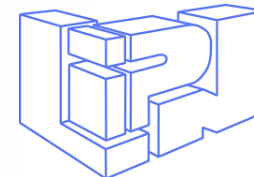


From analysis of requirements to the first experiments in cloud data management and scientific programs

What happens in the USPC french university consortium?

PREDONx 2015 : Atelier sur la
Préservation des Données Scientifiques
Mercredi 9 décembre 2015,
Observatoire Astronomique de Strasbourg

leila.abidi@lipn.univ-paris13.fr
christophe.cerin@lipn.univ-paris13.fr
marie.lafaille@univ-paris.13.fr



Context



Scientists are spending most of their time manipulating, organizing, finding and moving data, instead of researching. And it's going to get worse.

" (DoE Office of Science Data Management Challenge) "



All communities are impacted !!

Major preoccupation

- CPU Annual Conference

University 3.0 : new challenges, new scales
in the digital era

***"Establishing infrastructure to deal with
public data produced by research"***

- French Digital Council

Actions to promote the digital Republic

***"Strengthen digital mediation in order to
promote its use by private individuals"***



USPC is in the race!

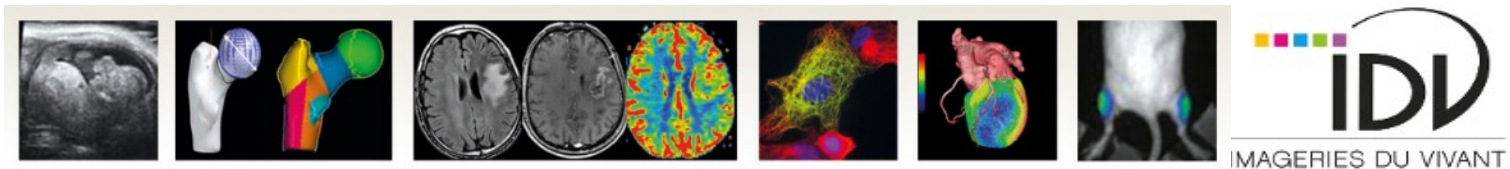
■ USPC platform CIRRUS

To federate the 3 big infrastructures dedicated to research (MAGI P13, S-CAPAD IPGP, Cumulus P5); 3300 cores, 2PB, 500 VM



● Life-imaging program (about 1 year ago)

Atlas creation of medical images (Cumulus, P5)





**How raising awareness and
educating academic communities
on the uses of digital technology?**

I. Preliminary analyses

- Who are the users?
- Data life-cycle?
- Work habits?
- Needs?
- Expectations?
- Fears?



Survey: a powerful tool



- Relatively easy to administer, convenient data gathering
- Can be developed in a short time
- Large amounts of information can be collected....
- from a large number of people (high representativeness)...
- in a short period of time and...
- in a relatively cost effective way.
- Numerous questions can be asked about a subject, giving extensive flexibility in data analysis.
- Advanced statistical techniques available in survey software
- Standardized surveys are relatively free from several types of errors

« Business process » survey (IDV)

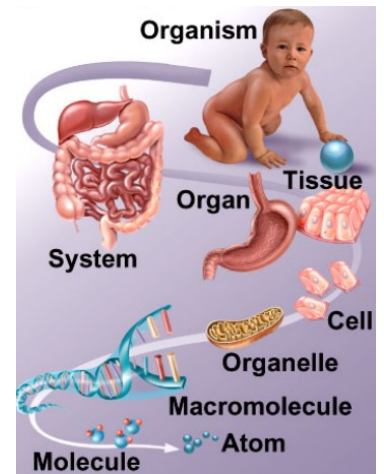
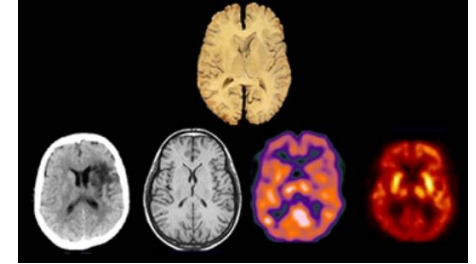
- Atlas: construction of a smart multi-modality multi-scale encyclopedia

Software & infrastructure as services in the distributed architecture

Pre-defined templates of VMs containing pre-installed tools

Possibility of customizing every working environment

Securing communication protocol [VMs \rightleftharpoons Data center]



Equipement: storage capacity?

Formats/software for image display and analysis?

Actual practices?

« Business process » survey



**Survey
software**

Responses (as of March 30th 2015):

- 19 laboratories (image processing, biology, physics, chemistry, psychology, pharmacy)
- 25 teams/imaging platforms



« Business process » survey

Top priorities:

■ Inventory of instruments

- Formats?
- Software?

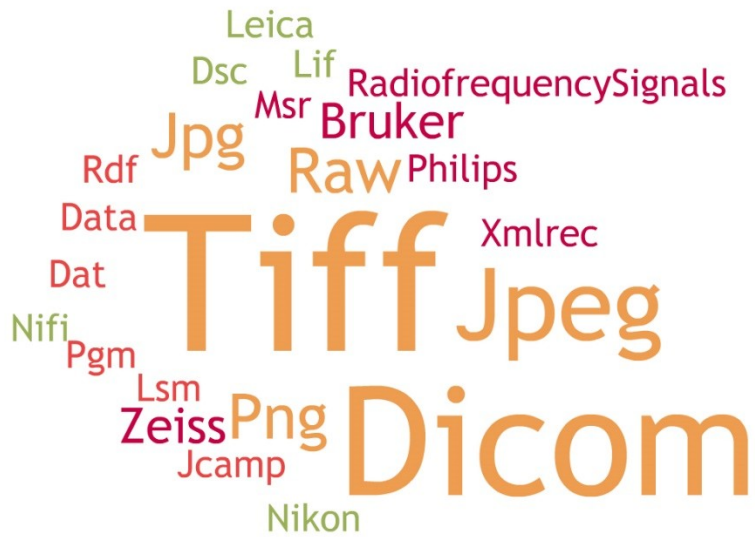
■ Data life-cycle

- Volume?
- Storage?
- Sharing?
- Archiving?

