Etude sur les besoins de Stockage de Laboratoires

Analyse de l'enquête

Ecole Normale Supérieure de Lyon

préparé par	Emmanuel Quemener
contrôlé par	
approuvé par	_____ Date: _____
reference	ENSL-Storage4labs-100415
version	0 draft
date de version	15 avril 2010
nom de document	ENSL-Storage4labs-100415.odt

Destinataires du document

Version							
Date de distribution							
Numéro d'ordre		n/a					
Nom	**Service**						
Chantal Rabourdin Combe	Direction de la recherche						
Véronique Queste	Pôle Recherche						
Véronique Chrétiennot	Direction des Systèmes d'Informations						
Hervé Brunet	Direction des systèmes d'Informations						
Informaticiens	Tous les laboratoires						

Mise à jour

Date	Version	Etat	Pages	Raisons du changement
Initiale	0	Brouillon		
Revue	0.1	Brouillon		Revue par Véronique Queste
Revue	0.2	Final		Remarque de Antonio Munoz

Table des Matières

Index des graphiques

Index de tableaux

1 Introduction

Fin 2009, la direction de la recherche a reçu de la part des laboratoires de biologie de l'établissement une demande de financement pour une infrastructure de stockage.

Dans ce cadre, l'auteur de ce document été mandaté (par une lettre de mission de la direction de la recherche) pour effectuer une évaluation des besoins de stockage, de sauvegarde et d'archivage de tous les laboratoires de l'établissement, site Jacques Monod.

2 Documents applicables

Lettre de mission E. Quemener	http://perso.ens-lyon.fr/emmanuel.quemener/documents/lettre_mission_e.quemener.pdf

3 Glossaire

CQQCOQP	Allographe de QQOQCCP (Qui fait Quoi ?, Où ? Quand ? Comment ? Combien ? et Pourquoi)
LBMC	Laboratoire de Biologie Moléculaire de la Cellule
UVH	Unité de Virologie Humaine
IGFL	Institut de Génomique Fonctionnelle de Lyon
RDP	Laboratoire de Reproduction et de Développement des Plantes
CRMN	Centre lyonnais de Résonance Magnétique Nucléaire
LST	Laboratoire de Sciences de la Terre
LJC	Laboratoire Joliot Curie
LIP	Laboratoire de l'Informatique et du Parallélisme
IXXI	Institut des Systèmes Complexes
UMPA	Unité de Mathématiques Appliquées
BDD	Bases des Données
LimeSurvey	Logiciel d'enquêtes en ligne : http://www.limesurvey.org/ (utilisé pour la seconde série d'enquêtes)
phpESP	Logiciel d'enquêtes en ligne : http://sourceforge.net/projects/phpesp/files/ (utilisé pour la première série d'enquêtes)
MB/s	Mega Bits par seconde
Mo	Mega Octet (normalement 2^20 soit 1048576 octets, mais standardisé en 2007 à 1 million d'octets)
Go	Giga Octet (normalement 2^30 soit 1073741824 octets, mais standardisé en 2007 à 1 milliard d'octets)
To	Tera Octet (normalement 2^40 soit 1099511627776 octets, mais standardisé en 2007 à mille milliards d'octets)
Po	Peta Octet (normalement 2^50 soit 1125899906842624 octets, mais standardisé en 2007 à un million de milliards d'octets)
VA	Volt.Ampère (unité estimée de consommation électrique)
BTU	British Thermal Unit (Unité d'énergie définissant une quantité de chaleur). Généralement, BTU = 3.41 Volt.Ampère
GE	Gigabit Ethernet (standard réseau permettant un transfert de données symétrique à une vitesse de 125 Mo/s)
GTI	Garantie Totale d'Intervention

4 Démarche

Cette étude se déroulera en 6 étapes, avant la remise du rapport à la direction de la recherche fin mars (fin février sera marqué par la remise d'un rapport préliminaire sur les expressions de besoins des laboratoires) :

1. réalisation d'un questionnaire permettant d'établir un état des lieux de l'existant et des besoins associés (semaine 1) ;
2. expédition du questionnaire aux directeurs de laboratoire et invitation à rediffuser l'information aux personnels concernés (fin semaine 1)
3. analyse des retours de questionnaires (semaine 9)
4. consultation des directeurs, de leurs responsables de plate-forme (semaine 10);

5. synthèse des besoins, analyse et rédaction d'un cahier des charges (semaine 13) ;
6. proposition d'une infrastructure adaptée dans le rapport final (semaine 14).

