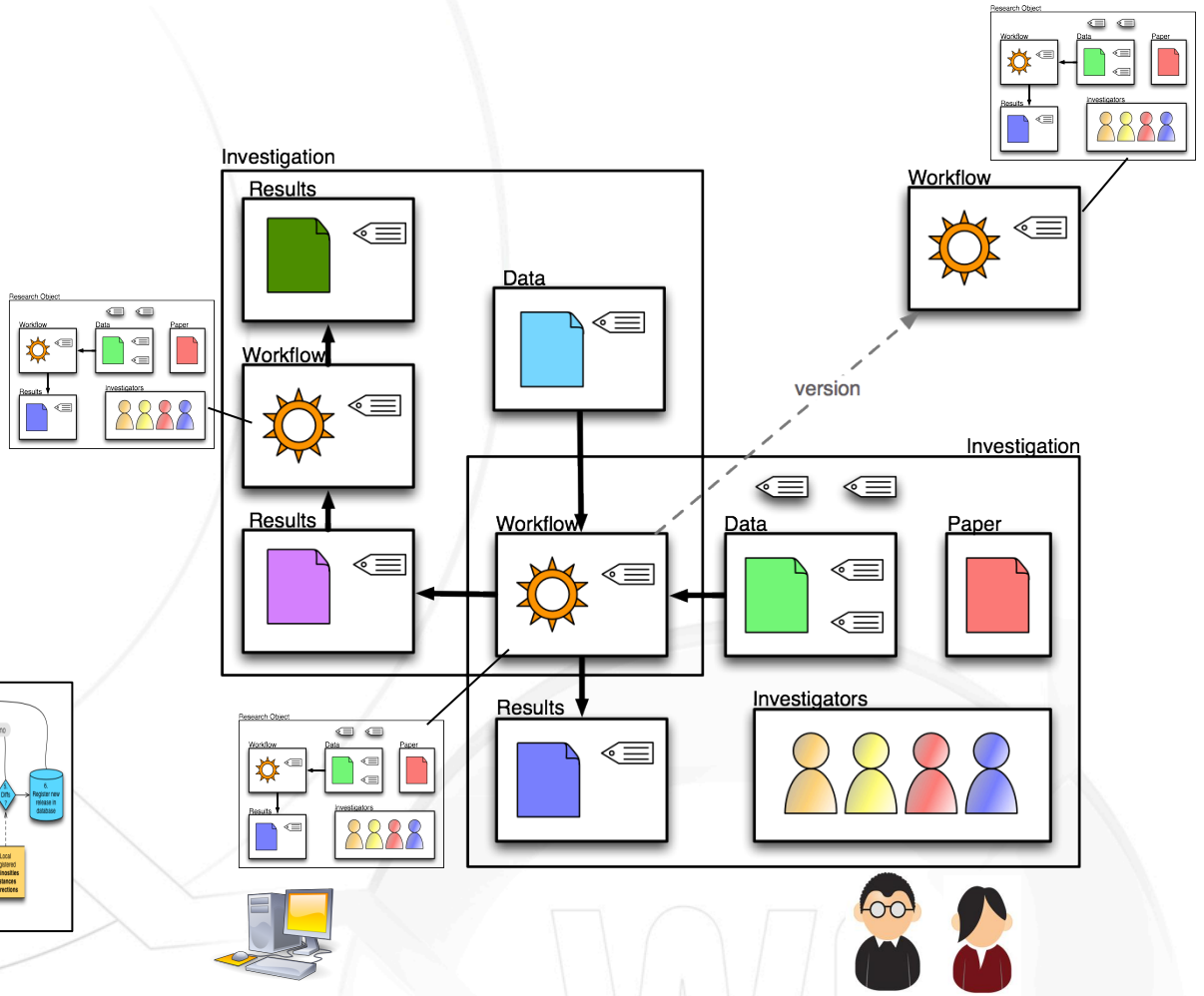


Assistive building
Completeness evaluation

Research Objects

Expose experiment in a structured way in order to be understood





Technical Objects

Distributed

Social Objects

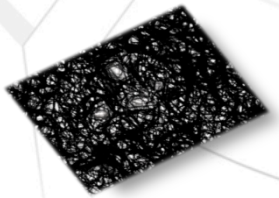
Similar initiatives in Astronomy

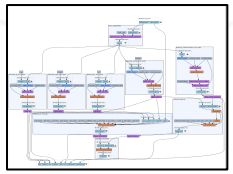
- **Semantic curation** of digital objects
  - CDS Centre Données Strasbourg
  - US Virtual Astronomical Observatory
  - SAO/NASA ADSLabs
- **Workflow users platforms**
  - Cyber-SKA
  - IceCore
  - Montage
  - Astro-WISE
  - Helio-VO
- **Semantically auto descriptive WS**
  - Workflows VO-France