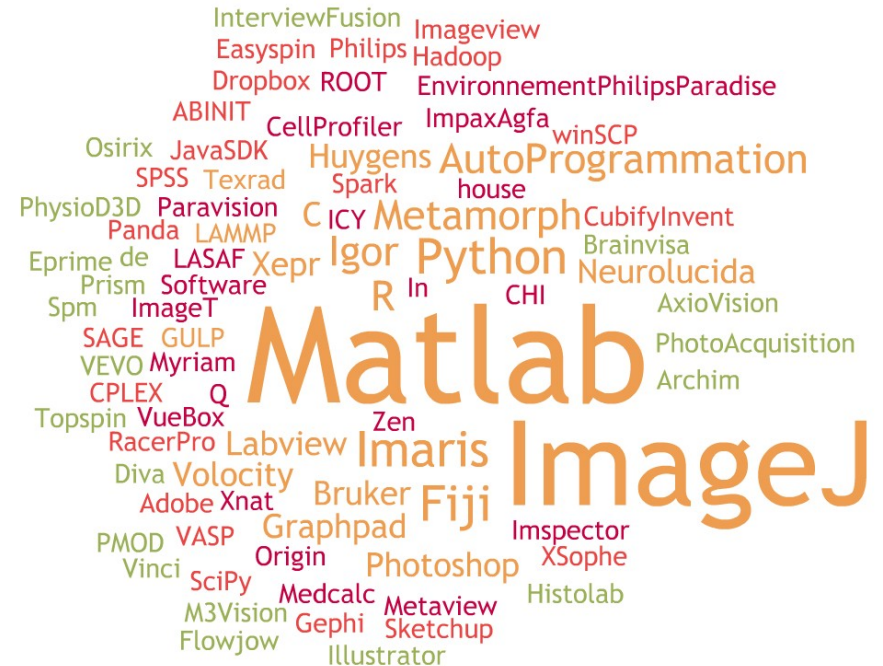


Instruments

Formats



Software



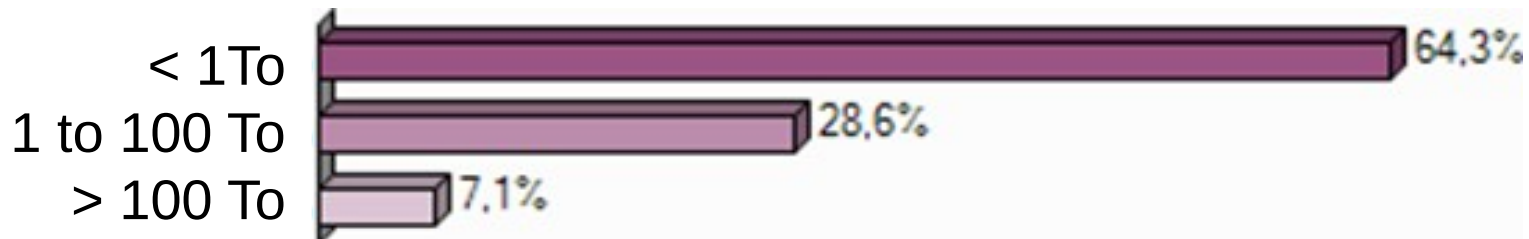
Cloudify the most frequently software

OsiriX (DICOM viewer)

ImageJ (displaying, editing, analyzing, processing, saving)

Data life-cycle

Annual volume of data



➔ Storage capacity estimated: 300 TB

Data preservation (>10 years)



➔ Possible improvement
Loss of potential valuable data to feed the atlas

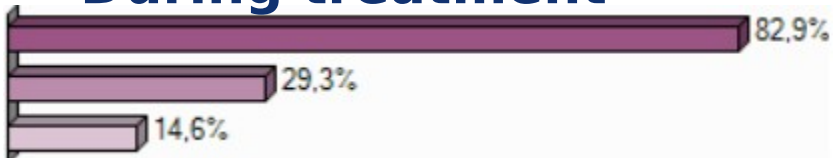
Data life-cycle

Local (hard drive) Network neighbourhood Internet

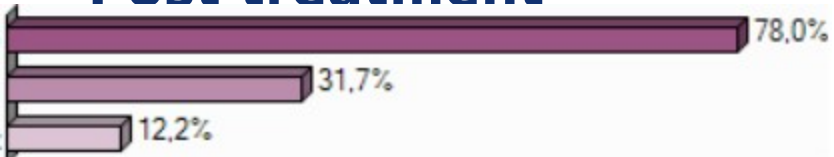
Pre treatment



During treatment



Post treatment



➔ Spatial fragmentation = under-exploited data

➔ Hard drive = high-risk location

VM - IDV

Demo VM IDV

cumulus.parisdescartes.fr

Apple Yahoo! Google Maps YouTube Wikipédia Informations Divers

www.comput... Sunstone Lo... Sunstone: Cl... IEEE Cloud C... Partagés ave... Hosting data... Laboratoire

UNIVERSITÉ PARIS DESCARTES idv-user OpenNebula

VMs Templates Services

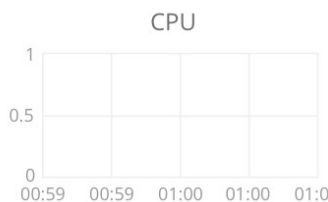
VM-filezilla-transfert

OFF

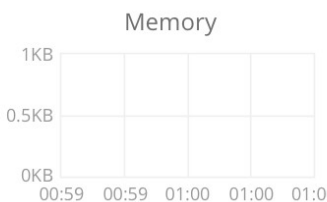
x8 - 8GB - Ubuntu-IDV-disk

172.17.34.65

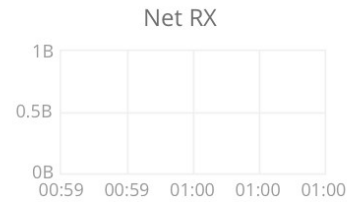
idv-admin 1 Nov - ID: 66



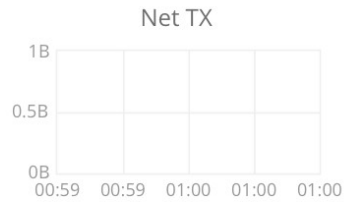
CPU



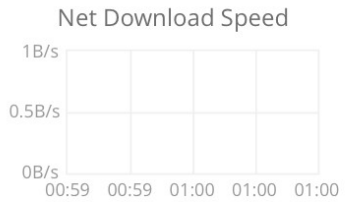
Memory



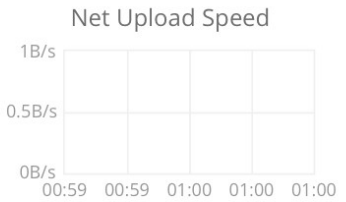
Net RX



Net TX



Net Download Speed



Net Upload Speed

Browser address bar: cumulus.parisdescartes.fr

Navigation menu: Apple, Yahoo!, Google Maps, YouTube, Wikipédia, Informations, Divers

Page tabs: www.comput..., Sunstone Lo..., Sunstone: Cl..., IEEE Cloud C..., Partagés ave..., Hosting data..., Lak






User profile: idv-user, OpenNebula

Logos: Université Paris Descartes

Navigation: VMs, Templates, Services

Create Virtual Machine

Select a Template

System	Group	Saved
<input type="text" value="Search"/>		
IDV Ubuntu template  ubuntu ...	DEBIAN 8 x64 - Minimal  debian ...	WIN2K12R2 x64 English 
IDV Spark template  ubuntu ...	IDV ePad template  ubuntu ...	

Corbeille

Info for Cardio.dcm

Info for Cardio.dcm

File Edit Font

0002,0002 Media Storage SOP Class UID: 1.2.840.10008.5.1.4

0002,0003 Media Storage SOP Inst UID: 1.2.276.0.7230010.3

0002,0010 Transfer Syntax UID: 1.2.840.10008.1.2.2

0002,0012 Implementation Class UID: 1.2.840.10008.1.2.1

0002,0013 Implementation Class UID: 1.2.840.10008.1.2.1

File Edit Font

0002,0002 Media Storage SOP Class UID: 1.2.840.10008.5.1.4

0002,0003 Media Storage SOP Inst UID: 1.2.276.0.7230010.3

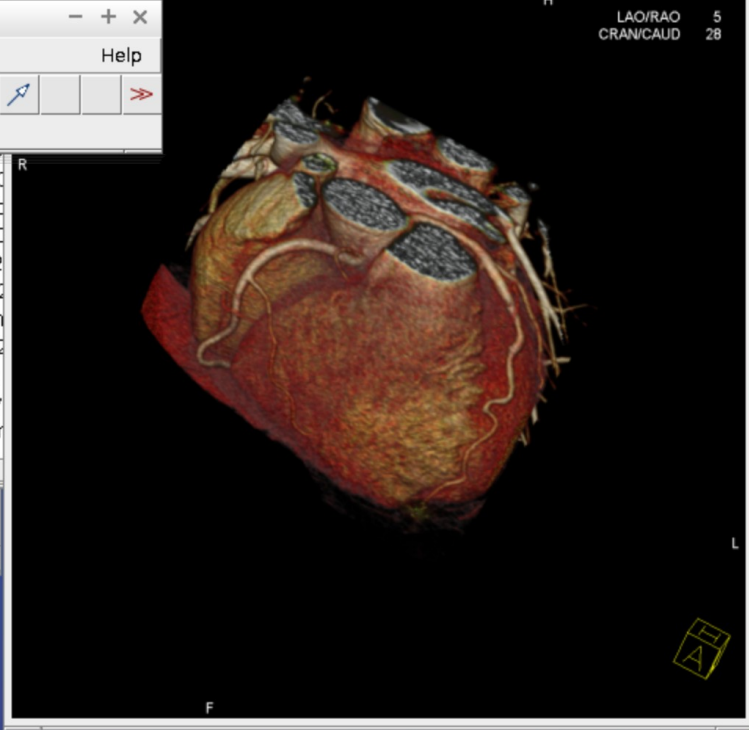
0002,0010 Transfer Syntax UID: 1.2.840.10008.1.2.2

