Transversal Bio Image Informatics – IPDM Node

DIGEST FROM VISITS OF FBI NODES

Paul-Gilloteaux Perrine | FBI bio image informatics- visit digest | 14 janvier 2016

Table of contents

Purpose and context of this document

In the context of reinforcing the operating phase of France Bio Imaging, visits to FBI nodes by the administrative project manager and the project manager for the IPDM transversal node has been organized on site. Regarding image processing and data management, the purpose of the visit and to discuss the needs from the facility staff and users in this regards, in order to gather elements for a roadmap for IPDM , and in particular the blocking points for setting up data management plans and tools on the different sites. It was to get the occasion of advertising the activities of the node and to get needs for transversal activities for this node.

Calendar of visits

29-30 of September 2015: Montpellier node (IGH, CBS Mars)

13-11 of November 2015: Paris Sud (Imagerie Gif, Bioemergences, LOB ecole Polytechnique)

8-9 of December 2015: Marseille node (CIML, IBDML, Fresnel)

14 of December 2015: Paris Node (Paris Descartes, IBENS, Imachem)

16-17 of December 2015 : Bordeaux node (BIC, LP2N, IINS)

Needs from facilities

IT INFRASTRUCTURES

The IT infrastructure is unequal between the nodes for now. For example in Bordeaux, the network is only 1G max, but should be upgraded to 10G while moving to their new building. In Marseille, most of the network lines are at 1Go but with a network centre at 10Go, making it theoretically possible to extend new 10Go lines inside the buildings. Projects to do so are already discussed, in particular for microscopes producing big data (e.g. SPIM). Elsewhere, **most of the connections are in 10Go.**

When a node is on several site, in particular attached to different research institutes, **the different sites has no network communications**

- In Bordeaux Inra (LP2N) and IINS are fully disconnected while having a common facility (BIC).
- In Marseille, CIML and IBDML are also disconnected, while a common campus IT service is existing, but dealing with the network only until it reaches the buildings. An FTP service has been set up in order to deal with data exchange between these networks and with external users/collaborators.
- In Gif Sur Yvette (Paris Sud), several networks are coexisting, with some bridges.
- The different nodes in Paris are on full different network.

In all places (Bordeaux) a centralized server with various size is set up, but not in regular used, mainly due to the difficulty of changing the habits, the lack of information.

DATA MANAGEMENT, VALORIZATION AND STORAGE

New systems are being acquired by the facilities (example: **Serial Block Face in EM in Marseille, Block Face imaging in biphoton, SPIM in different nodes**, …), and three main problems are expected, for which no solutions are proposed for now:

1. A **big jump of data production** is expected, but **difficult to quantify before feedback** on the actual usage
2. Data transfer to storage place requires **a dedicated infrastructure** in most of the cases
3. **Software in use are inadequate for processing and visualization of large data**, in particular volume data. It may be due to both hardware limitation or software limitation (due to internal management of memory and of image data access)

In addition, some facilities are also in the process of a quality approach, in which a **data management plan** may be needed. Some facilities has set up some rules (such as automatic archiving on another server after 3 years), some other are just calling to cleaning when the local facility servers are full.

In all sites, **external hard drives stay the favourite way of transferring data for users**, even when a centralized system is set up. Centralized storage is underused in most of the places. When there is a risk of it appears that duplication of data maybe one of the main reasons.

Several facilities (starting from one site of a node most of the time) have set up a data management solution where the data are stored in a centralized server.

- In Marseille **Omero** is in place, but only one team is actively using it. (40To, 50% used for CIML, 220To (90% used) for IBDML.
- In Bordeaux, the centralized server is not used due to network speed.
- In Montpellier, the home made solution WIDE is now replaced by **OMERO**, with an FBI engineer to develop missing tools (such as interaction with image processing tools). 85 To (24% used). All data will have to go to Omero , data on the previous ftp server will be kept only 6 months.
- In Paris-Sud , Bioemergences has about 70To (unknown % in used) , but in Paris Sud, storage is done on local computers. A file transfer system IRODS is in place. Their own data base system (**bioemenergences**) is in place
- In Paris Ibens, **Omero** has been set up, with 2Pbytes (unknown % in use)
- In Paris Curie, Cid iManage has been set up, with 2 Pbytes (unknown % in use)

The **structuration of data in terms of common semantic** seems a **practical issue** preventing a broader use of data management systems. This aspect appears to be less complicated in HCS where experiments and measured phenotypes are more structured.

However, **some facilities consider that providing a secured access to their data to users is not part of their mission**, while others considers that it is. For example, most of the facilities do not invoice (directly) the storage.

In addition, needs for big data valorisation (meaning both its diffusion with correct curation and its exploitation, and its convenient visualization) is underlined. **Tools in use in facilities are not adapted to big data human assimilation.**

Another need expressed by the facilities are to reduce the burden of development by having **interoperable software tools** (for example through a common interface Knime), including the compatibility of results such as ROI.

The **3D visualization and 3D representation** is still a problem (with richer view than isosurface or simple volume rendering)

ACCESS TO COMPUTE NODES

For most of the facility nodes, this appear as **not been a priority**, the access to tools been able to **process large data is not comprehended in this way**. In addition most of the nodes have access to a compute grid (Marseille have access to CCPM grid, Bioemergences to IN2P3 grids and European grids in addition to their own cluster, IBENS and Curie have their own cluster) but only for developers due to the **lack of user friendly interface** to run preinstalled software, apart in Curie where access is also provided through the image data base to preinstalled software, or to run workflow in Icy.