5 Modélisation du circuit des données numériques

« Un bon croquis vaut mieux qu'un long discours. » disait Napoléon Bonaparte.

Un schéma synoptique (Illustration 1, page 11) permet de visualiser rapidement le circuit suivi de leur genèse à leur diffusion.

Plutôt que de s'axer dans un premier temps sur la nature des données, considérons tout d'abord les processus, abordons ensuite les données, puis les actions

5.1 Les processus

Quatre processus manipulent ces données à l'aide de plates-formes :

- l'expérience : la « génération » des données, leur indexation, … ;
- le traitement : leurs réduction, synthèse, analyse, indexation, …
- la valorisation : leur transformation en contenu scientifique. Cela constitue le “cœur de métier” du chercheur”.
- l'exploitation : leur diffusion sur tous les médias scientifiques

5.2 Les données

Ainsi, les données, à la source ou destination de ces processus sont finalement de 3 natures distinctes :

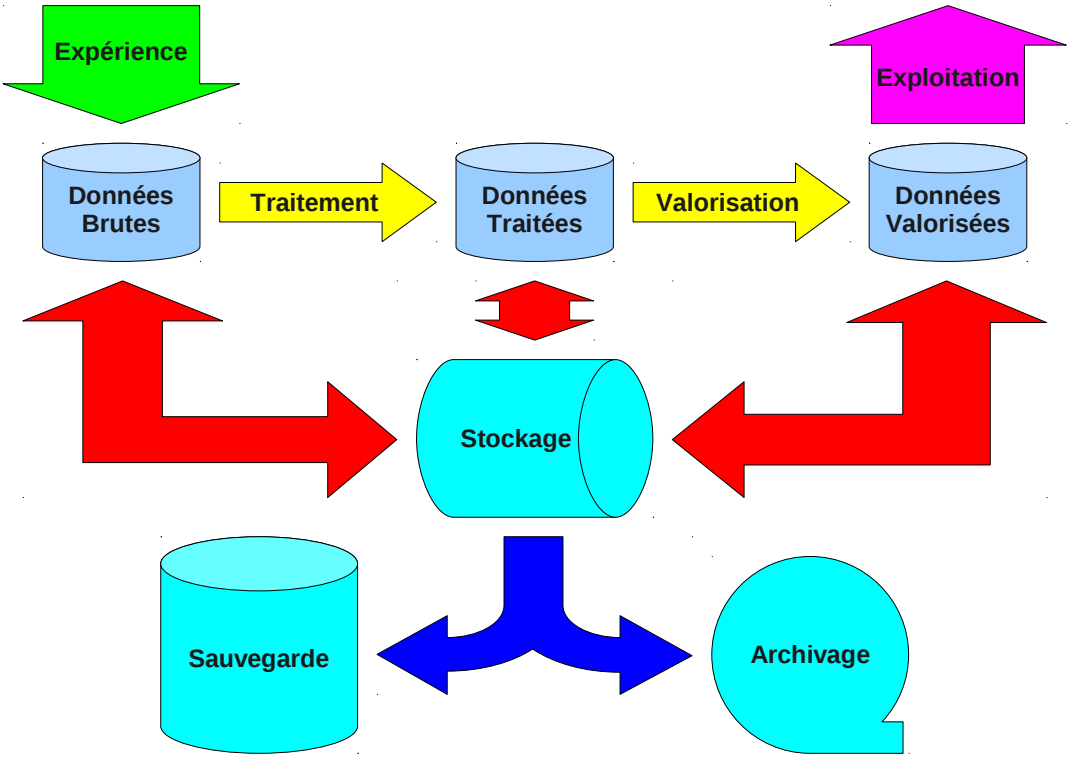
- les données « brutes » : directement issues des expériences ;
- les données « traitées » : premiers résultats ;
- les données « valorisées » : résultats à destination des travaux de publication.

5.3 Les « formes » d'entrepôt de ces données

De plus, le schéma synoptique présente un entrepôt de ces données avec une triple nature :

- le stockage : accès direct aux données
- la sauvegarde : duplication complète/partielle, synchrone/asynchrone de l'espace de stockage dans un endroit apportant une sûreté aux données en cas de perte du stockage
- l'archivage : état du stockage complet ou partiel, à un instant dans le passé. Ces états sont d'une fréquence et d'une pérennité à définir

Illustration 1: Modélisation du circuit des données numériques



Pour qu'aucune information importante n'échappe à l'étude préliminaire, a été choisie la collecte basée sur une démarche analytique classique. Elle propose de répondre aux sept questions élémentaires : Pourquoi ? Quoi ? Qui ? Quand ? Où ? Combien ? Comment ?