0002,0012 Implementation Class UID: 1.2.840.10008.1.2.1

0002,0013 Implementation Class UID: 1.2.840.10008.1.2.1

Cardio.dcm (50%)

1000x1000 pixels; RGB; 3.8MB



ImageJ

File Edit Image Process Analyze Plugins Window Help

ImageJ toolbar icons including selection tools, crop, rotate, and zoom.

0008,0008 Image Type: DERIVED(SL

0008,0016 SOP Class UID: 1.2.840.1

0008,0018 SOP Instance UID: 1.2.27

0008,0020 Study Date: 20070208

0008,0021 Series Date: 20070208

0008,0022 Acquisition Date: 200702

0008,0023 Image Date: 20070208

0008,0030 Study Time: 120810.671

0008,0031 Series Time: 125519.859

0008,0032 Acquisition Time: 121338

0008,0033 Image Time: 125520.062

0008,0040 Data Set Type:

0008,0041 Data Set Subtype:

0008,0050 Accession Number: 0

0008,0060 Modality: CT

0008,0020 Study Date: 20070208

0008,0021 Series Date: 20070208

0008,0022 Acquisition Date: 200702

0008,0023 Image Date: 20070208

0008,0030 Study Time: 120810.671

0008,0031 Series Time: 125519.859

0008,0032 Acquisition Time: 121338

0008,0033 Image Time: 125520.062

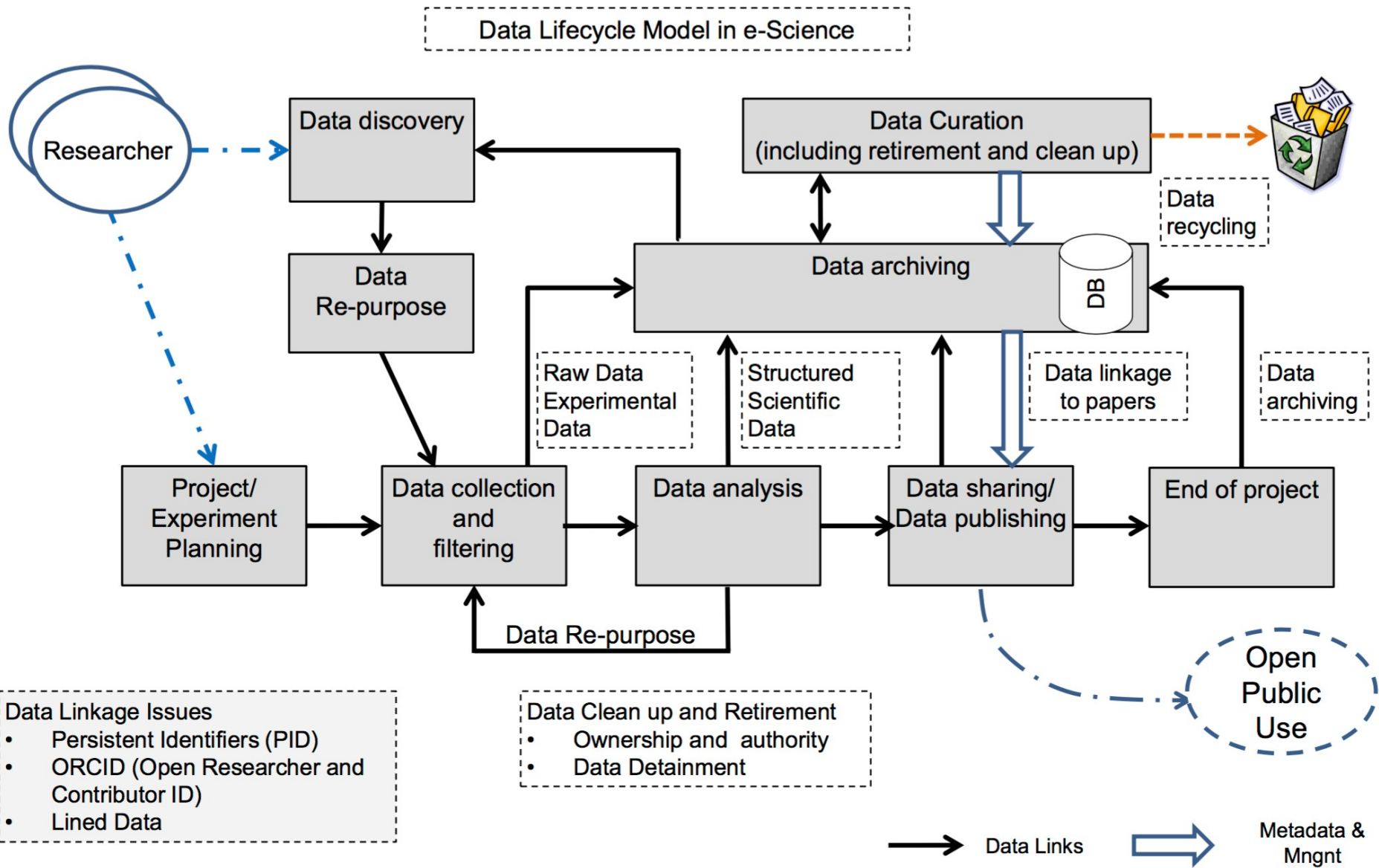
0008,0040 Data Set Type:

0008,0041 Data Set Subtype:

0008,0050 Accession Number: 0

0008,0060 Modality: CT

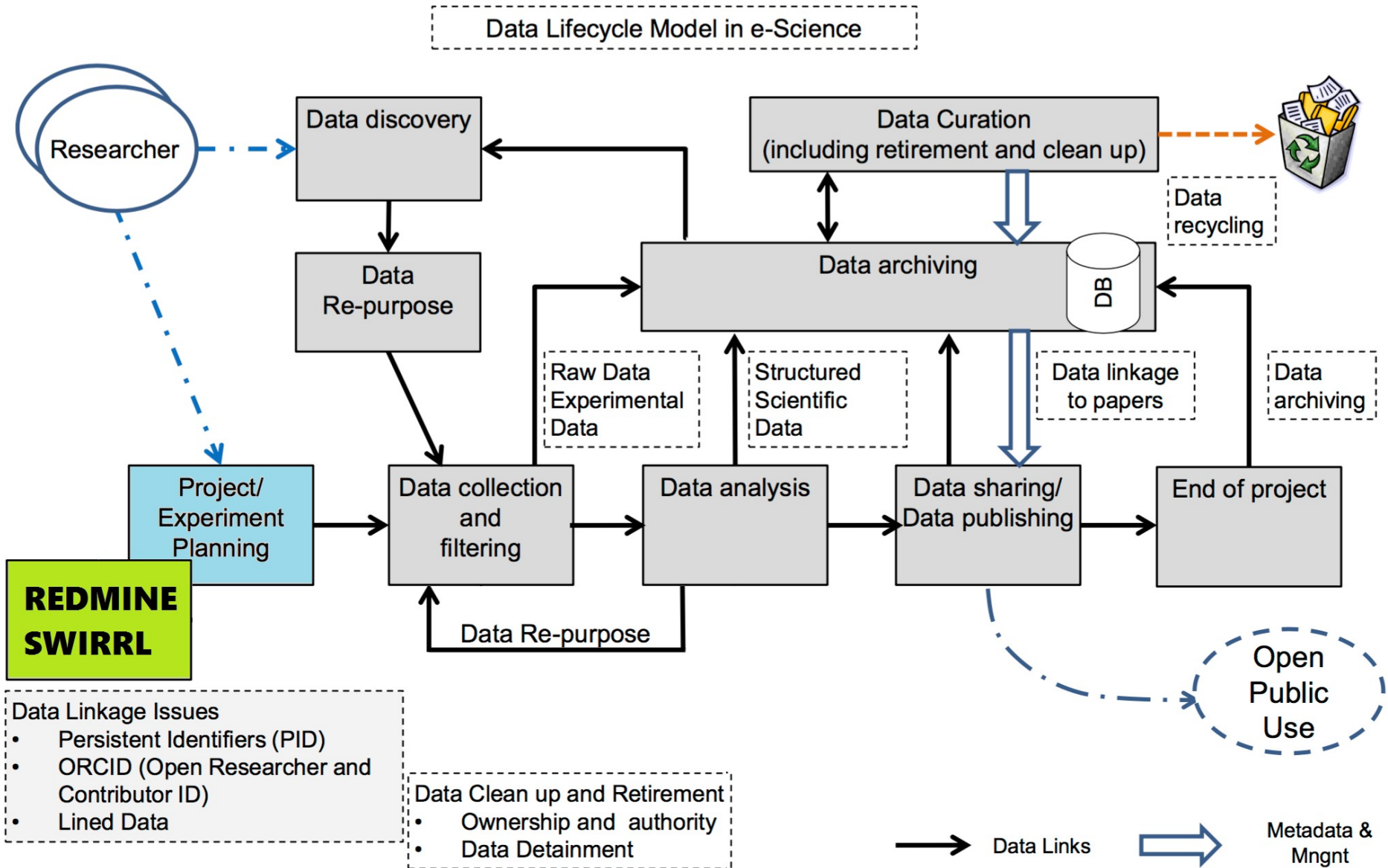
II. The long-term view



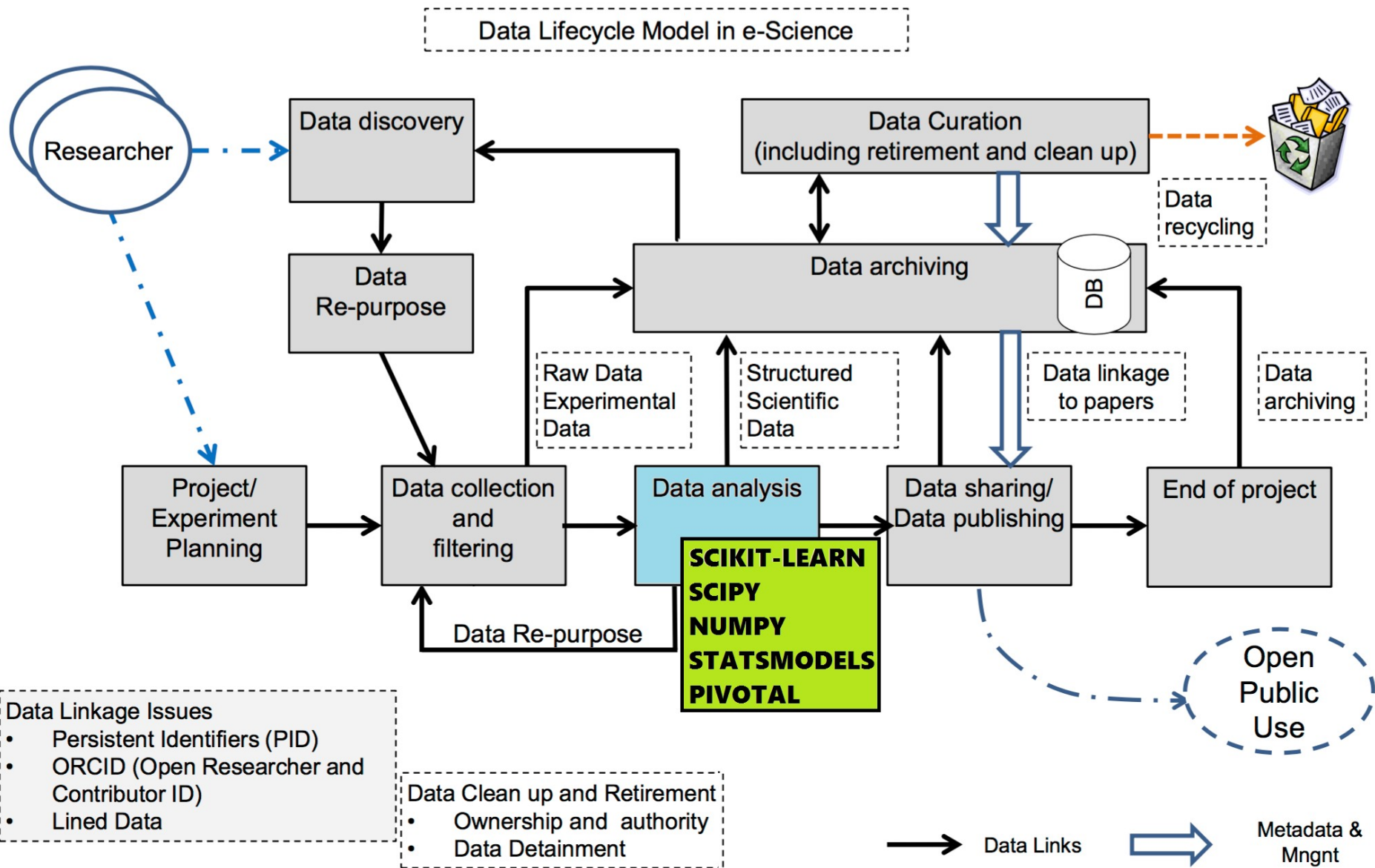
- Data Linkage Issues**
- Persistent Identifiers (PID)
 - ORCID (Open Researcher and Contributor ID)
 - Lined Data