ADSLabs Initiative

ADO Linked Components

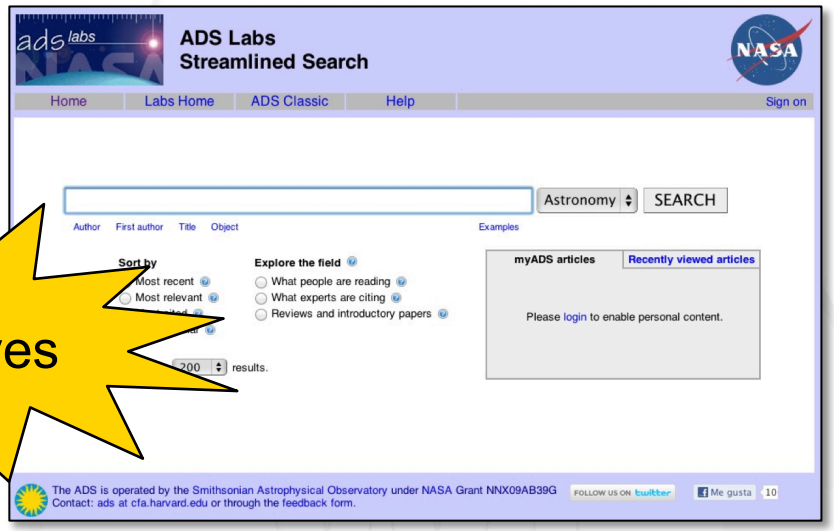
- Authors
- Publications
- Journals
- Objects SIMBAD
- Tabular data behind the plots CDS
- ASCL reference of used software
- Observing time Proposals
- Used facilities, surveys or missions





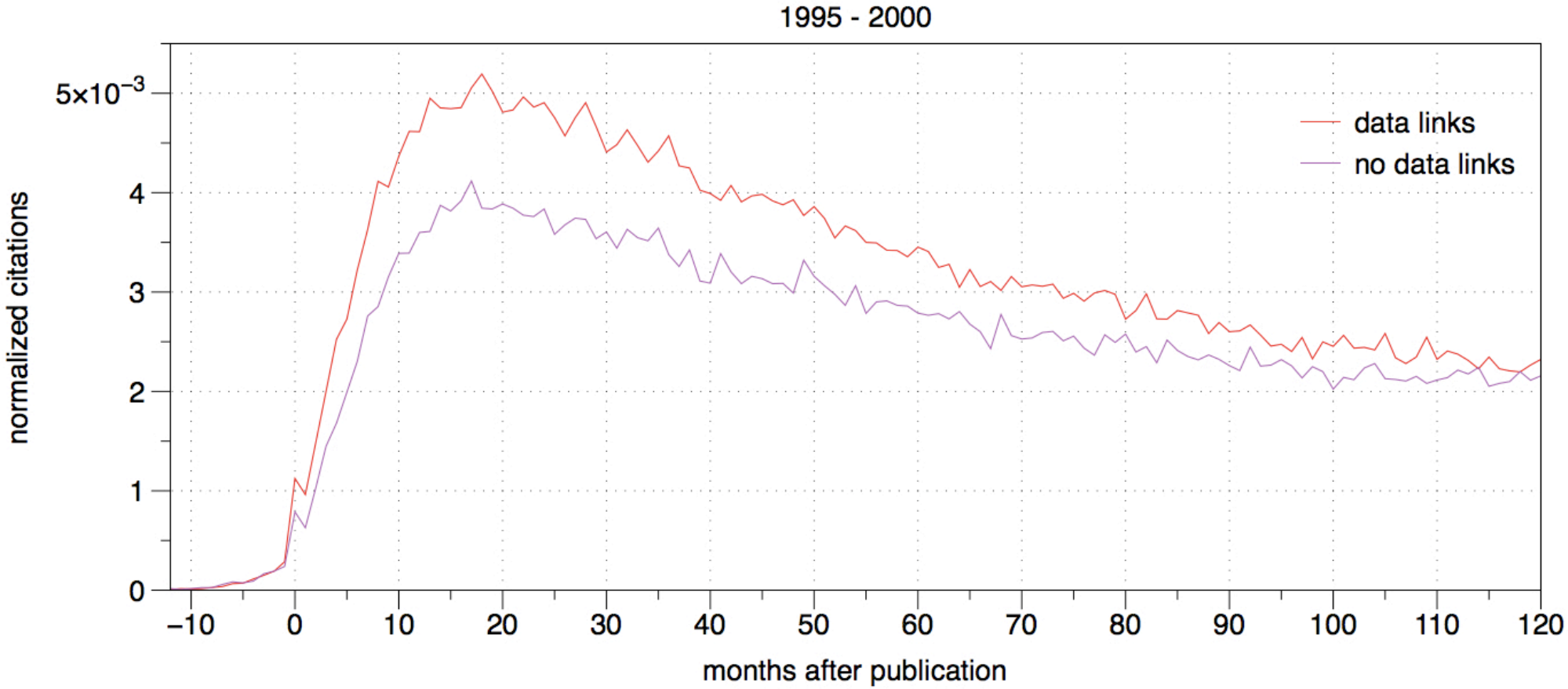
Incentives

http://labs.adsabs.harvard.edu/



The Incentive

Papers with data links are cited more than those without

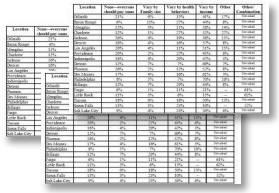


Effect of E-printing on Citation Rates in Astronomy and Physics
2006. Edwin A. Henneken et al.