6 Questionnaires

6.1 De la démarche analytique aux questionnaires

La démarche analytique utilisée, plus connue sous l'allographe CQQCOQP, a été "transformée" en questions simples les plus générales possibles pour que les utilisateurs de plate-forme puissent, quelque soit leur discipline, répondre simplement.

Le Laboratoire Joliot Curie, par l'aide précieuse de son informaticien, a été le terrain d'expérimentation utilisé pour tester les premières versions de questions. L'avantage, de plus, de ce laboratoire, était de disposer d'une population de scientifiques physiciens, chimistes et biologistes . Leurs remarques, très judicieuses, ont été intégrées dans la réalisation des questionnaires.

Les trois questionnaires sont disponibles en annexe de ce document :

- l'expérience (page 43, chapitre 14.1)
- le traitement (page 44, chapitre 14.2)
- l'exploitation (page 44, chapitre 14.3)

6.2 Une mise en forme informatique en deux temps

Deux versions de questionnaires en ligne ont été utilisées phpESP d'un côté et LimeSurvey de l'autre.

Le premier, phpESP, a été utilisé au lancement de l'étude. Cependant sont apparus rapidement de gros soucis à l'exploitation : des questions avaient été modifiées à la mise en ligne et des manipulations des navigateurs étaient

indispensables pour répondre plusieurs fois au même questionnaire (dans le cas d'un changement de plate-forme). Pour finir, le questionnaire n'était accessible qu'à l'intérieur du réseau informatique de l'ENS-Lyon, site Monod. Il était donc difficile (mais possible) d'y répondre.

Le second, LimeSurvey, a été diffusé en ligne mi-février. Il évitait tous les errements à l'usage de phpESP, permettait un usage en ligne de partout sur Internet, et, surtout, permettait facilement d'ajouter des légendes expliquant les questions et donnant des exemples de réponses.

6.3 Le « pourquoi » d'une exclusion de la valorisation

Le processus de valorisation, cœur de métier du chercheur, ne fait pas l'objet d'un questionnaire. Cependant, étant donné les usages des outils informatiques, leur impact dans le volume de stockage à mettre à disposition sera non négligeable.

Le volume nécessaire pour cette étape de valorisation sera rajouté à partir d'un montant estimé des besoins d'un utilisateur dans le cadre de son travail quotidien (hors manipulation de gros volumes liés au processus de traitement ou d'exploitation).

7 Résultats préliminaires

7.1 La chasse aux résultats

Fin janvier 2010, seule une douzaine de questionnaires avaient été complétés. De plus, seuls 4 laboratoires sur la douzaine de l'établissement avaient répondu.

Par la suite, une relance directe de directeurs et des informaticiens de laboratoires a été réalisée.

7.2 Réponses aux questionnaires

7.2.1 Premier bilan

Au 19 mars 2010, date de la fermeture des enquêtes, une exportation des résultats présentaient les réponses suivantes :

- 46 réponses aux questionnaires sur le processus « expérience » ;
- 45 réponses aux questionnaires sur le processus « traitement »
- 11 réponses aux questionnaires sur le processus « exploitation ».

7.2.2 Éléments d'analyse

L'analyse se déroule en deux temps.

D'abord, un premier temps consacré à l'examen des questions connexes liées aux contextes dans la gestion des données numériques : l'analyse qualitative. Elle comprend cependant un certain nombre de données quantitatives relatives au contexte.

Ensuite, un second temps destiné à analyser, à partir des estimations numéraires, les volumes nécessaires en stockage, leur projection par extrapolation sur les trois prochaines années.