- Data Clean up and Retirement**
- Ownership and authority
 - Data Detainment

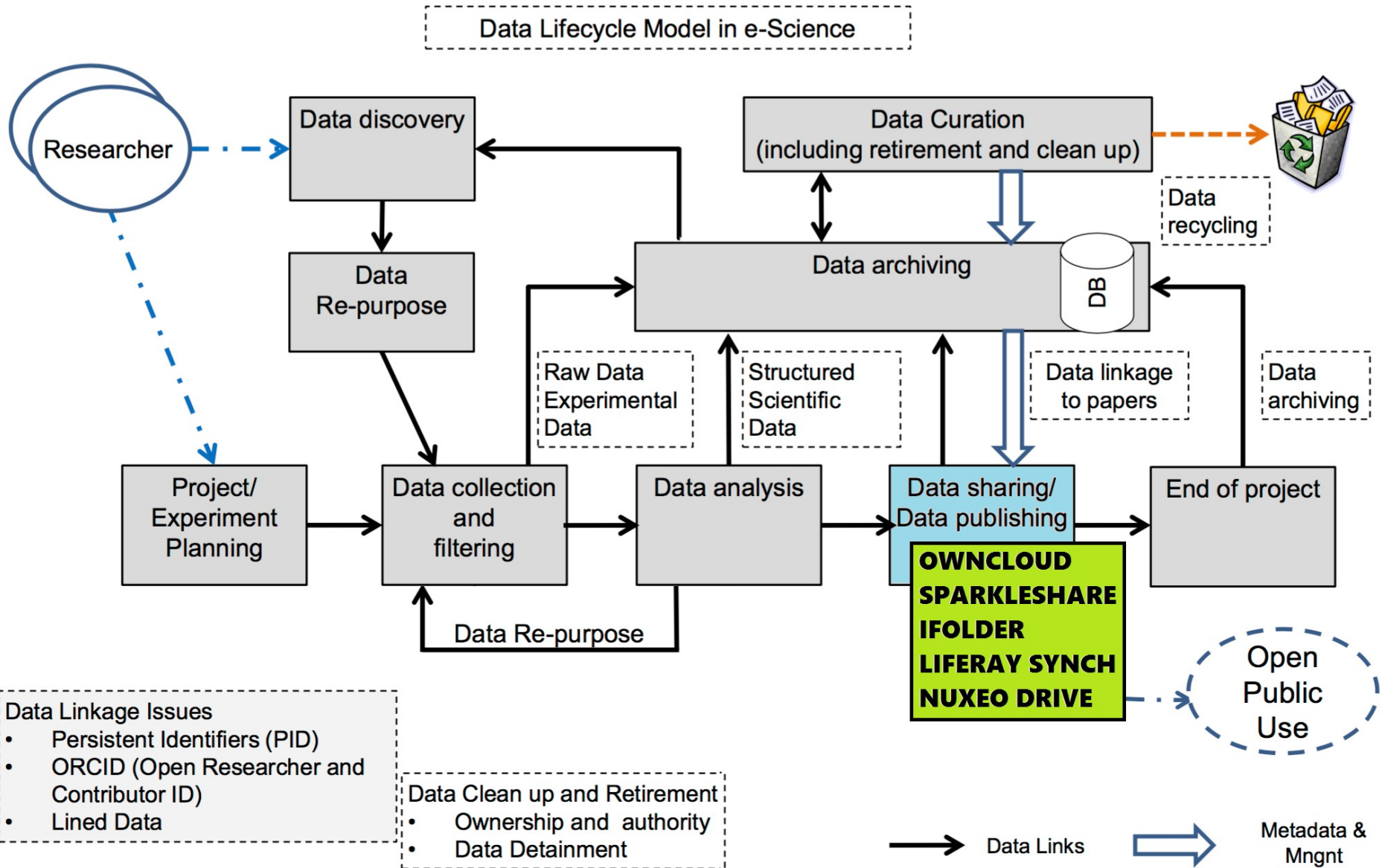
Data life-cycle model



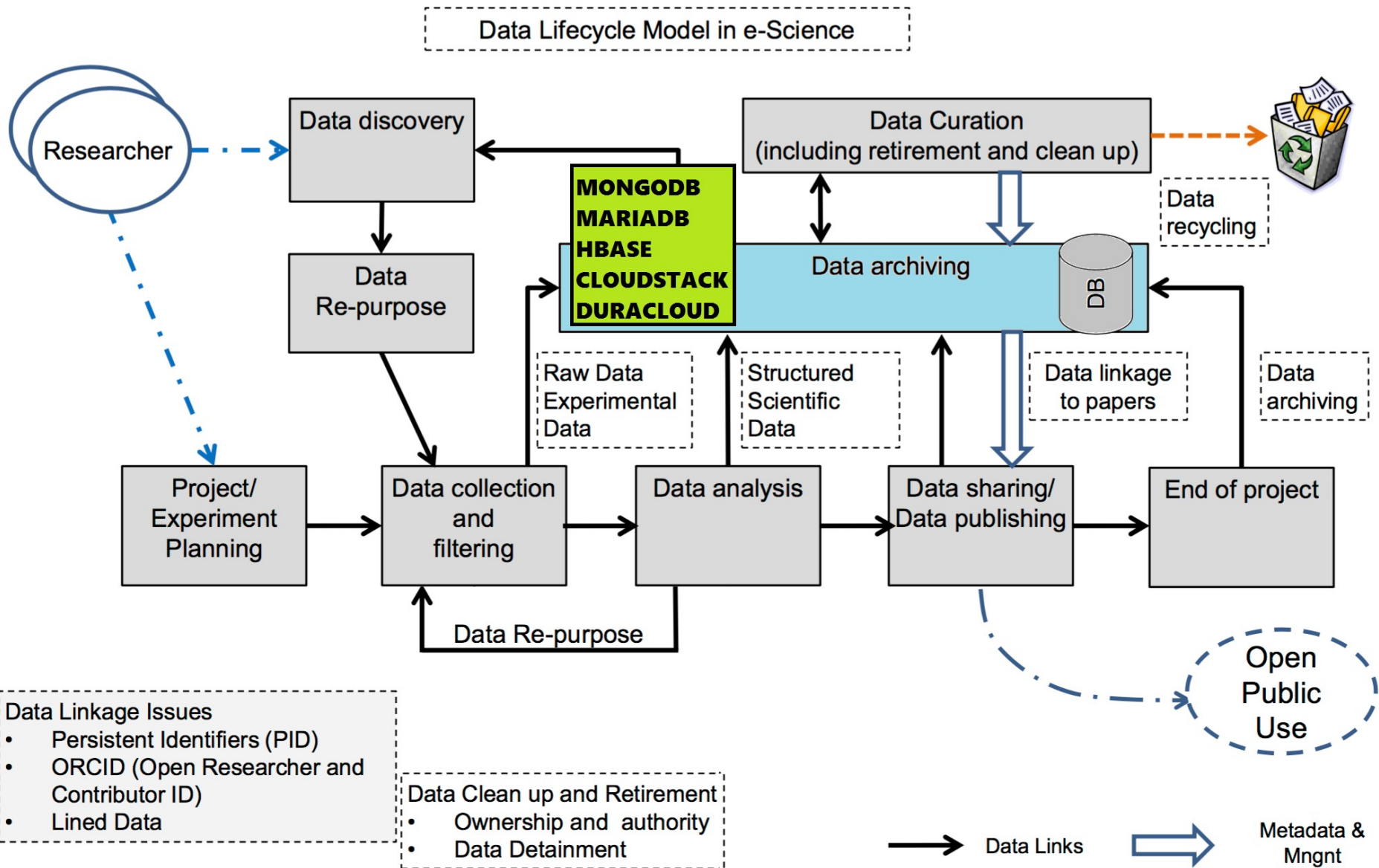
Data life-cycle model



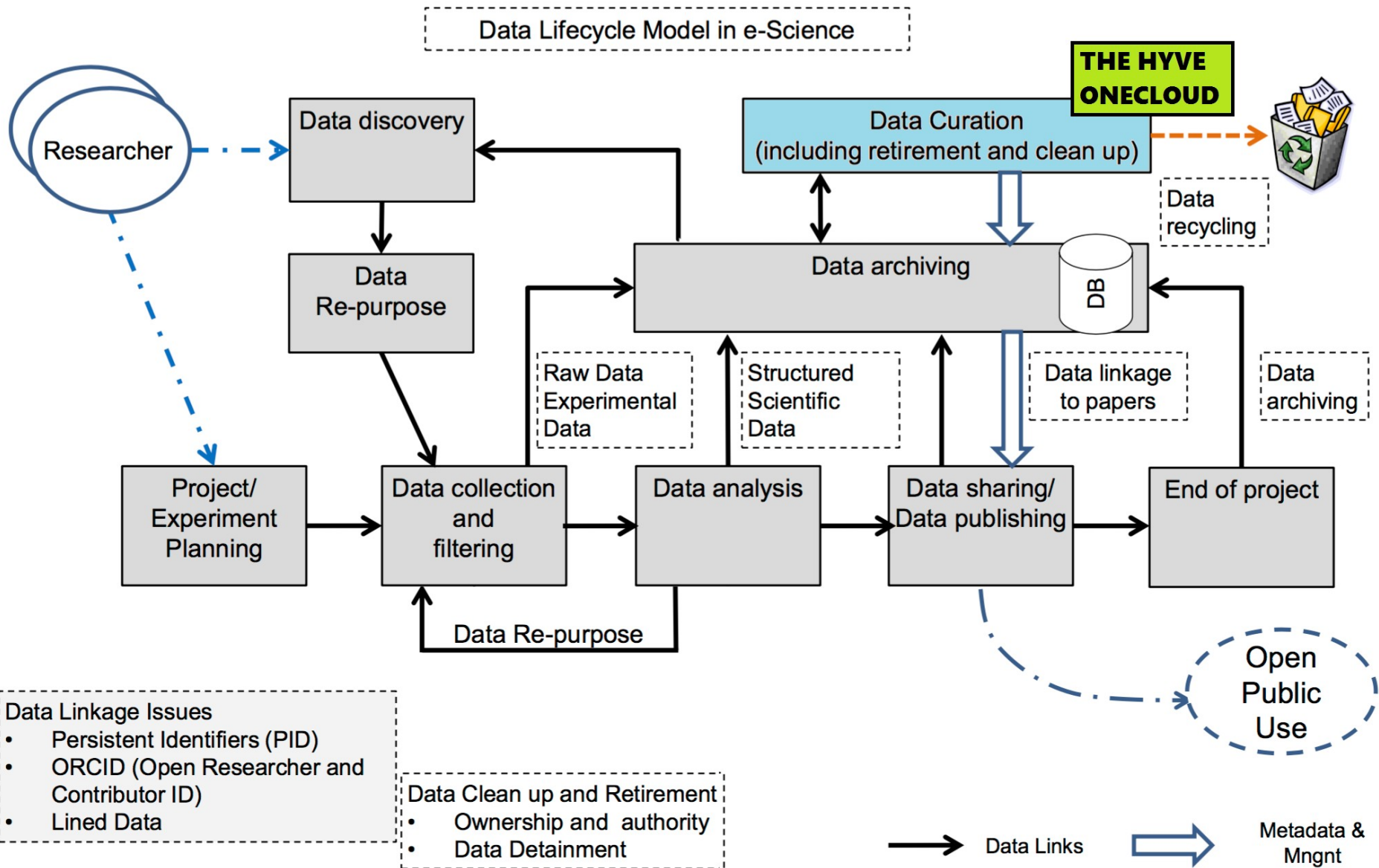
Data life-cycle model



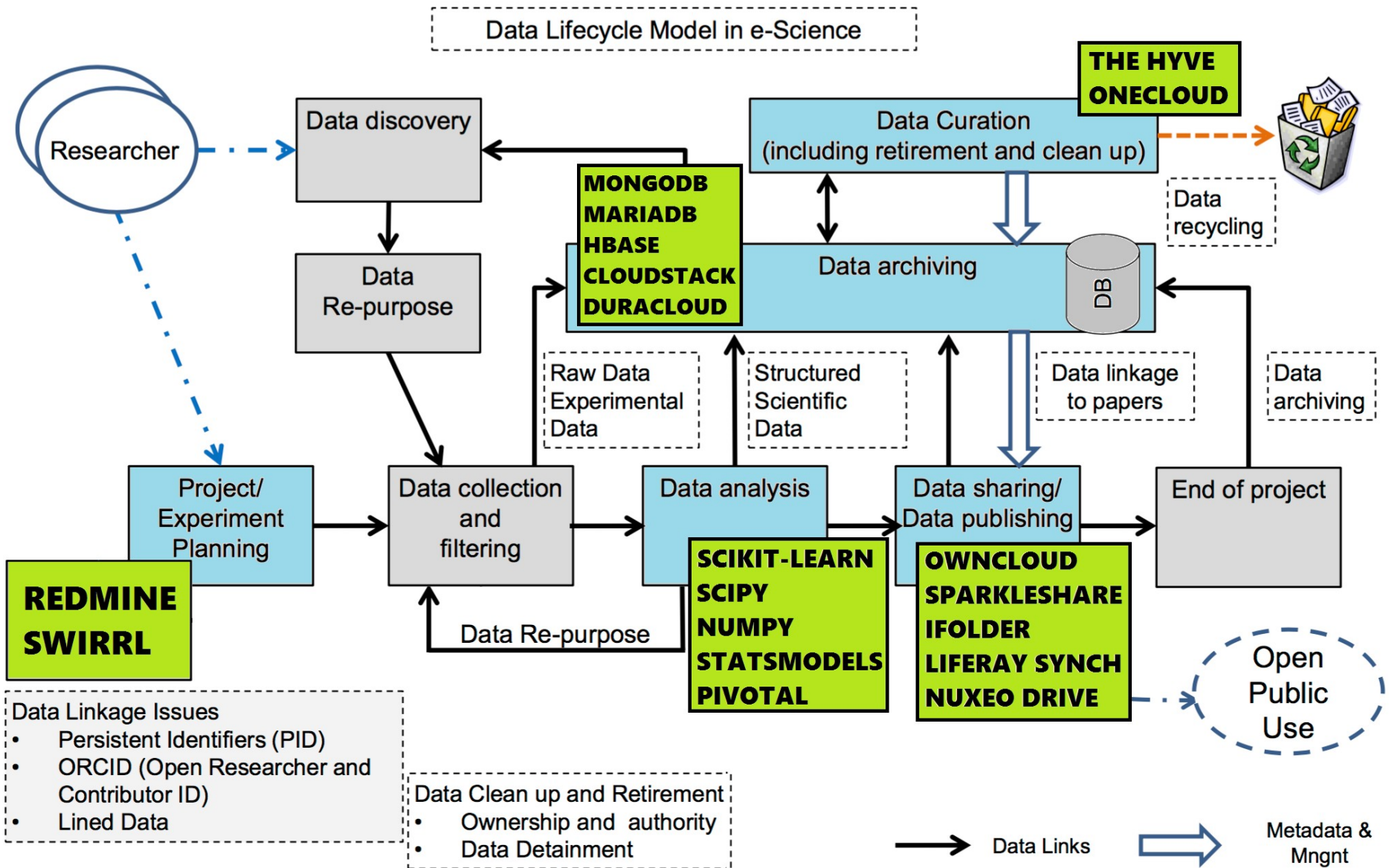
Data life-cycle model



Data life-cycle model

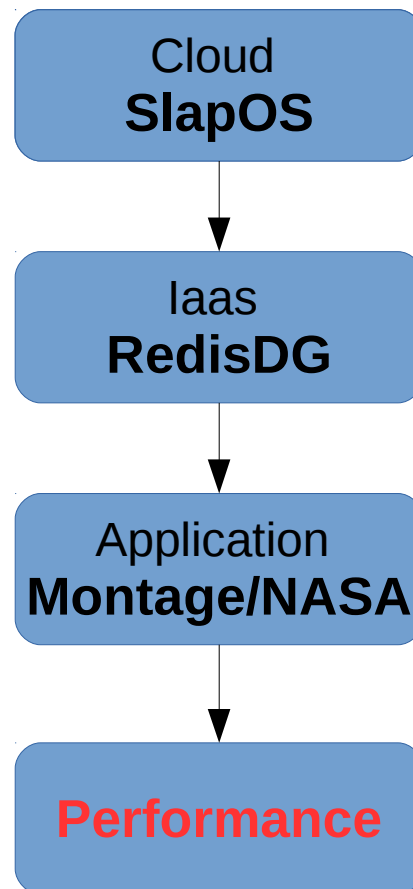


Data life-cycle model



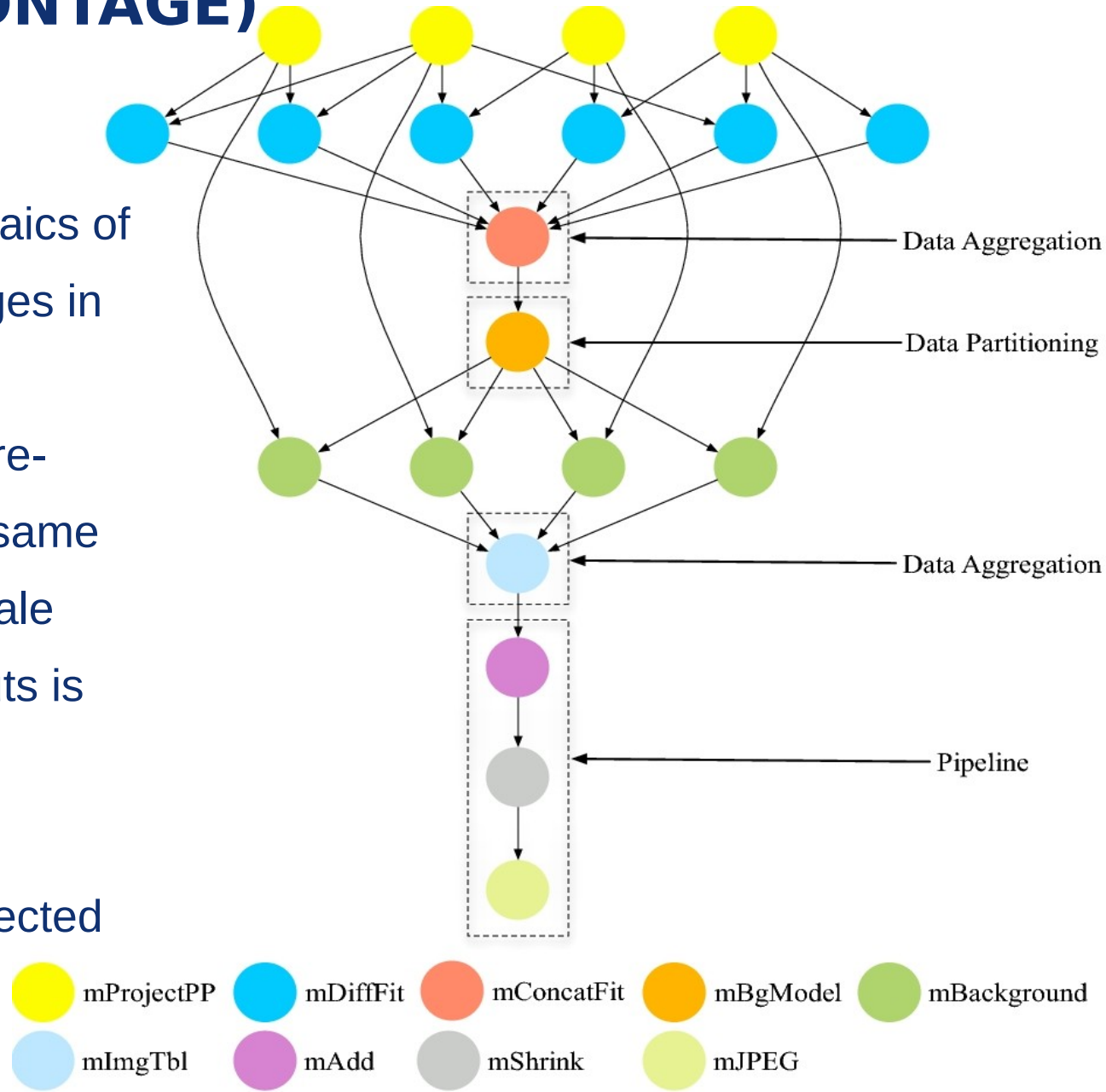