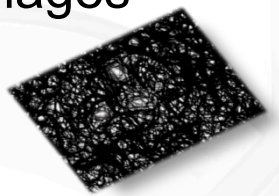
The Wf4Ever Project

- Development of **AstroTaverna plugins** to access and manage VO data
- Development **Golden Exemplars** of astrophysical Workflows and Research Objects that use the Wf4Ever technological support

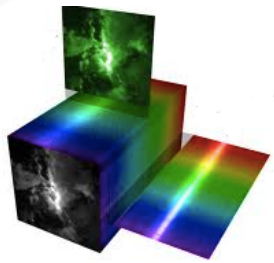
  - **Curation** of physical quantities in 1D catalogues
    - Data retrieved from external repositories and stored locally
    - Only local processes for calculations



  - **Environment and Modelling** from 1D catalogues and 2D images
    - Data retrieved from external repositories (SDSS DR7)
    - Local software and external web services as processes

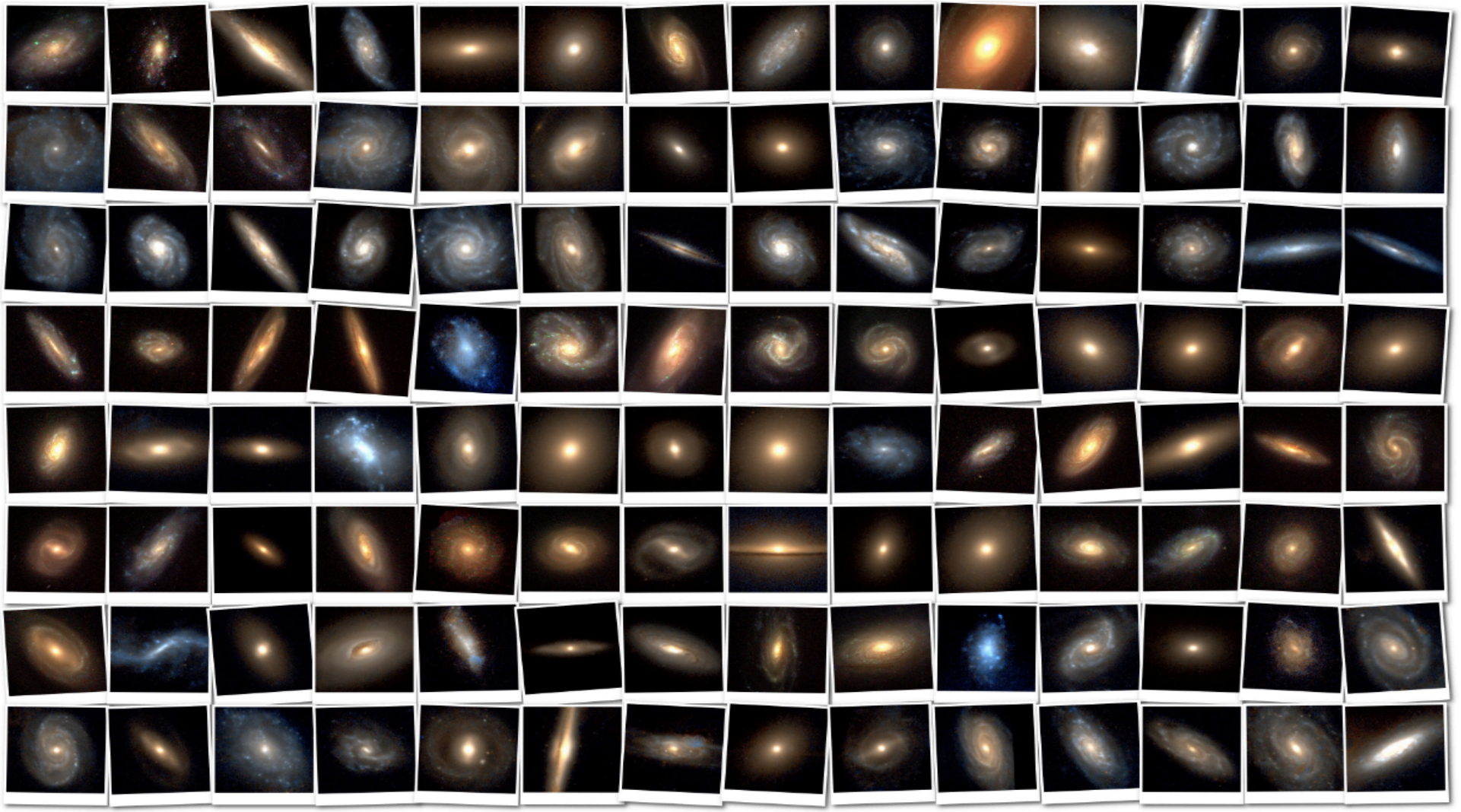


  - **Modelling and Analysis** of 3D formatted data
    - Only external data and processes



Astronomical Research Objects in Action

Curation by inspecting propagation of changes in quantities



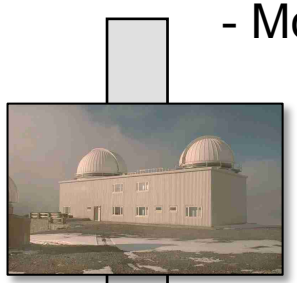
Credit: Zsolt Frei and James E. Gunn. The Galaxy Catalog

AMIGA Catalog

Panchromatic properties for a sample of the most isolated nearby galaxies

How is the User DB affected ?