8 Analyse qualitative

8.1 Sur les besoins de stockage dans les plates-formes expérimentales

8.1.1 Sur les laboratoires

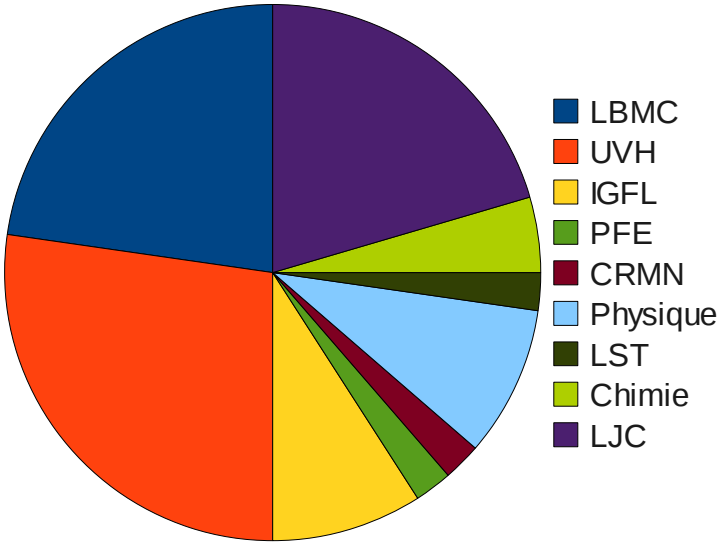


Illustration 2: Réponses des laboratoires sur les expériences

Remarques générales :

- sont représentés tous les laboratoires du site Monod à l'exception du RDP, du LIP, du CRAL, de l'UMPA et l'IXXI ;
- les laboratoires de biologie (à l'exception du RDP) se sont fortement mobilisés pour cette étude ;
- le laboratoire Joliot Curie,malgré son faible nombre de chercheurs, a contribué pour près du quart des réponses.

8.1.2 Sur la nature des données

Remarques générales :

- les multiples réponses étaient possibles dans cette question ;
- essentiellement, images et des vidéos sont en sortie des manipulations.

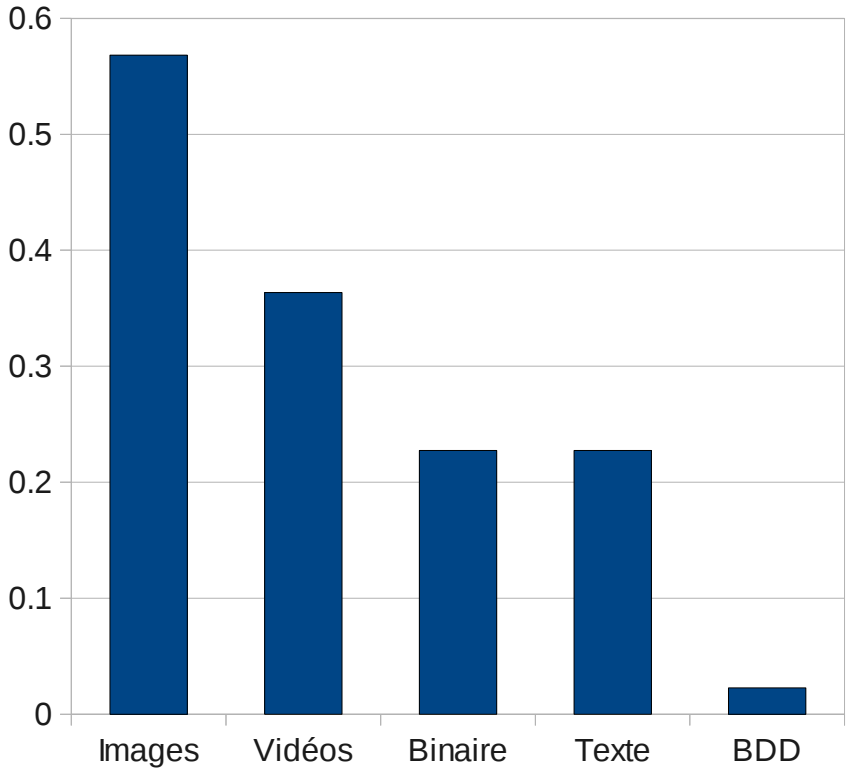


Illustration 3: Nature des signaux stockés

8.1.3 Sur le « qui réalise » et « qui exploite » les expériences »

Remarques générales :

- essentiellement, ce sont les doctorants et les chercheurs qui réalisent et exploitent les expériences
- toutes les catégories participent, sans exception, aux expériences
- il existe une large adéquation entre ceux qui réalisent les expériences et ceux qui les exploitent

Spécifications fonctionnelles :

- large éventail de population exigeant la mise en place d'une gestion de droits d'utilisateurs ;
- des personnels temporaires réalisant les manipulations exigeant que leurs données restent encore accessibles à leur responsable après leur départ ;
- des prestataires externes ou des collaborateurs utilisant les plate-formes exigeant la mise en place d'un espace ouvert vers l'extérieur de l'établissement pour que ces derniers puissent récupérer le fruit de leurs manipulations.