Data Analysis platform

- Experimental validation of a research work :
RedisDG as a service

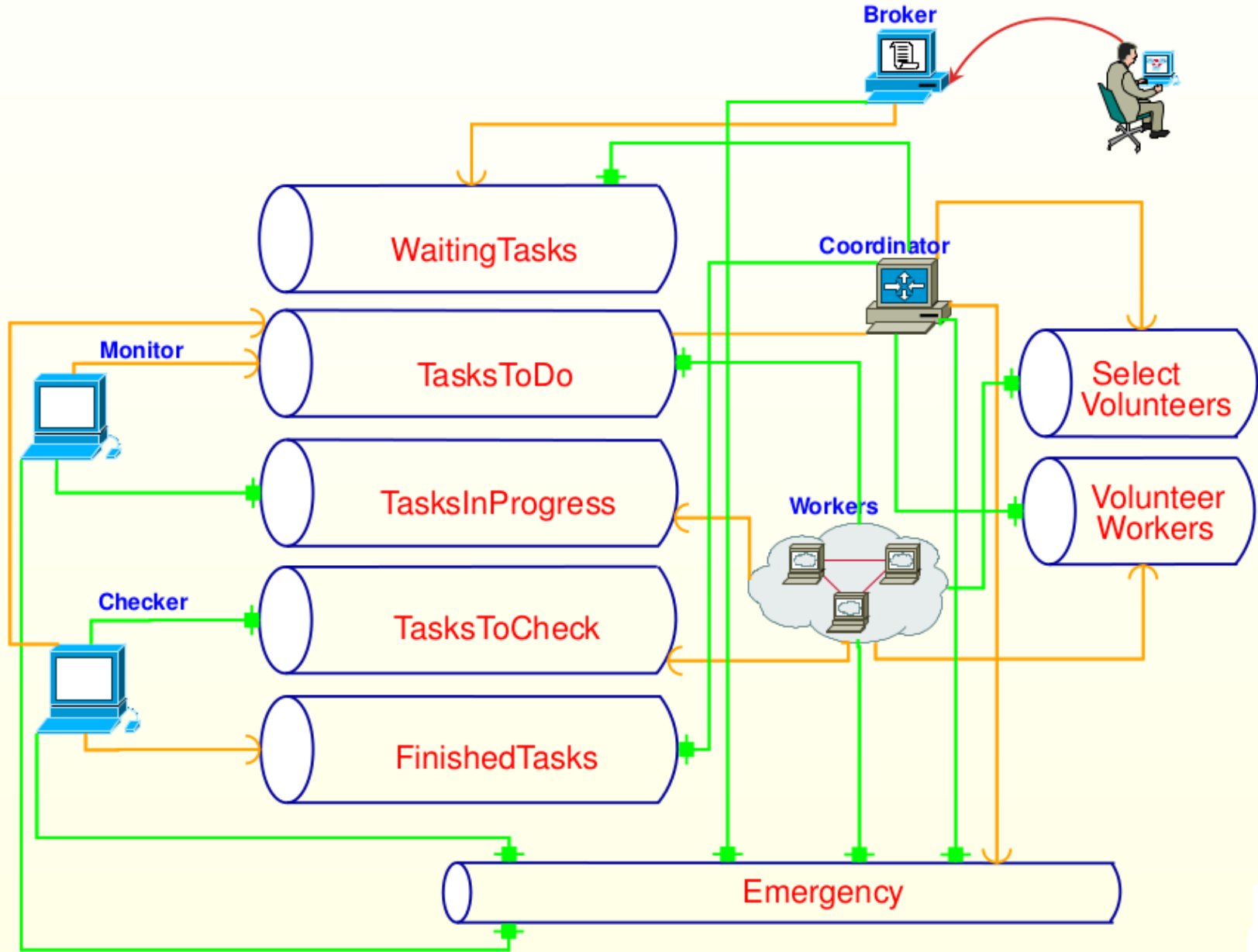


Workflow Scientific (example : NASA application-MONTAGE)

- Generate custom mosaics of the sky based on images in the format FITS
- The input images are re-projected to be in the same spatial rotation and scale
- Geometry of the outputs is calculated from the geometry of the inputs
- Re-projected and corrected images are merged



