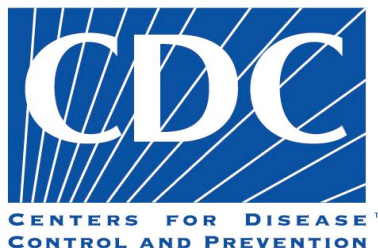




# g-INFO: Grid-based International Network for Flu Observation

Doan Trung Tung

Aurélien BERNARD, Ana Lucia DA-COSTA, Vincent BLOCH,  
Thanh-Hoa LE, Yannick LEGRE, Lydia MAIGNE, Jean  
SALZEMANN, Hong-Quang NGUYEN, Vincent BRETON



# Outline

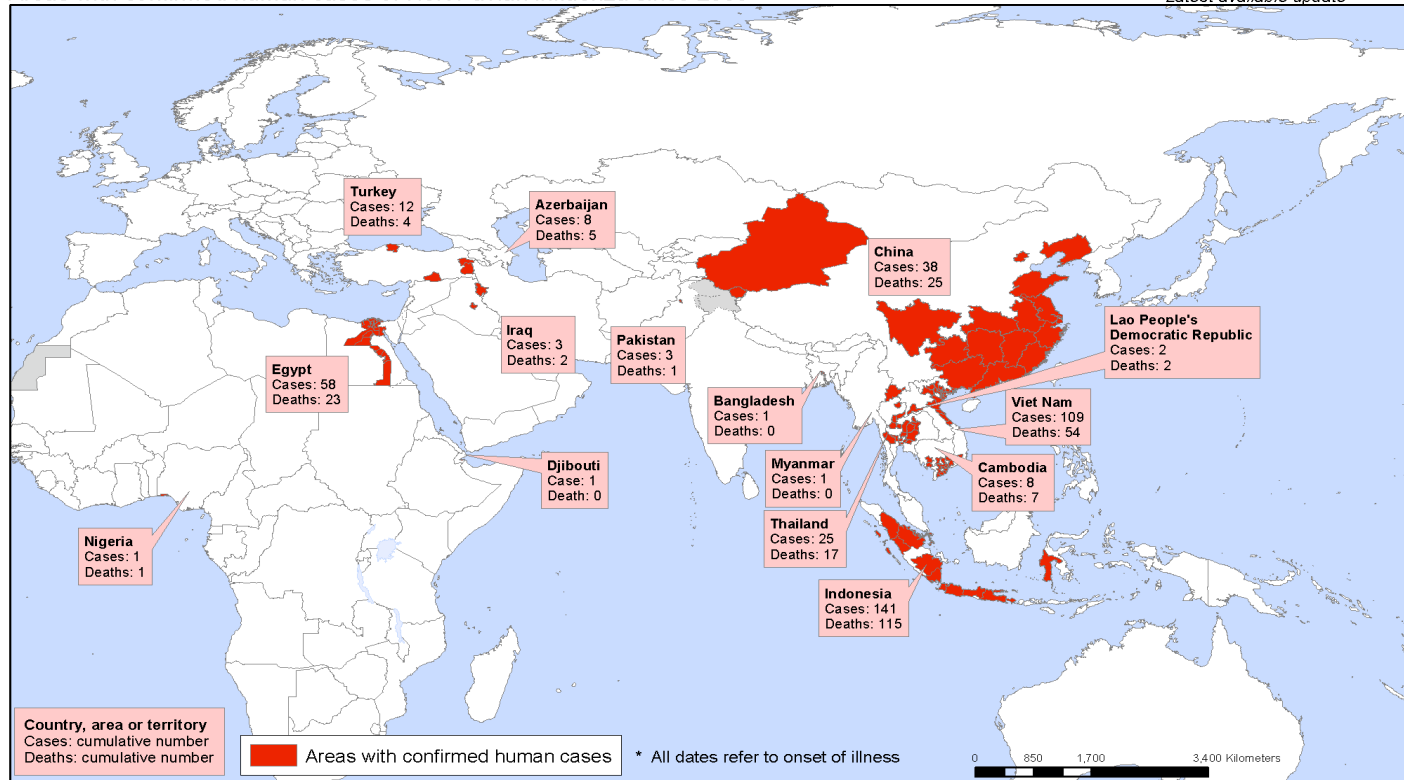
- ❖ Introduction
- ❖ Overview of g-INFO
- ❖ Implementation of g-INFO
- ❖ Conclusions and perspectives