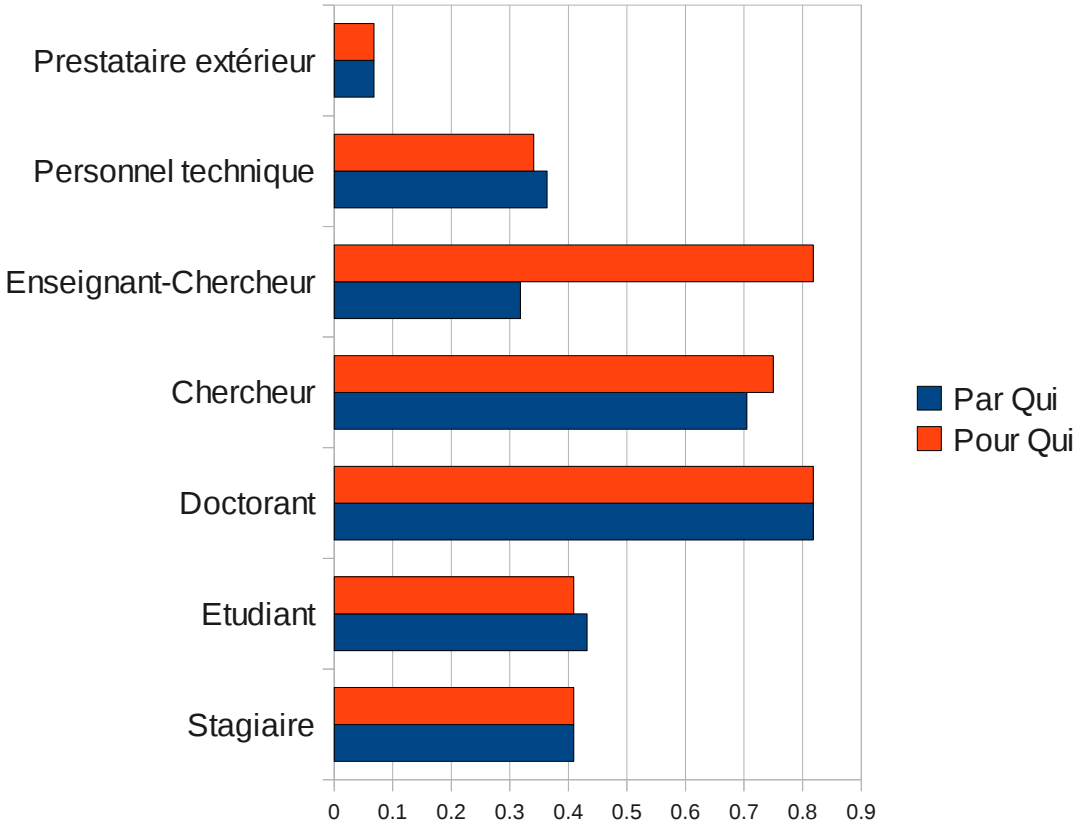


Illustration 4: Par qui et pour qui sont réalisées les expériences ?

8.1.4 Sur la durée moyenne des expériences

Remarques générales :

- les expériences durent de quelques minutes à un mois : cela constitue presque 4 ordres de grandeur

Spécfications fonctionnelles :

- le fait que les acquisitions aient des durées moyennes importantes exigeant de la part des équipements de stockage une disponibilité continue accrue, ou une reprise sur incident quasi-transparente pour la manipulation.

8.1.5 Sur le nombre d'expériences par semaine

Remarques générales :

- la fréquence des expériences varie de 1 par mois à 125 par semaine : plus de 2 ordres de grandeur
- ces fréquences sont largement corrélées aux durées moyennes des expériences et imposent donc les mêmes contraintes sur la disponibilité de l'espace de stockage

8.1.6 Sur le nombre annuel d'expériences

Remarques générales :

- le nombre d'expériences sur l'année est très différent selon les disciplines
- plusieurs laboratoires effectuent plus de 1000 expériences par an

Spécifications fonctionnelles :

- nombre de manipulations importants exigeant une indexation fine et des règles d'organisation strictes dans le stockage hiérarchique des données.