- Changes in External DB
- Modifications in Calculations



Evaluate variations with time

- Modified External Quantities
- Affected User DB Quantities

Update

External DB

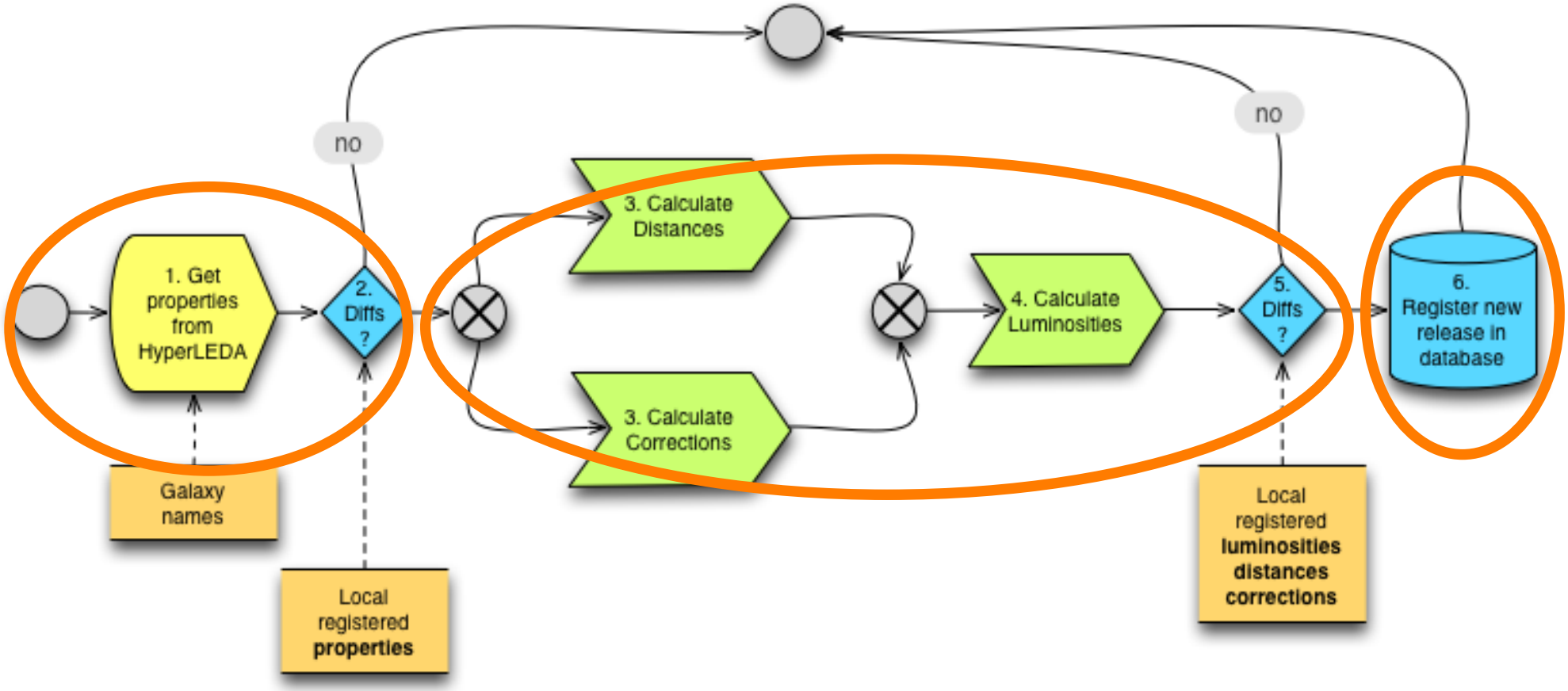
Calculations

User DB

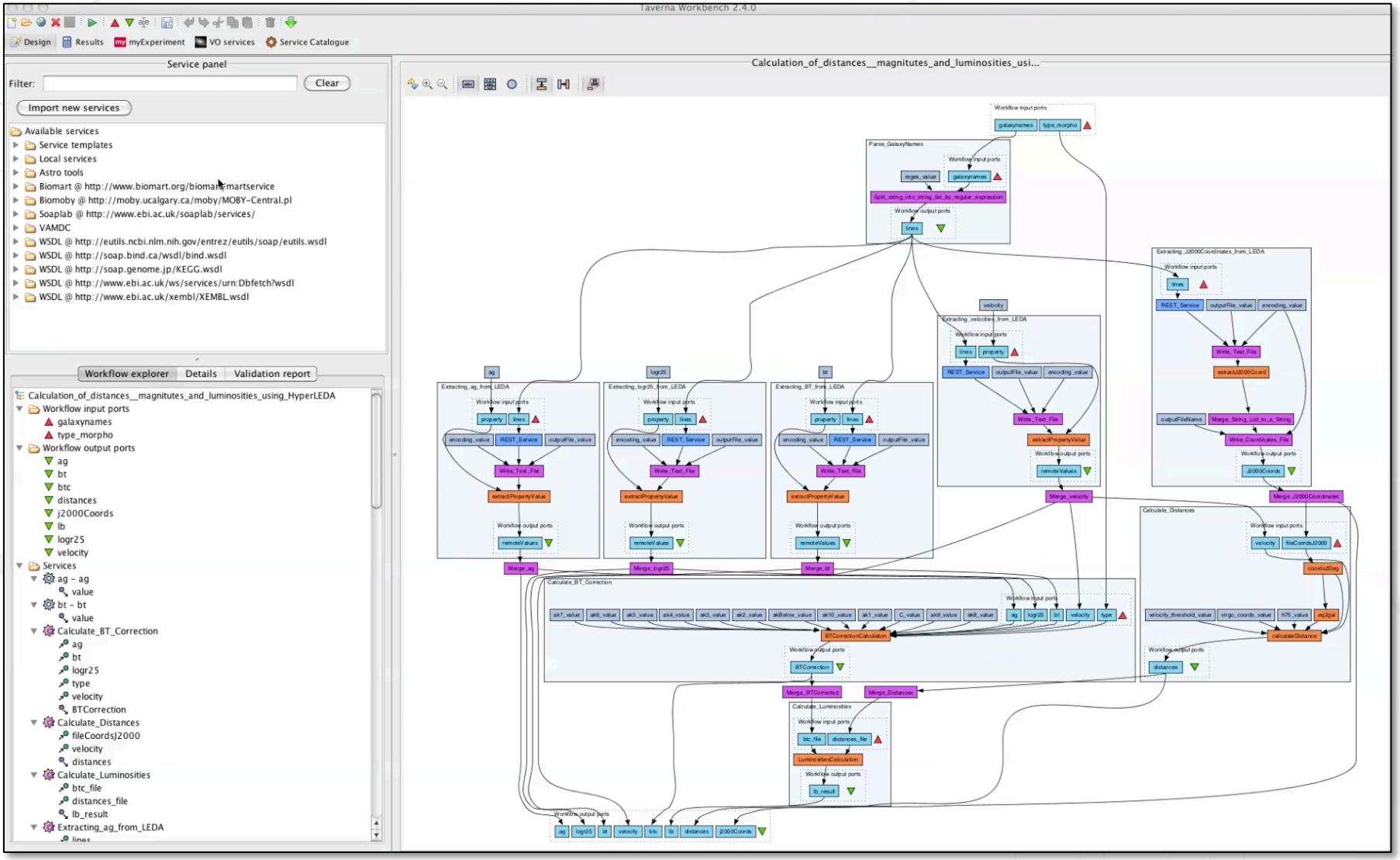
Astronomical Research Objects in Action

Curation by inspecting propagation of changes in quantities

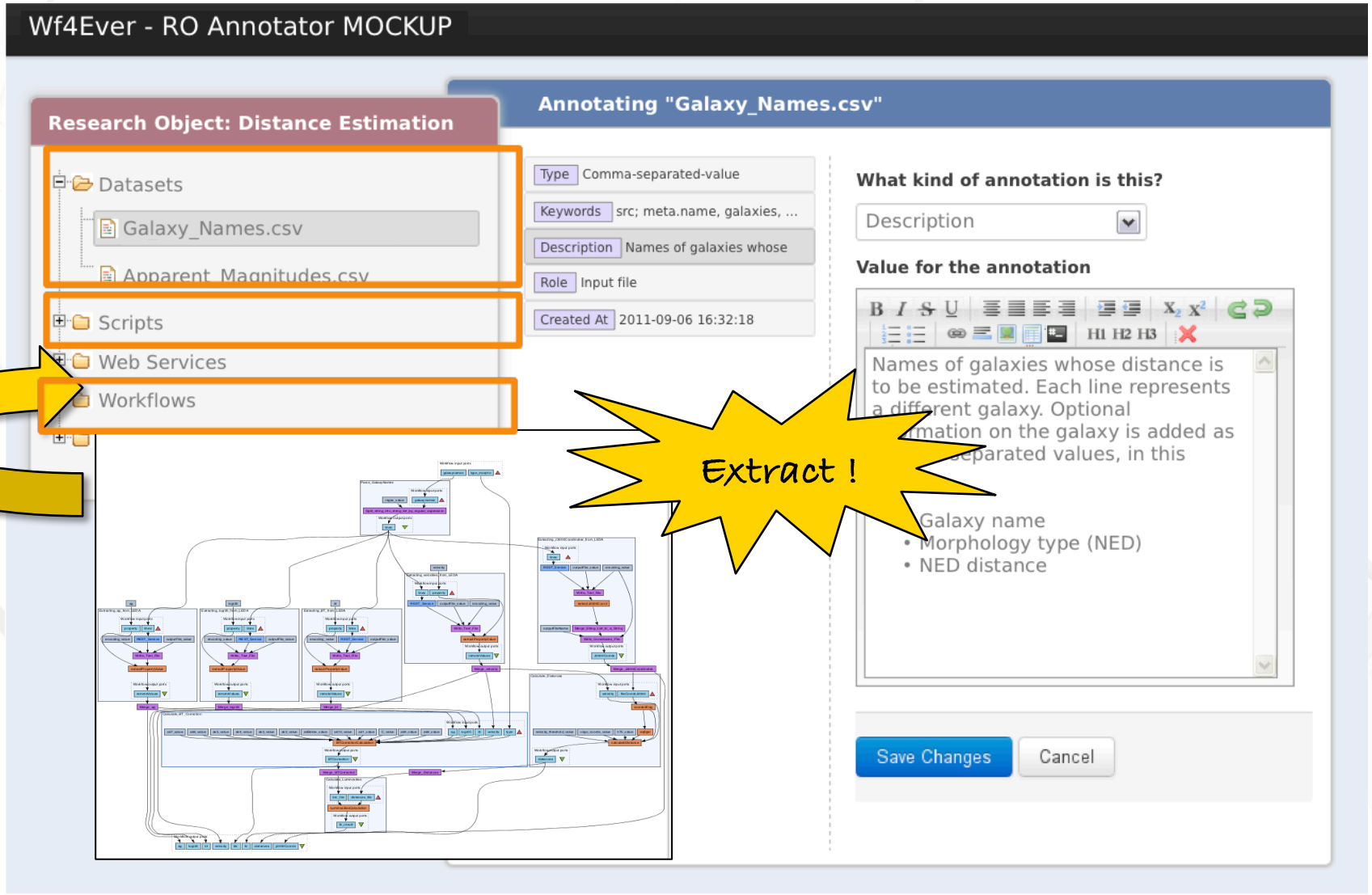