# Current status of Influenza A

## ❖ H5N1 (avian flu)

Areas with confirmed human cases of H5N1 avian influenza since 2003 \*

Status as of 11 March 2009  
Latest available update



262 deaths  
436 cases

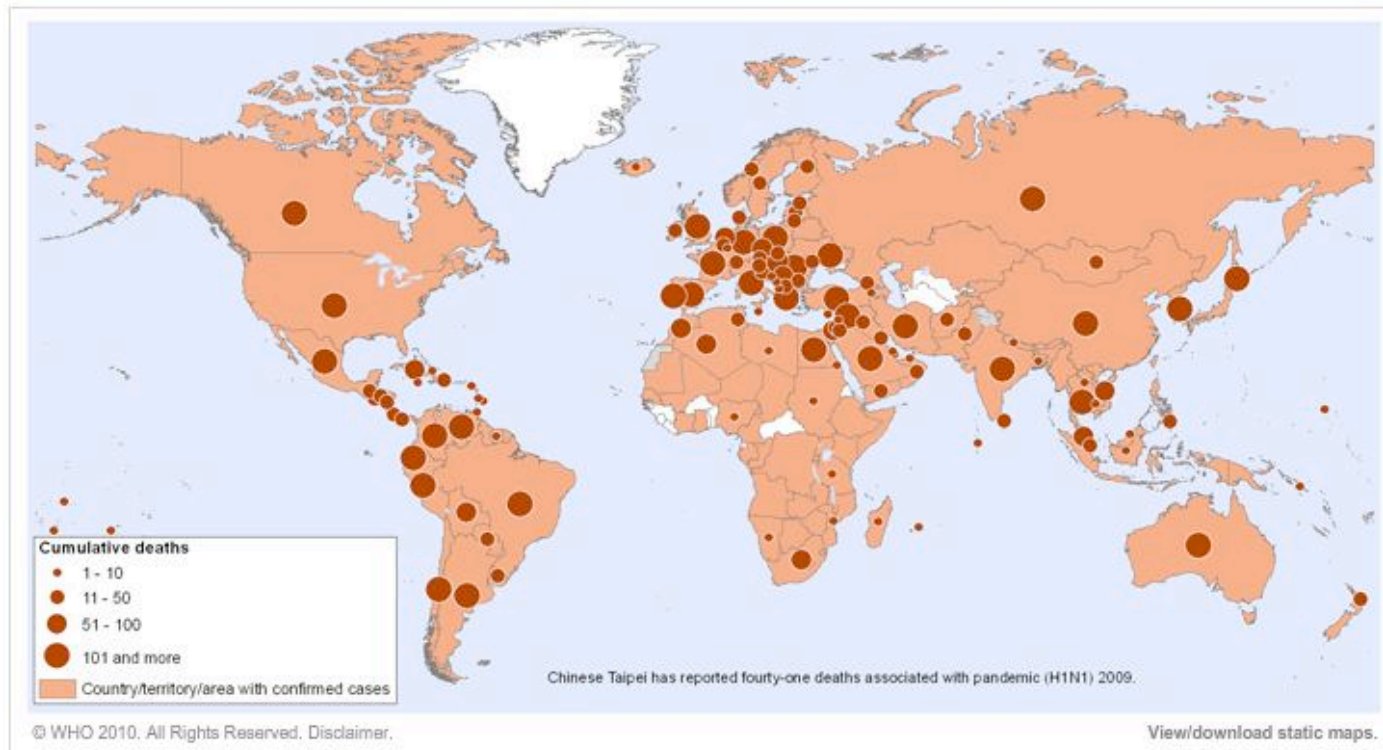
WHO - July 2009

287 deaths  
486 cases

WHO - March 2010

# Current status of Influenza A

## ❖ H1N1 (swine flu)



382 deaths  
89 921 cases

WHO - 3 July 2009

16455 deaths  
? cases

WHO – 28 Feb 2010

# Influenza surveillance

## ❖ Data collection

- BioHealthBase
- NCBI
- LosAlamos

## ❖ Data processing in batch mode

- General phylogenetic pipelines
- Specific phylogenetic pipelines

## ❖ Deployment of phylogenetic tools on clusters / grids

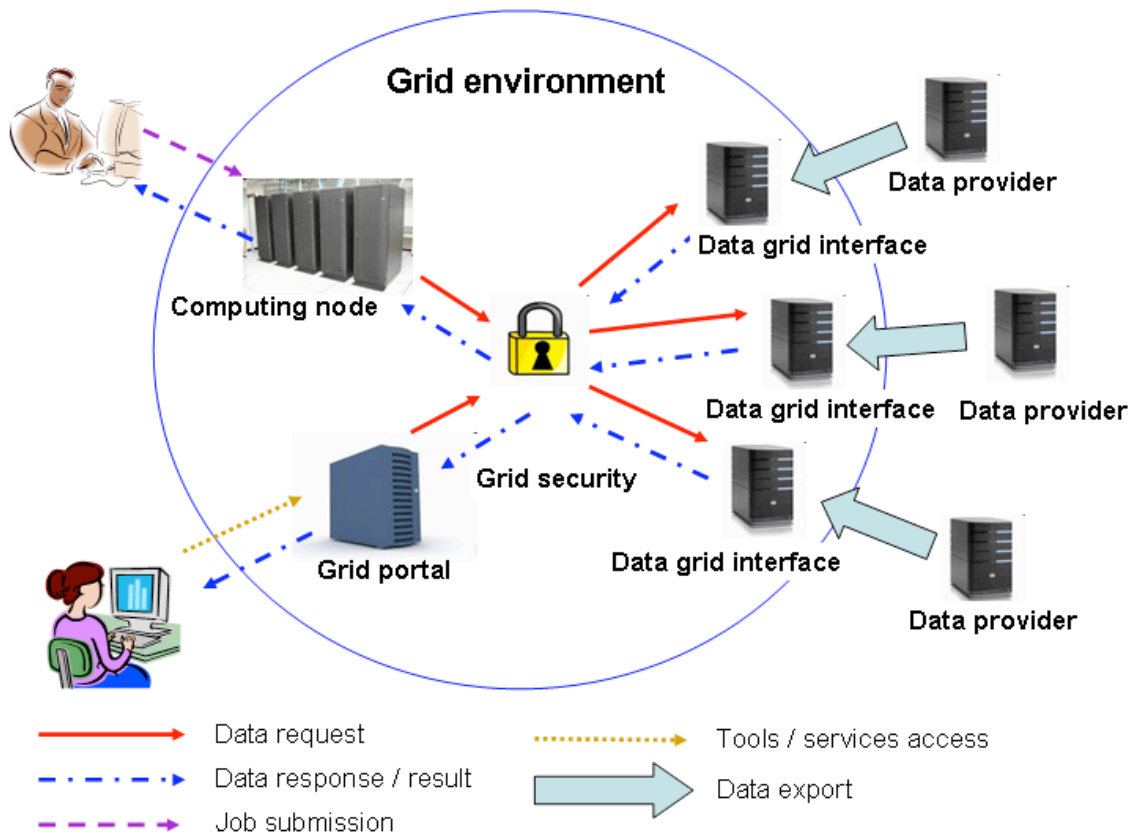
g-INFO: Grid-based  
International  
Network for Flu  
Observation

# **g-INFO's overview**

# g-INFO's goals

- ❖ Integration of influenza virus data sources into a federation of databases
- ❖ Automatic phylogenetic pipelines
- ❖ Specific molecular epidemiology studies