8.1.7 Sur la croissance du taux de manipulation

Remarques générales :

- plus de la moitié compte doubler leur activité
- près de 1/6 compte la quintupler
- une telle augmentation, sur la projection à 3 années, génère un besoin cumulé important. Un doublement sur 3 ans équivaut à 6 ans de cette année. Un quintuplement équivaut à 12 ans.

8.1.8 Sur la durée de conservation des données

Remarques générales :

- la conservation des données pour quelques jours est largement minoritaire ;
- la conservation au delà de plusieurs années dépasse les deux tiers. Cette longue conservation des données est également corrélée avec de grand nombre d'expériences générant de gros volume de données. Il peut donc être considéré que, dans l'estimation du besoin de stockage, les données sont stockées « aussi longtemps que possible » ;
- étant donné le grand nombre d'expériences réalisées, leur durée de conservation, une indexation associée à une gestion des droits d'accès, sera indispensable pour que ces données soient correctement accessibles dans le temps.

Spécifications fonctionnelles :

- une durée de conservation des données de manipulations exigeant une gestion des données et de leur accès largement supérieure à la présence de ceux les réalisant.

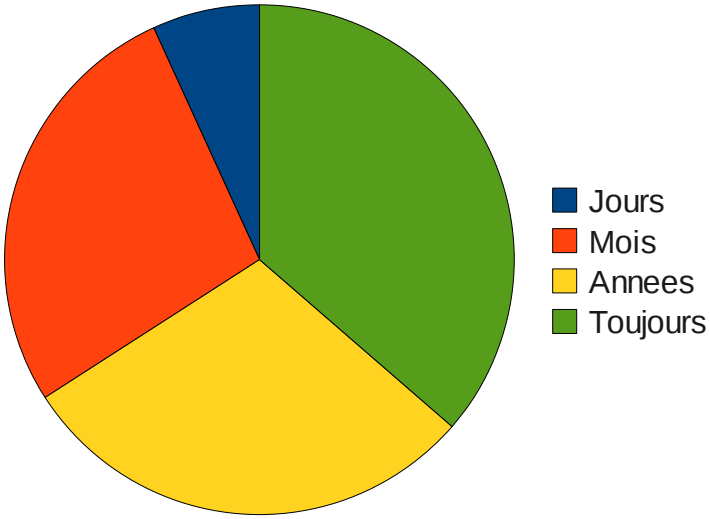


Illustration 5: Durée de conservation des données

8.1.9 Sur la situation de l'équipement

Remarques générales :

- les salles sont dédiées à l'expérience dans la moitié des cas et partagées dans un tiers des cas

Spécifications fonctionnelles :

- une forte pression sur l'utilisation des équipements exgieant, dans le cadre d'un stockage distant, une analyse de la connectivité réseau pour chacun d'eux. En effet, pour les différents locaux, il est indispensable que les adductions réseau ne soient pas le goulot d'étranglement dans la circulation des données.

8.1.10 Sur le stockage des données brutes

Remarques générales :

- les données sont stockées dans un peu plus d'un tiers des cas sur un serveur ;
- sur support amovible ou localement sur la machine expérimentale, les données ne sont vraisemblablement pas sauvegardées, ou, au moins, dupliquées dans un autre lieu.

Spécifications fonctionnelles :

- des pratiques sur l'usage de support amovible ou locaux exigeant une mise en place urgente de volumes de stockage adaptés.

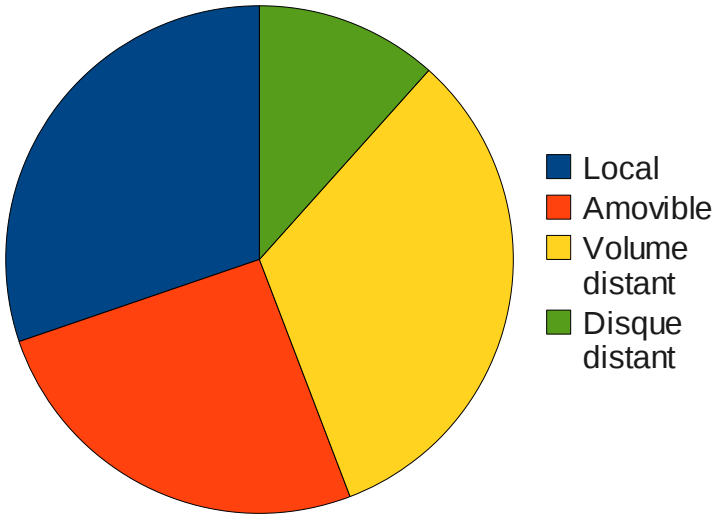


Illustration 6: Support des données stockées

8.1.11 Sur le volume moyen d'une expérience

Remarques générales :

- les expériences ont des volumes de données très divers, mais celles générant des volumes de quelques centaines de Mo à quelques Go sont majoritaires ;
- déplacer ces volumes précédents exige de plusieurs minutes à plusieurs dizaines de minutes sur un réseau à 100 MB/s.