Workflow Engine : RedisDG



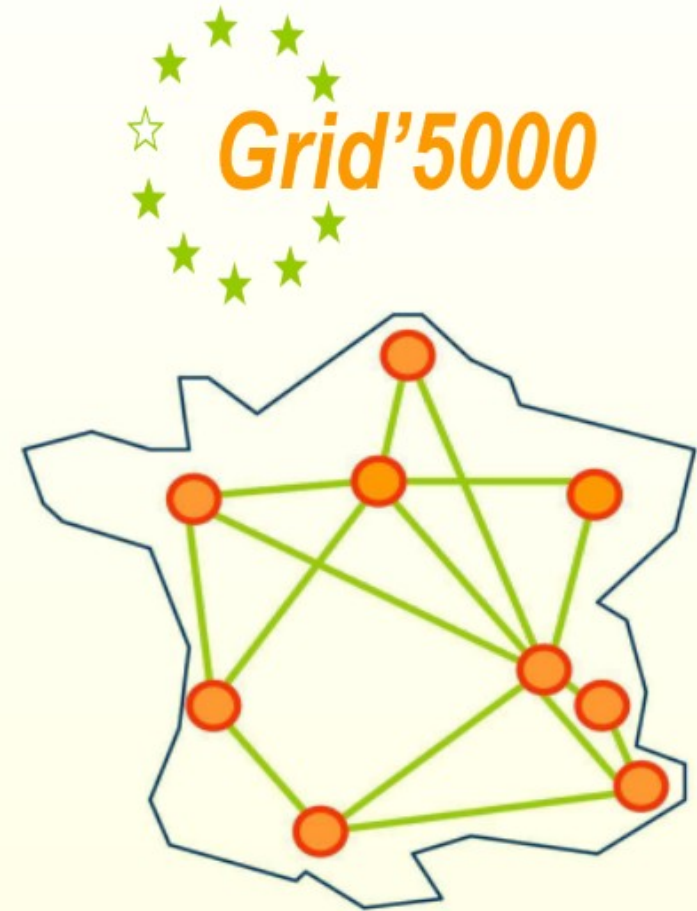
Execution of MONTAGE with RedisDG

Workflow

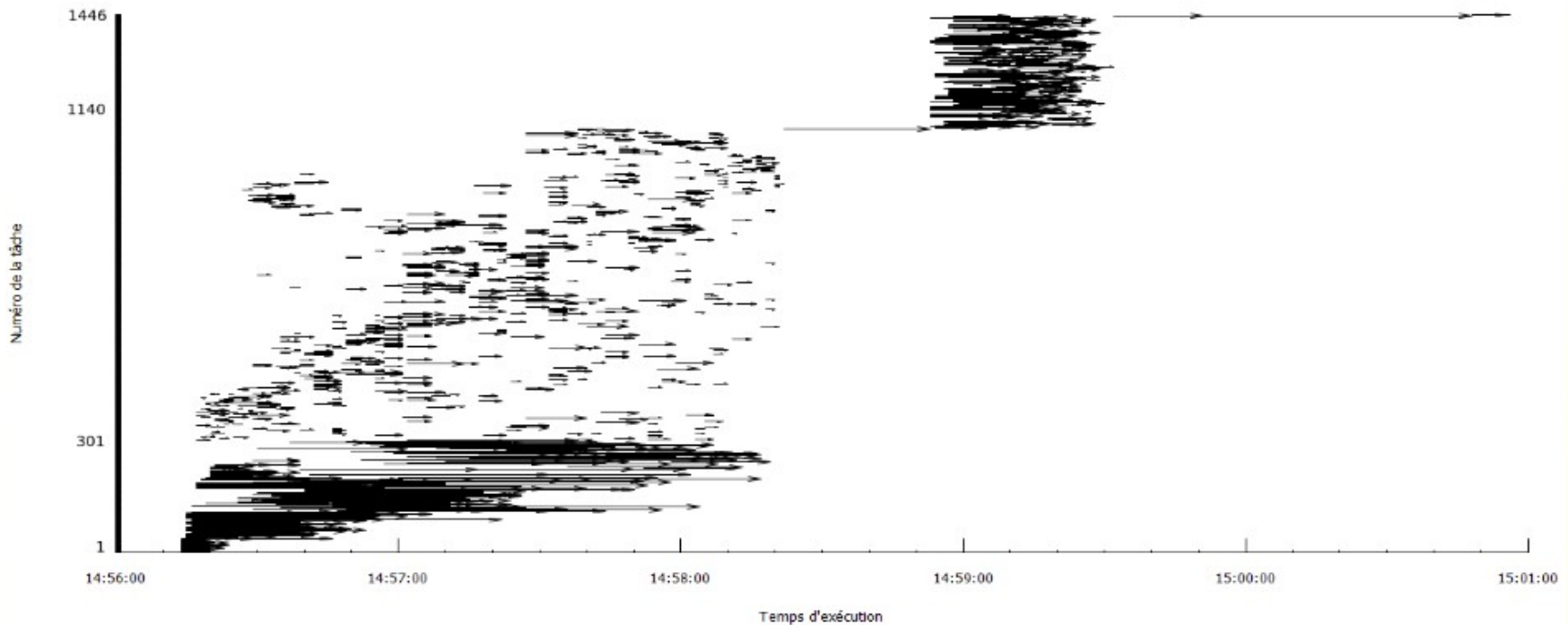
- 1446 tasks
- 3722 dependency links
- DAX with more than 20000 lines of code
- 9423 Input files (including intermediate files)
- 2889 Output files (including intermediate files)

Platform

- 3 sites : Lyon, Grenoble et Nancy
- Up to 340 nodes



Experiments



- 200 nodes (Nancy-Grenoble-Lyon)
- significant parallelism
- 200 workers participated to execution (100%)
- Round-Robin : **unfair scheduling**
- total execution time = 4 minutes

Continue performing a technical view of the data lifecycle model by :

- **integrating our workflow engine RedisDG**
- **working on the other boxes**

Thank you for your attention

Any questions?

leila.abidi@lipn.univ-paris13.fr
christophe.cerin@lipn.univ-paris13.fr
marie.lafaille@univ-paris.13.fr