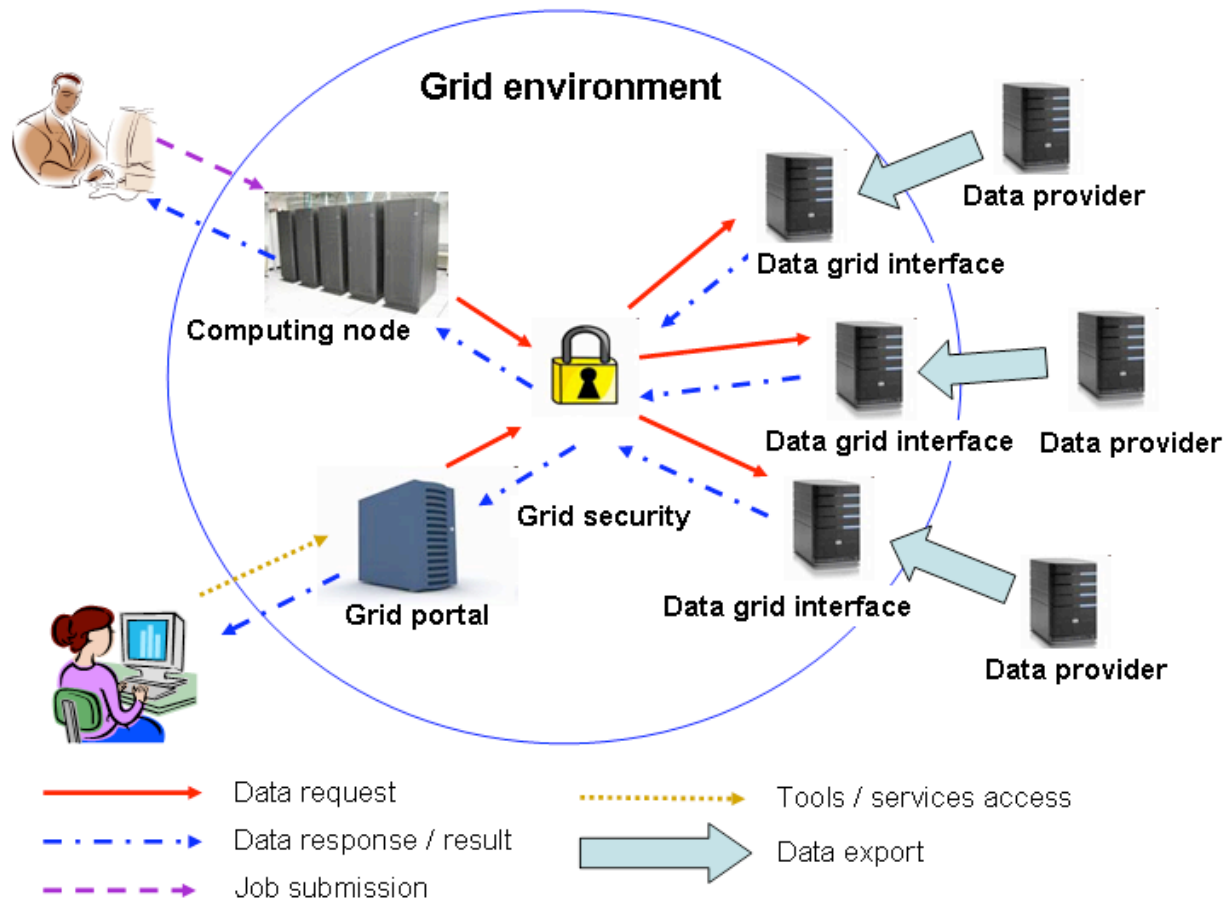
# Architecture of g-INFO



- Each data provider has its own server(s) to store his data
- Data provider export only selected data to a data grid interface server
- The data exported is integrated in a common schema on the interface servers
- Providers can keep the privilege of granting access rights to their data



# Architecture of g-INFO



▪ Epidemiologic pipelines will be deployed on the grid

- ❖ BLAST
- ❖ Alignment
- ❖ Phylogenetic trees
- ❖ Visualisation
- ❖ ... and more

# **g-INFO's implementation**

# Data collection

>ABV25634








MKAILLVLLCAFAATNADTLICIGYHANNSTDVTDVLEKNVTVTHSVNLLED SHNGKLCRLGGIAPLQLG  
 KCNIAGWLLGNPECDLLLTVSSWSYIVETSNSDNGTCYPGDFIDYEELREQLSSVSSFEEKFEIFPKTSSW  
 PNHETTRGVTAACPYAGASSFYRNLLWLVKKENSYPKLSKSYVNNKGKEVLVLWGVHHPPTSTDQQLYQ  
 NADAYVSVGSSKYDRRFTPEIAARPKVRGQAGRMNYYWTLLPEGDTITFEATGNLVAPRYAFALNRGSES  
 GIITSDAPVHDCDTHCQTPHGAINSSLPFQNIHPVTIGECPKYVKSTKLRMVTGLRNIPSIQSRGLFGAI  
 AGFIEGGWTGLIDGWYGYHHQNGQSGYAADQKSTQNAIDGITNKVNSVIEKMNTQFTVVGKEFNLER  
 IKNLNKKVDDGFLDVWTYNAELLVLENERLTDFHDSNVKNLYEKARSQLRNNAKEIGNGCFEFYHKCDD  
 ACMESVRNGTYDYPKYSEESKLNREEIDGVKLESMVYQILAIYSTVASSLVLLVSLGAISFWMCSNGSL  
 QCRICI

## FTP NCBI

Index de <ftp://ftp.ncbi.nih.gov/genomes/INFLUENZA/updates/2009-07-16/>

 Vers un rép. de plus haut niveau

### Nom

-  [genomeset.dat](#)
-  [influenza.cds](#)
-  [influenza.dat](#)
-  [influenza.faa](#)
-  [influenza.fna](#)
-  [influenza\\_aa.dat](#)
-  [influenza\\_na.dat](#)



### Taille Dernière modification

Taille	Dernière modification	
	16/07/09	08:04:00
105 KB	16/07/09	08:04:00
5 KB	16/07/09	08:04:00
42 KB	16/07/09	08:04:00
97 KB	16/07/09	08:04:00
13 KB	16/07/09	08:04:00
9 KB	16/07/09	08:04:00