Spécifications fonctionnelles :

- des volumes de données exigeant la mise en place d'une amélioration des éléments intervenant dans le transfert (disque « local », interconnexion réseau, volume « destination »).

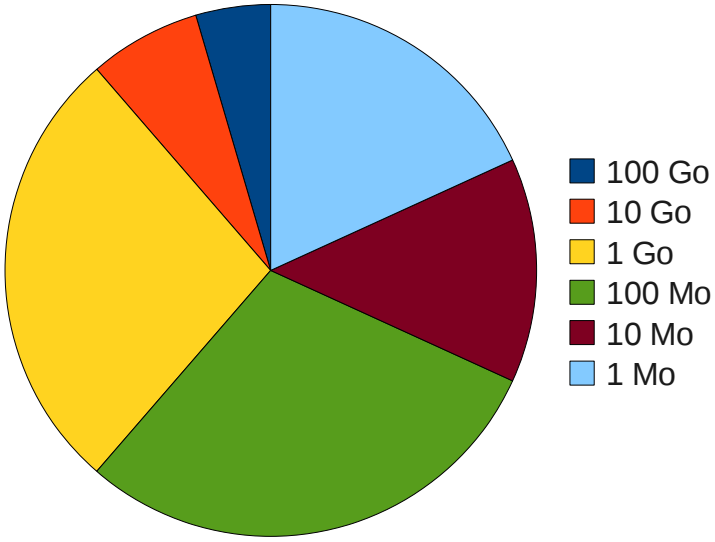


Illustration 7: Volume moyen d'une expérience

8.1.12 Sur le sous-dimensionnement du réseau pour transférer les résultats

Remarques générales :

- 40% des sondés trouvent le réseau sous-dimensionné : cela confirme le fait que le réseau est un facteur limitant dans le transfert des données
- la limitation du réseau est donc un des verrous à lever auprès des plateformes expérimentales.

Spécifications fonctionnelles :

- un jugement « insuffisant » du réseau informatique exigeant une amélioration de la liaison entre les postes de manipulations et les machines d'analyse sur le réseau informatique.

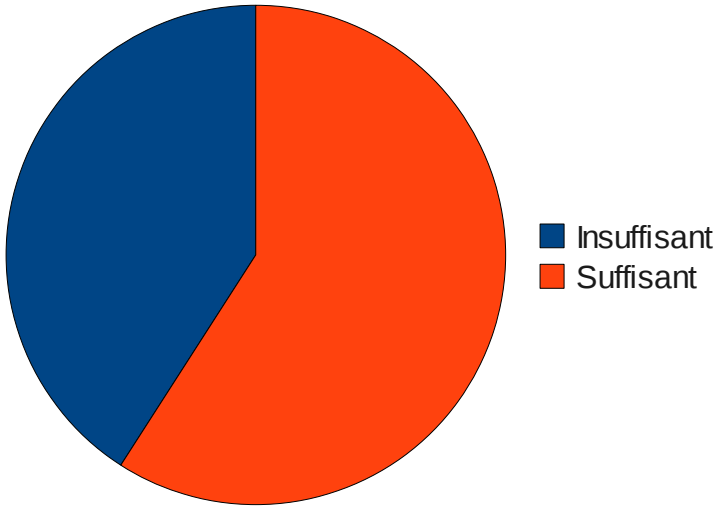


Illustration 8: Adaptation du réseau aux transferts de données

8.1.13 Sur les contraintes techniques d'exploitation

Remarques générales :

- près d'un tiers des utilisateurs sondés ne peut utiliser un répertoire distant
- le fait de ne pas pouvoir utiliser un répertoire distant pour les sorties de ses expériences a plusieurs origines :
  - stockage du serveur distant insuffisant,
  - vitesse d'écriture sur le serveur insuffisante (protocole utilisé ou matériel),
  - vitesse du réseau d'interconnexion,
  - contrainte de la plate-forme dans un réseau dédié ne communiquant pas directement avec le serveur ;
- ces contraintes militent pour une étude approfondie de solutions permettant de se libérer de ces contraintes.