Multi-workflow Research Object



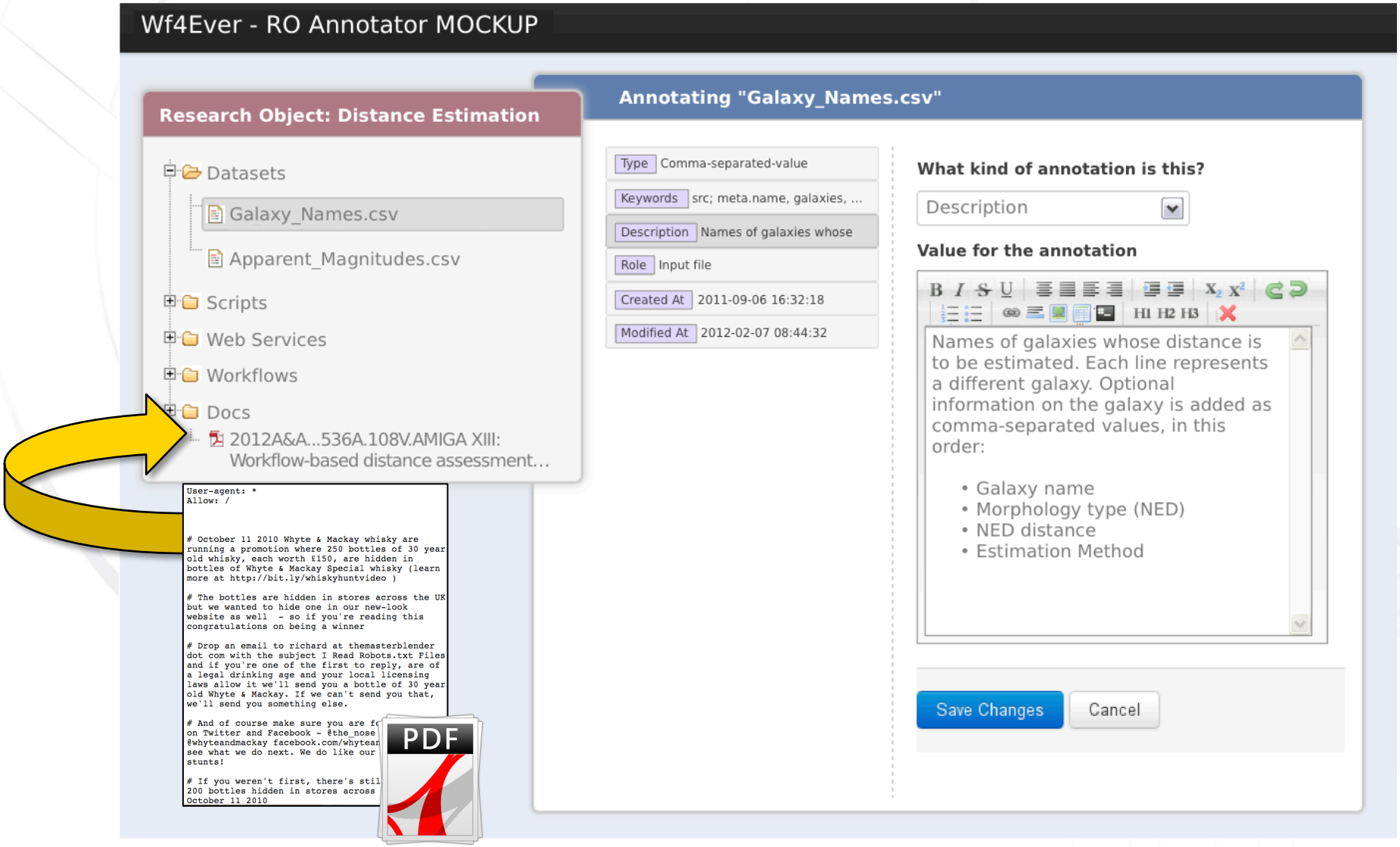
Create, annotate and run a workflow



Populate the Research Object and annotate

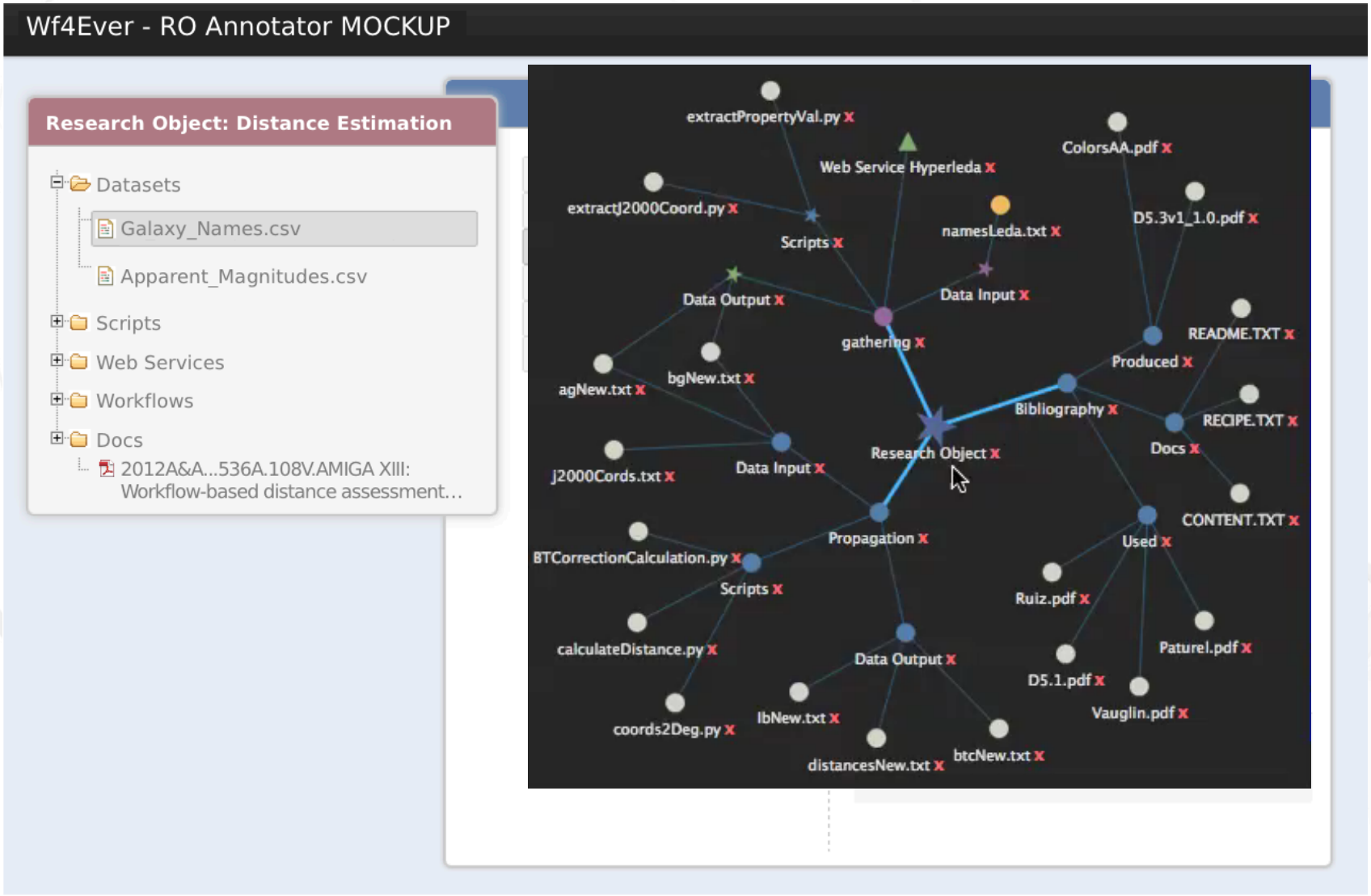


Add documents and references

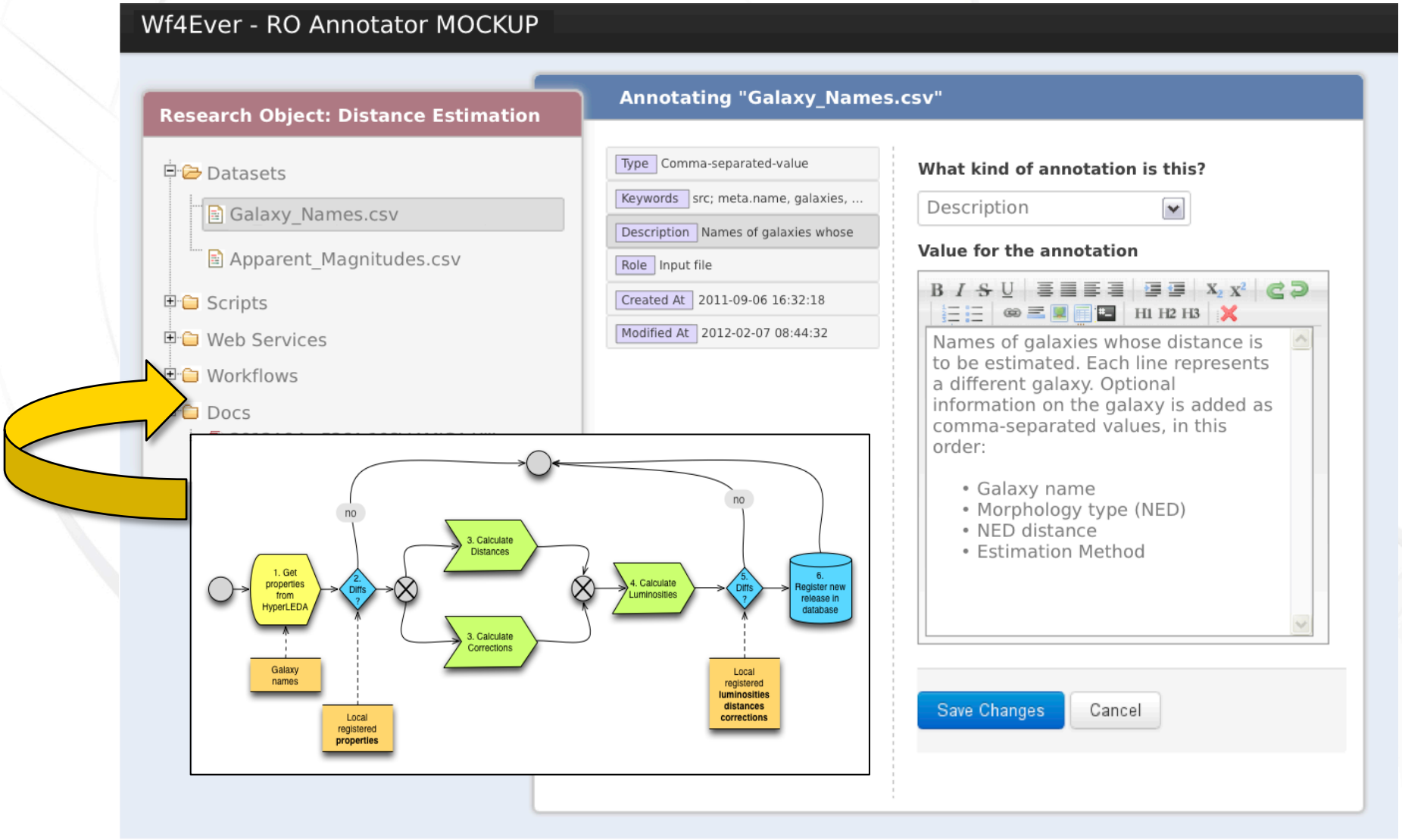


Astronomical Research Objects in Action

Create and explore relations among components

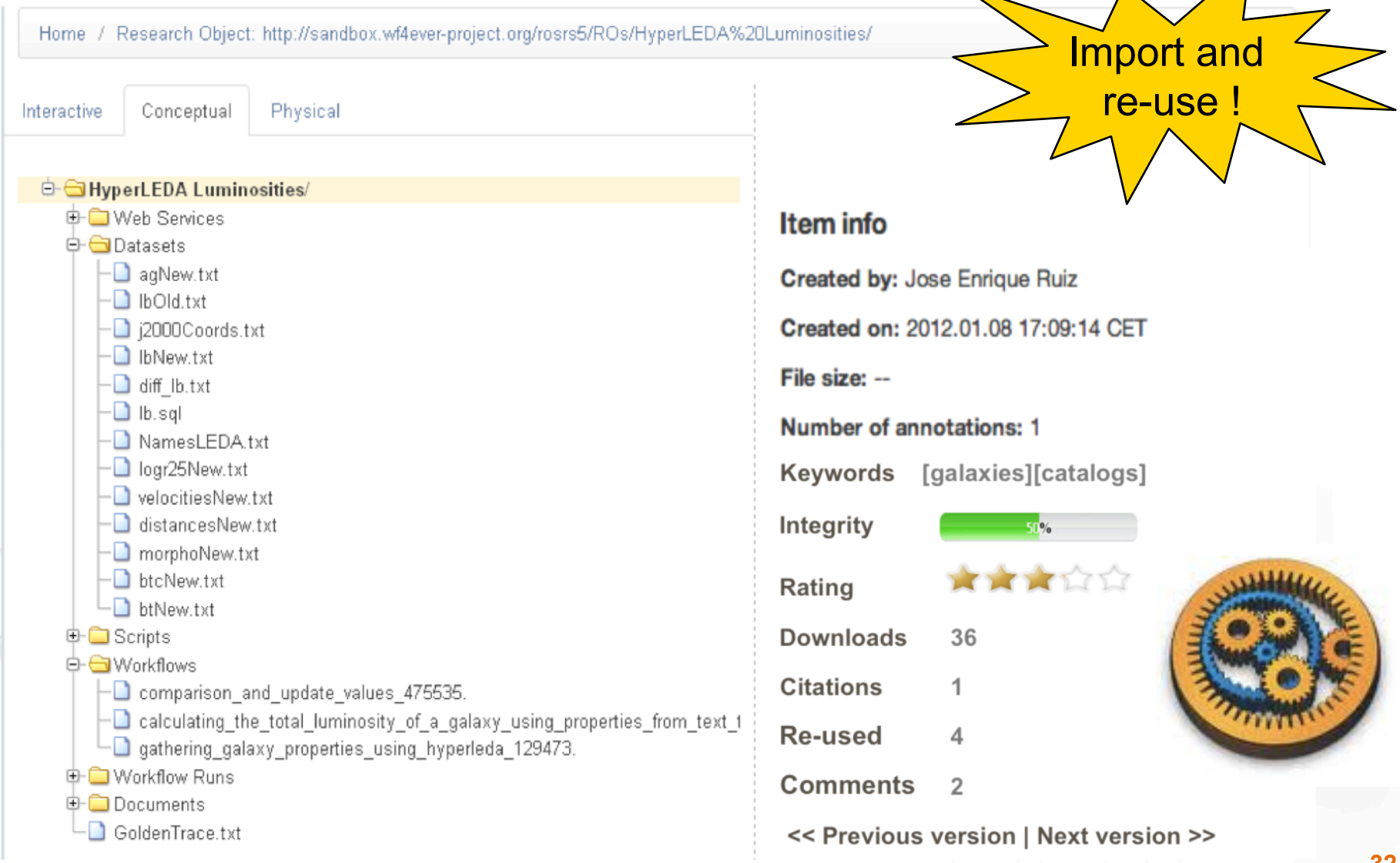


Add schema of the experiment



Astronomical Research Objects in Action

Publication for later discovery



Curation by inspecting propagation of changes in quantities

- Taverna 2.3
- MyExperiment Pack
  - *http://www.myexperiment.org/packs/231*

Related Publication

The AMIGA sample of isolated galaxies XI.

A First Look at Isolated Galaxy Colors

2012 A&A 540, A.47

How NOT to be a good e-astronomer

- Search the beautiful plot for high impact instead of real scientific results
- Write a obscure paper, do not say clearly how to reproduce the results
- Do things quickly and forget about them once you've submitted the paper
- Be untidy, spread your code and data in a variety of formats, folders and disks
- Practise the "data mine-ing" – input data are mine
- Practise the "data flirting" – call me if you would like to have more
- Do not provide data results, including the plots is just fine
- Always cite the same authors and papers or those that cite you
- Do not cite other resources than papers, neither provide their URL links
- Do not search info on Internet with other tools than ADS or arXiv
- Work alone and email/phone one friend if you have any doubt

http://amiga.iaa.es/p/212-workflows.htm
http://www.wf4ever-project.org
jer@iaa.es
bultako