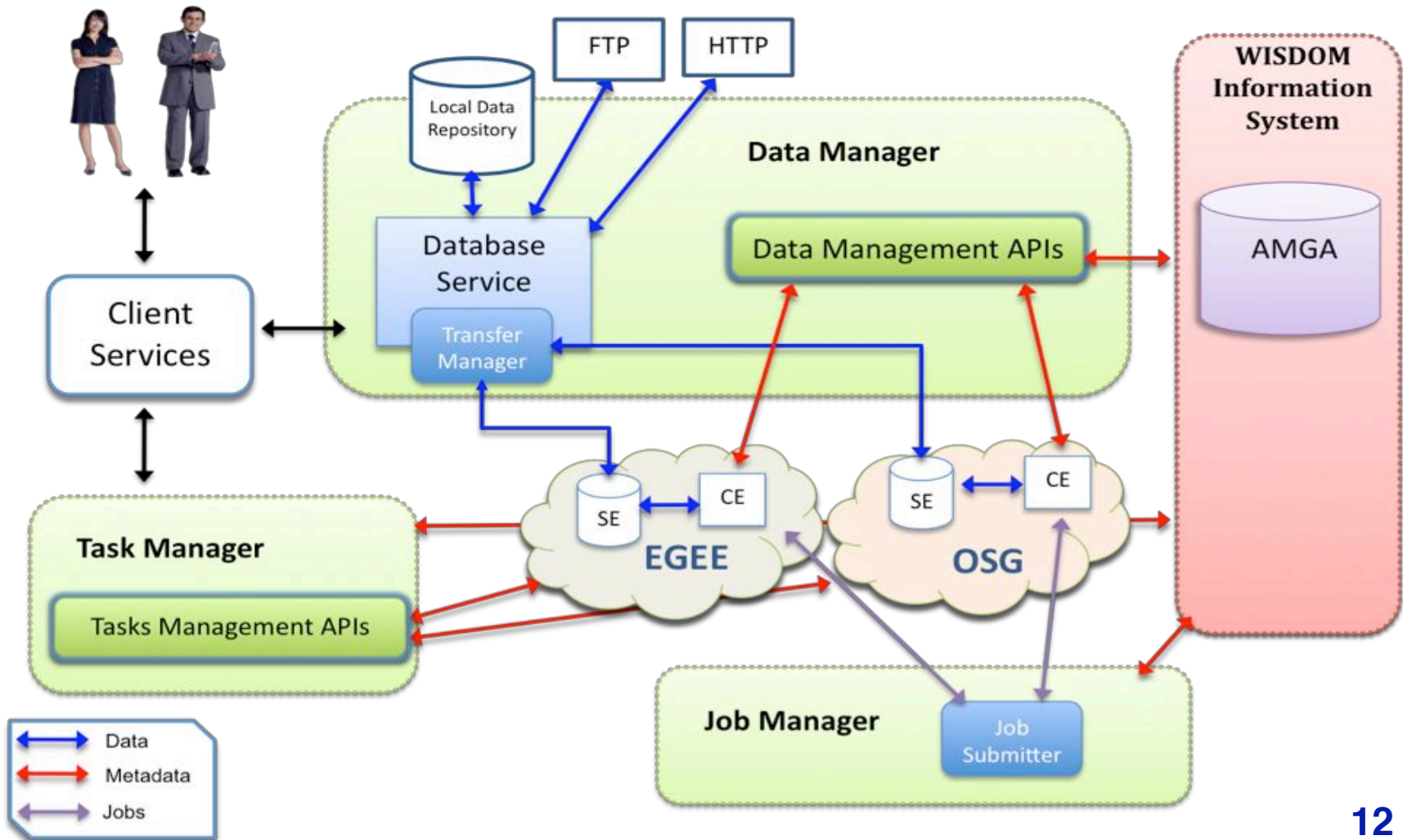
## Grid DB

Metadata

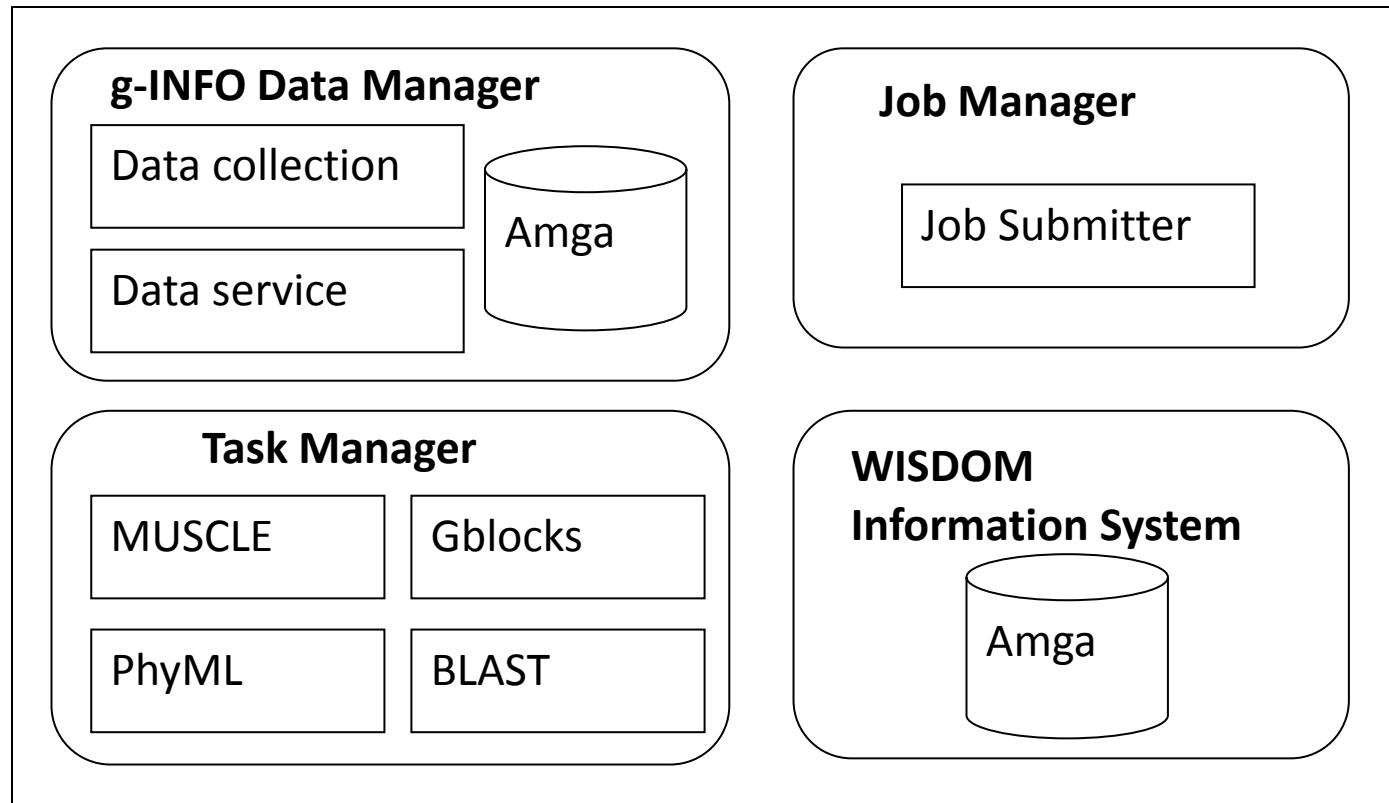
Sequences  
*Protein, Nucleotide,  
 Coding region*