The need to access cluster grids could be raised by the more extensive use of image data base and its direct link with image processing tools, or to process big data if the access to this grids and communication with the data sets is eased.

Needs from associated teams

Some teams (Fresnel Marseille) have sufficient access to local clusters, but most of them would be greatly interested in accessing grids (with different scenarii, ranging from developer access to running a software with simple user interface)

Most of the research teams do not need to have access to storage and have their own server (ranging from 40 To to 80To for some teams) , but will consider to share their data in the purpose of facilitating development of image processing tools or validation/cross comparisons of data.

In particular teams at the end of transfer technological are now considering the aspect of data sharing and transfer for new collaborators.

All IPDM actions regarding the help for diffusion of their software tools, or diffusion of their needs in processing to image processing teams seems to be of interest for these associated teams.

Proposals for IPDM transversal node

1. Organizing a **meeting between IT proximity engineers or technician** when existing, IT department delegates, animated by IPDM node with the purpose of exchanging on current hardware infrastructure for data storage and transfer, or facility applications such as booking system, often developed in a parallel way by these teams (similarly to the EUlife IT meeting).

2. Setting up a **centralized access** in Curie Data Center for data publications or sharing associated to **working groups** in FBI.

    a. WG having expressed the interest in a centralized server to share and present their data
    b. WG1 (benchmarking going on)
    c. WG3 (to demonstrate or document probes to potential users)

3. Setting up a **centralized access** in Curie Data Center also for **FBI users to publish their data or gold standard data.**

4. The **centralized repository** would have a purpose of **exchanging data for users of multiple nodes** when nodes have no solutions in place.

5. The **centralized repository** could be used to present **new data modalities to image processing teams**, together with challenges encountered on these data.

6. Proposing templates of **data management plan** on both a facility quality approach view and a project based view from the users (to help proposals and follow up of grant proposals)

7. Organizing a **hackaton/coding party** around **tools to facilitate access to software on grids** (Galaxy/Mobyle, OpenMole)

8. Organizing **Coding party for teams and facilities** in order to facilitate the 'technological transfer' toward users (**diffusion**) through the development of **user friendly tools**. Some users/teams are not convinced by the impact of Icy and are asking for this coding parties for imageJ.

9. A working group in IPDM specialized in **SPIM data processing** would be of interest for several nodes where a SPIM is installed, to coordinate with existing related initiatives. More generally, **tools adapted to processing and visualization of big data** sets are required.

10. Several people would be interested in participating in the project of **metrology/facility monitoring from image data base**.

11. **A coding party for interfacing software tools with data base** (Omero, CiD iManage) would be of interest (for example for deconvolution,..). People using Omero are particularly interested by the development of block processing under ICY developed for CiD iManage.

12. **On-demand focalized training in Icy or on thematic image processing notions** needed in a specific WG, in co-organisation with IPDM would be of interest for most of the facilities/research teams.

13. Defining a **common policy of FBI nodes regarding data responsibility** for external users**.**

14. Organising a **training for facility people for data curation and annotation**.

Success stories from other domain

OpenGIS : regarding interoperability between data processing through **standards**
http://www.opengeospatial.org/ogc

CERN data center: from one equipment LHC -> access and sending data to the data center

http://information-technology.web.cern.ch/about/computer-centre

Proposals not directed related to IPDM

1. The rules and wording of acknowledgements of France Bio Imaging in publication is asked by most of the sites. In particular the differences between using the facility, and using an equipment co-funded by FBI.

2. Reanimating the workgroups forums to facilitate the communications and following of members

3. Adding expertise as an entry was suggested, in particular regarding WP3. This concept could be extended to other working groups.

4. From the user point of view (nuc), a simple way to strengthen the link to users will be to have the FBI link the FBI instrument offer on the booking system of every facilities in the FBI nodes.

5. The centralized directory could be used to present demo images of advanced system to potential users.

ANNEXES (in French) FBI coordination internal use only

MINUTES FROM MARSEILLE VISIT

CRFBIMarseille.pdf

MINUTES FROM BORDEAUX VISIT

CRFBI_Bordeaux.pdf

MINUTES FROM MONTPELLIER VISIT

CRFBIMONTPELLIER.pdf

MINUTES FROM PARIS SUD VISIT

CRFBIParisSud.pdf

MINUTES FROM PARISE CENTRE I VISIT (NOT DONE)

MINUTES FROM PARIS CENTRE II VISIT

CRFBI_Paris2.pdf

LIST OF IT PEOPLES IN NODES

In facilites :

- Marseille IBDML/CIML : Dominique Mondelli
- Bordeaux : IINS : Laurent Chambon
- Montpellier : MRI-DEV (Volker Bäcker) and MRI-NET (Olivier MIQUEL)
- Paris Sud : A. Martel
- Bioemergences : Bernard Martin
- IBENS : Pierre Vincent
- Curie: Sébastien Goud

In associated research team :

- LOB : Simon Dadoun
- Paris Descartes : Vincent de Sars