IDs

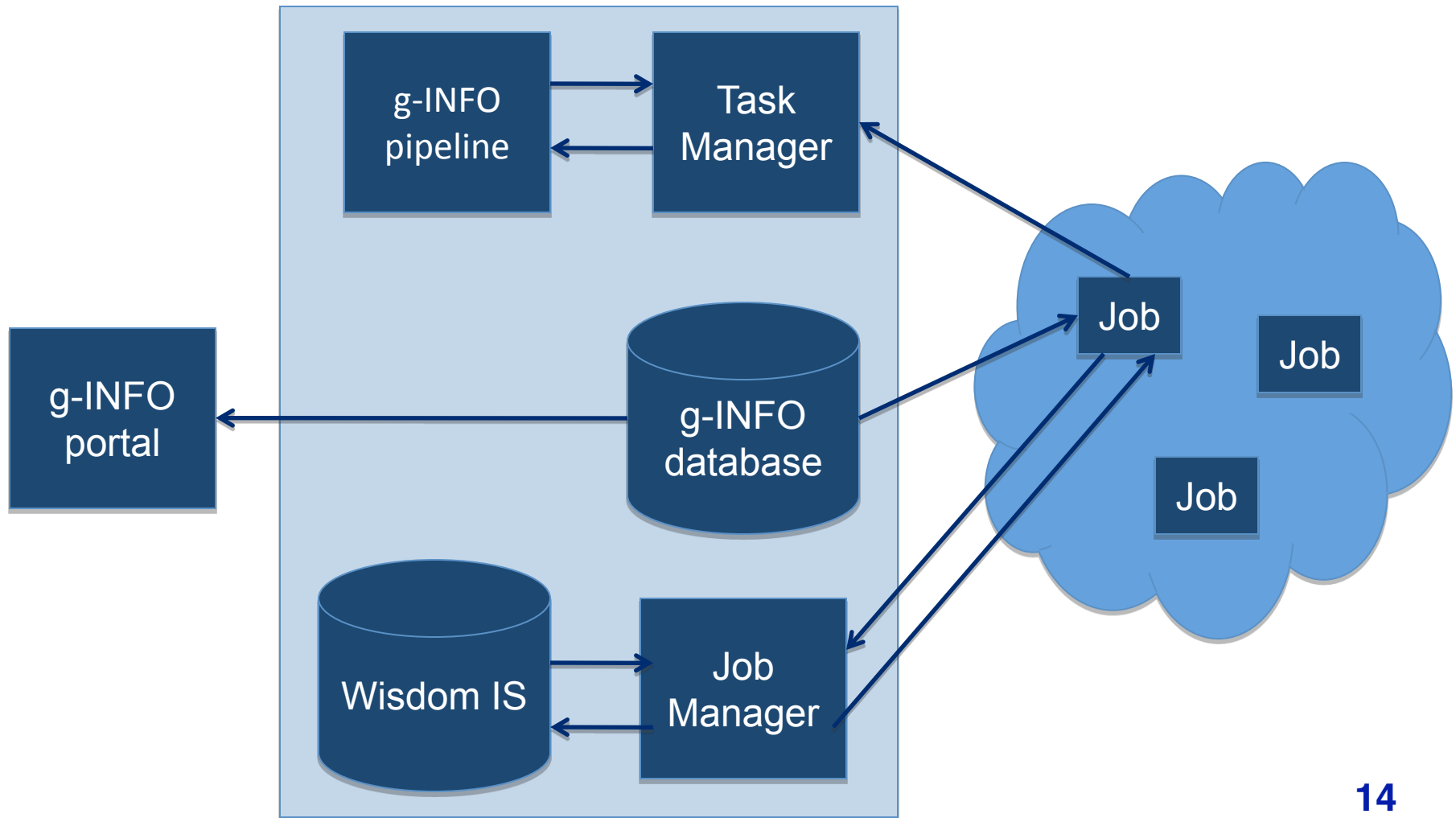
# WISDOM Production Environment



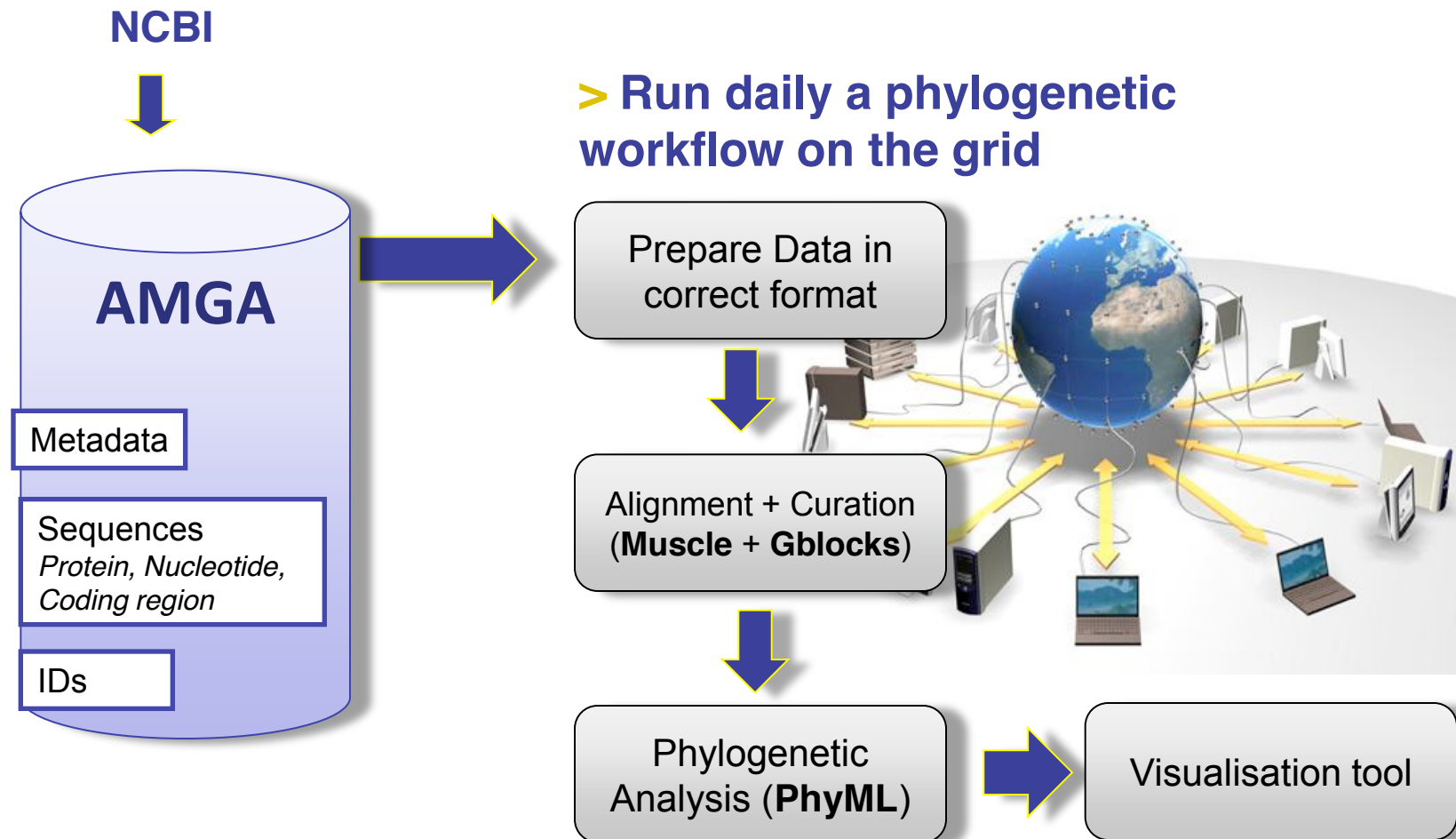
# Integration of g-INFO into WPE



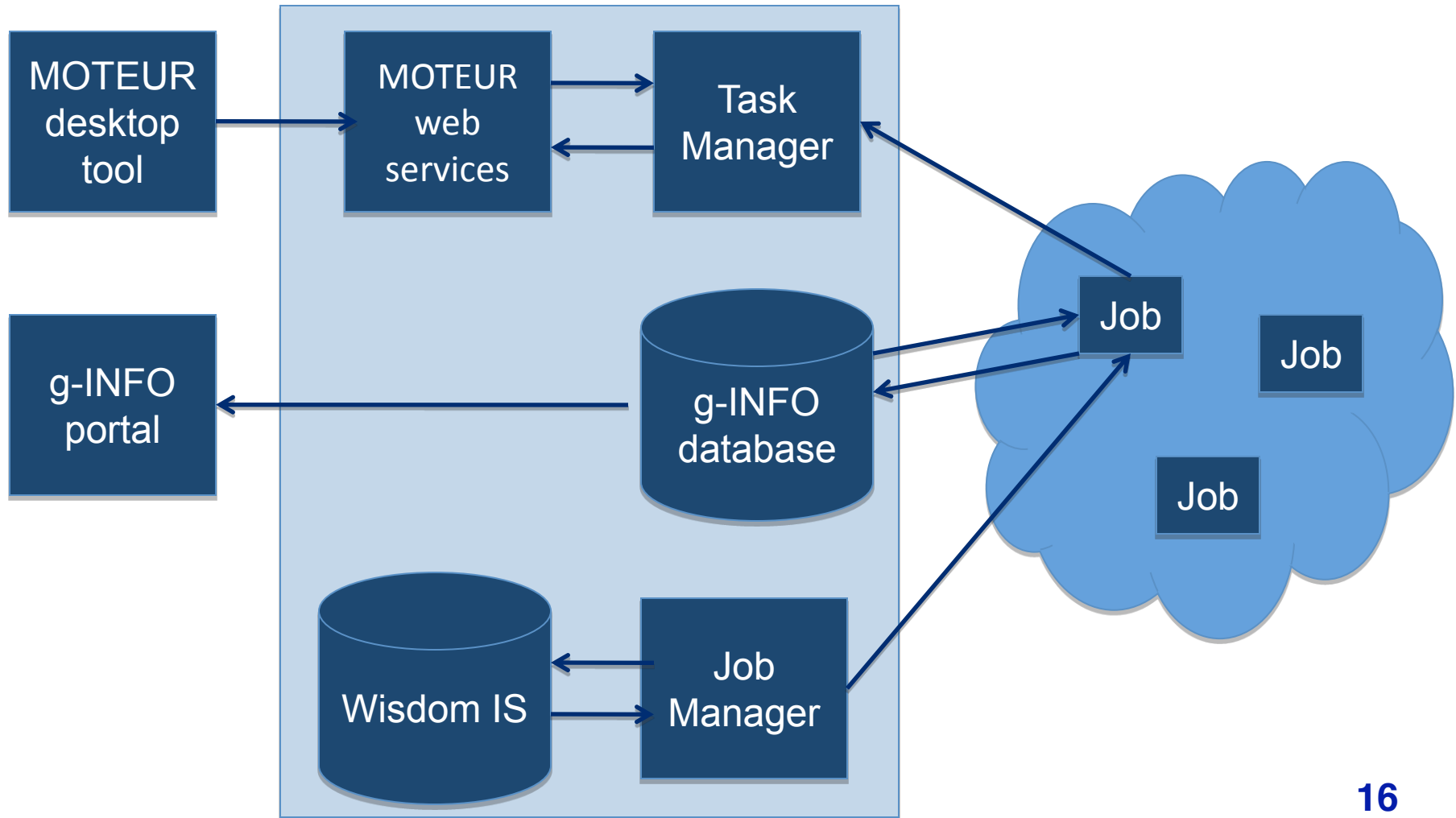
# Automatic phylogenetic pipeline



# Automatic phylogenetic pipeline

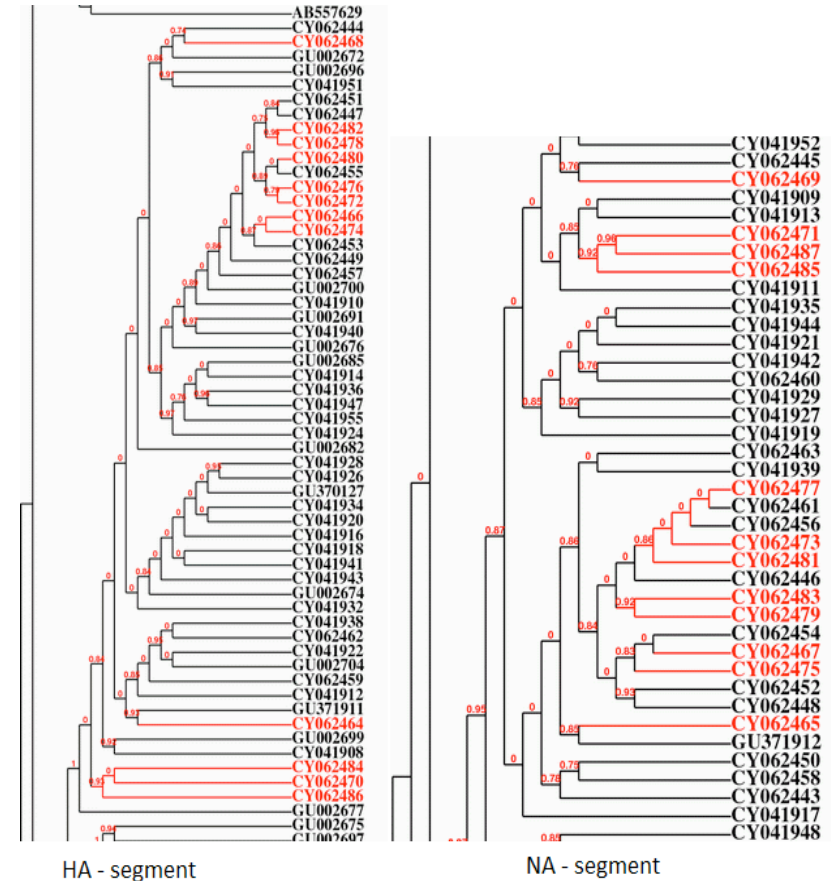
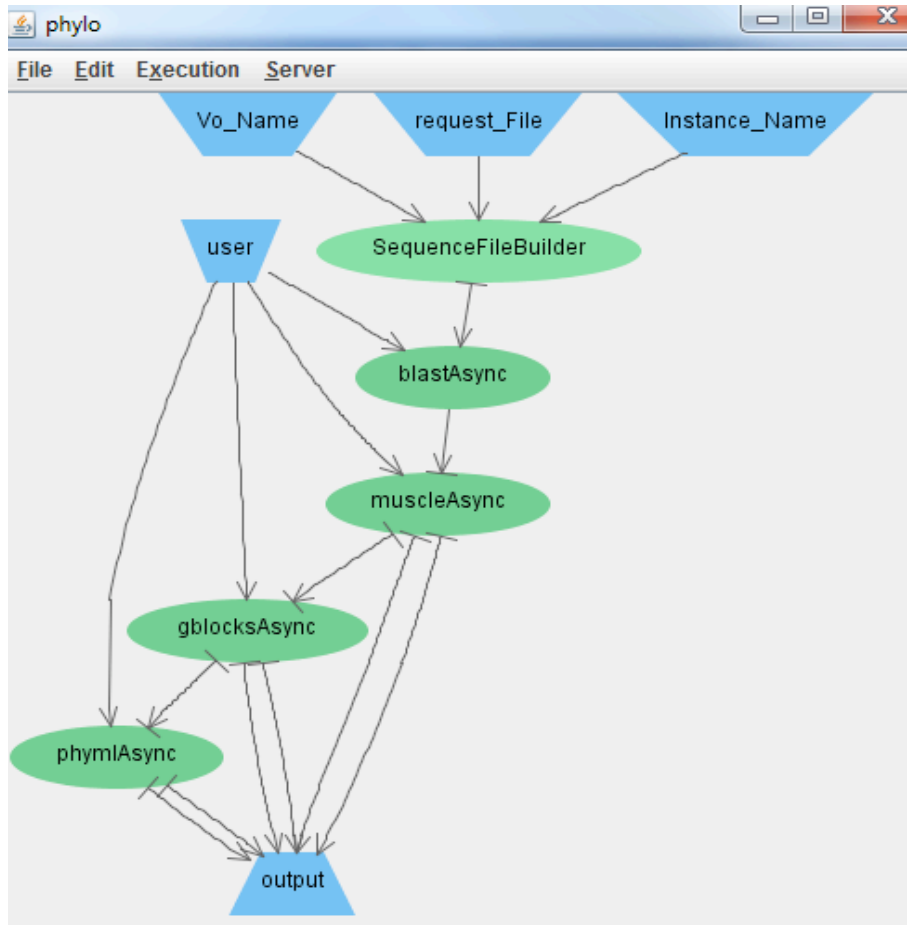


# Manual phylogenetic workflow

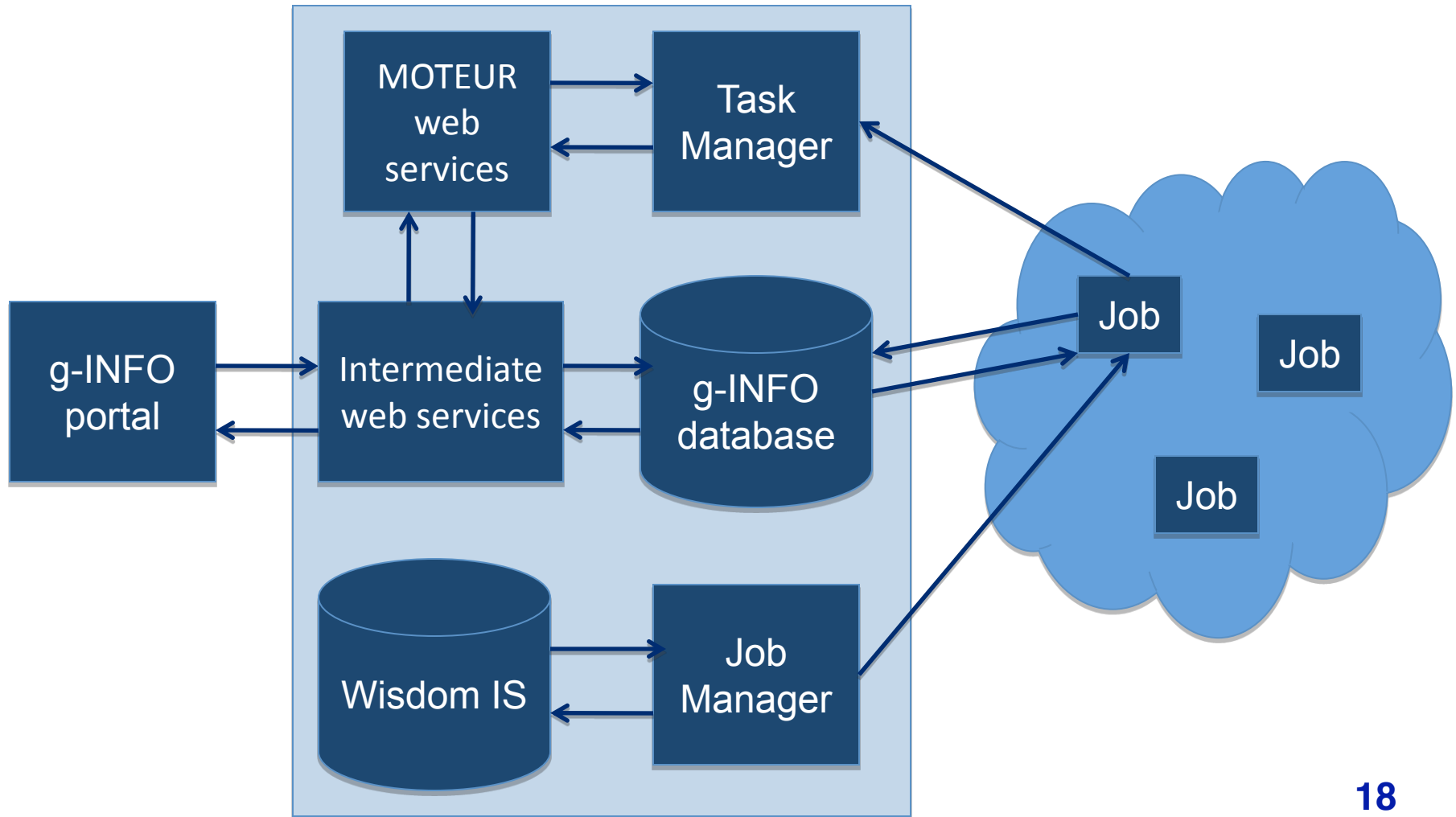





# Workflow execution example



# g-INFO portal (under development)



# g-INFO website



grid-based  
International  
Network for  
Flu  
Observation

[Home](#)  
[Flu Virus](#)  
[Databases](#)  
[Analysis](#)  
[Results](#)  
[Contact](#)

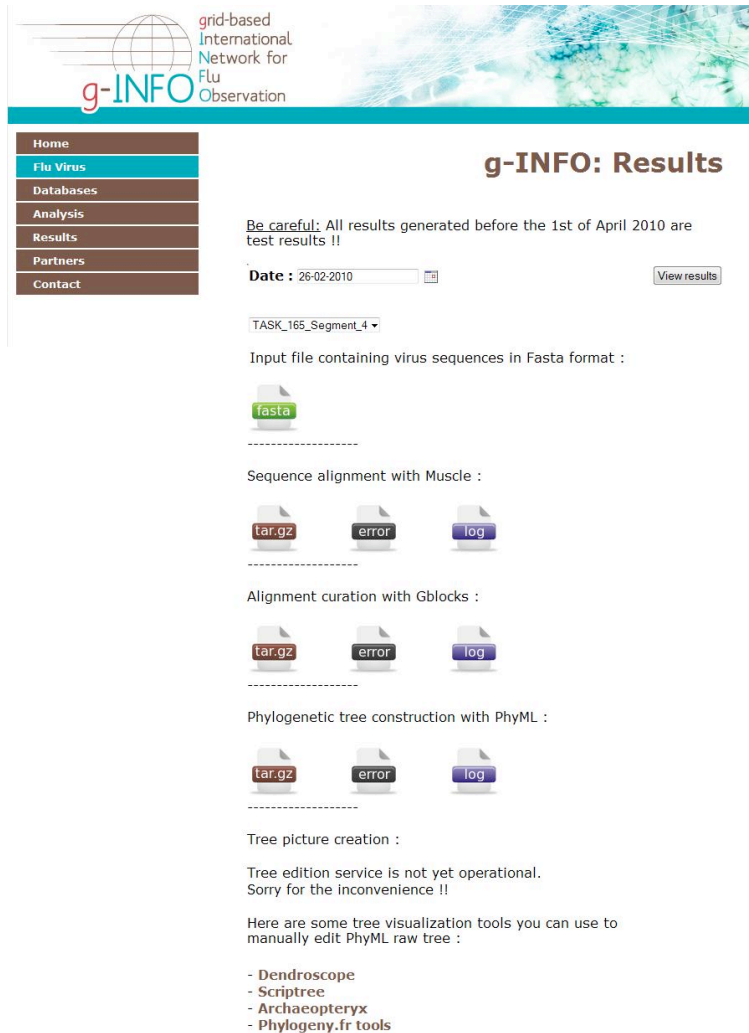
## g-INFO: Home

Recent years have seen the emergence of diseases which have spread very quickly around the world, either through human travel, like SARS and SIV(H1N1), or animal migration, like avian flu (H5N1) or more recently, the swine flu outbreak that has been classified as a "pandemic" by WHO in response to its world-wide geographic Spread.

Among the biggest challenges from emerging infectious diseases, is the relation to early detection and surveillance of the diseases, as new cases can appear anywhere. This is due to the globalization of exchanges and the circulation of people and animals around the world, as recently demonstrated by the avian flu epidemics. An international collaboration of research teams in Europe and Asia has been exploring some innovative in silico approaches to better tackle flu, taking advantage of the very large computing resources available on international Grid infrastructures. Based on current H1N1 pandemic example, it is expected to have an impact by adding a new weapon to researchers' arsenal: the grid.

Existing data sources have been integrated towards a global surveillance network for molecular epidemiology, based on Service Oriented Architecture (SOA) and Grid technologies. The idea is to dynamically analyze the molecular biology data, made available on public databases using computing, storage and automatic updating services offered by grid technology.

# g-INFO website



The image shows a screenshot of the g-INFO website interface. The header features the g-INFO logo and the text "grid-based International Network for Flu Observation". A navigation menu on the left includes links for Home, Flu Virus, Databases, Analysis, Results, Partners, and Contact. The main content area is titled "g-INFO: Results" and contains a warning: "Be careful: All results generated before the 1st of April 2010 are test results !!". Below this, there is a "Date" field set to "26-02-2010" and a "View results" button. A dropdown menu shows "TASK\_165\_Segment\_4". The section "Input file containing virus sequences in Fasta format:" includes a "fasta" file icon. The "Sequence alignment with Muscle:" section has "tar.gz", "error", and "log" file icons. The "Alignment curation with Gblocks:" section also has "tar.gz", "error", and "log" file icons. The "Phylogenetic tree construction with PhyML:" section has "tar.gz", "error", and "log" file icons. Below these, there is a "Tree picture creation:" section, a note that "Tree edition service is not yet operational. Sorry for the inconvenience !!", and a list of tools for manually editing PhyML raw trees: Dendroscope, Scriptree, Archaeopteryx, and Phylogeny.fr tools.

grid-based  
International  
Network for  
Flu  
Observation

**g-INFO: Results**

Be careful: All results generated before the 1st of April 2010 are test results !!

Date : 26-02-2010

TASK\_165\_Segment\_4

Input file containing virus sequences in Fasta format :

Sequence alignment with Muscle :

Alignment curation with Gblocks :

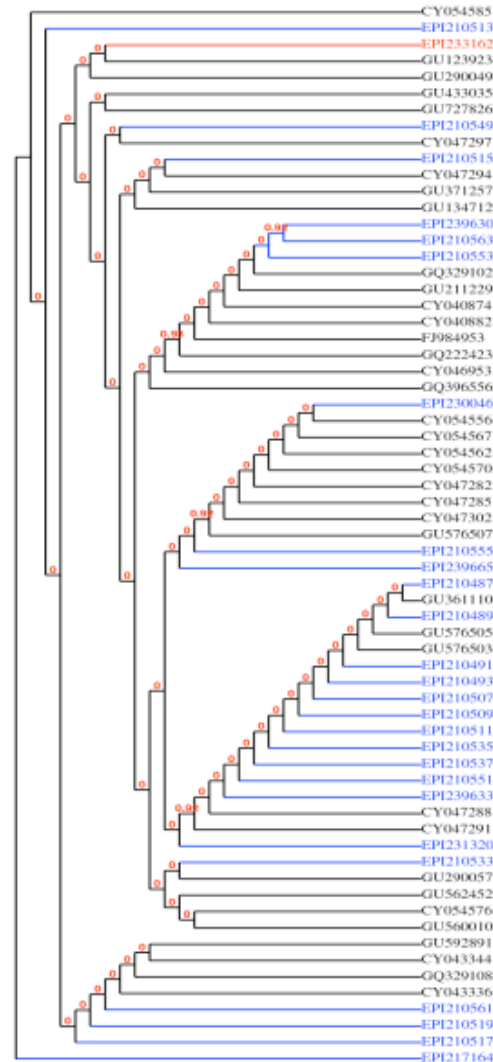
Phylogenetic tree construction with PhyML :

Tree picture creation :

Tree edition service is not yet operational.  
Sorry for the inconvenience !!

Here are some tree visualization tools you can use to manually edit PhyML raw tree :

- Dendroscope
- Scriptree
- Archaeopteryx
- Phylogeny.fr tools



# Conclusions

- ❖ A success in terms of international collaboration
- ❖ A complementary service for the public health research community
- ❖ The adoption of grids for pandemics monitoring represents a step forward responding to the needs of the research community concerning the federation of all the influenza data sources

# Perspectives

- ❖ Provide more tools and pipelines
- ❖ Import other database resources
- ❖ g-INFO portal's users can create their own pipelines and store specific analysis results
- ❖ We are expecting the research community to contribute with more useful tools
- ❖ Can be applied for other emerging diseases

Thank you!