Spécfications fonctionnelles :

- des contraintes exigeant pour les plate-formes expérimentales citées une étude de la chaîne de transmission.

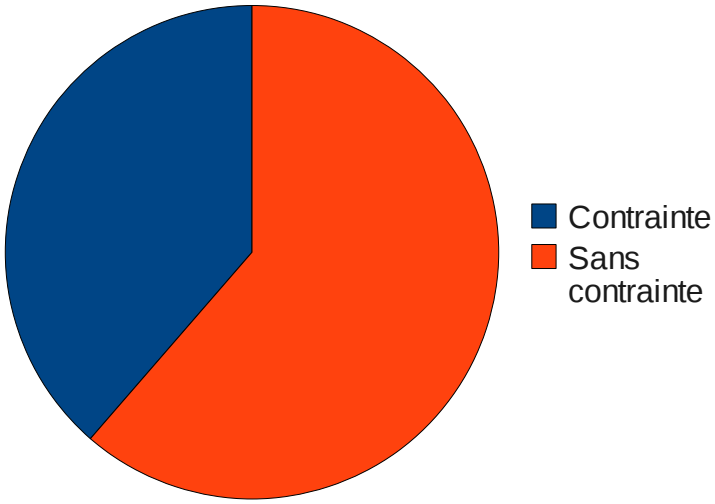


Illustration 9: Contraintes techniques sur le stockage

8.1.14 Sur la nature du logiciel

Remarques générales :

- une immense majorité des plateformes utilise des logiciels propriétaires « fermés »
- la nature « fermée » des logiciels imposent des usages exigeant une adaptation de la politique générale du laboratoire : l'équipement peut être « isolé » parce que les programmes propriétaires ne respectaient pas les règles les plus élémentaires de sécurité (fonctionnement systématique en administrateur, etc)
- les logiciels d'acquisition font partie intégrante de la plateforme et sortent de la gestion informatique des applications classiques du laboratoire.

Spécifications fonctionnelles :

- des solutions propriétaires fermées, souvent « mal conçues » d'un point de vue informatique (gestion des droits et des accès), exigeant la mise en place d'une solution de stockage la plus proche du matériel (donc si possible se comportant le plus possible comme un composant interne à la machine).

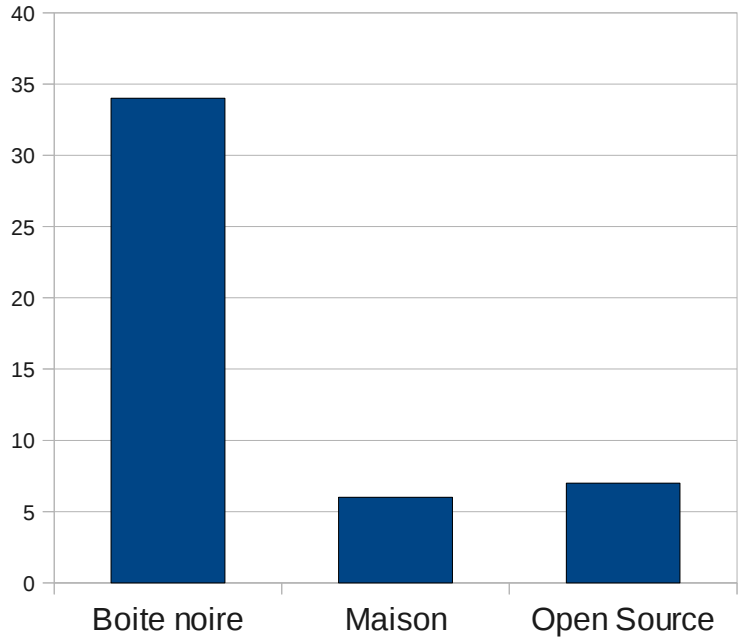


Illustration 10: Nature du logiciel d'acquisition

8.2 Sur les besoins de stockage dans les plates-formes de traitement

8.2.1 Sur les laboratoires

Remarques générales :

- tous les laboratoires sont représentés, à l'exception du laboratoire de physique, du RDP, l'UMPA et l'IXXI ;
- comme pour le questionnaire sur les expériences, les laboratoires de biologie se sont bien mobilisés ;
- le LJC a poursuivi sur sa lancée en représentant ici également près du sixième des retours ;
- le LIP a finalement répondu aux questions relatives aux traitements.

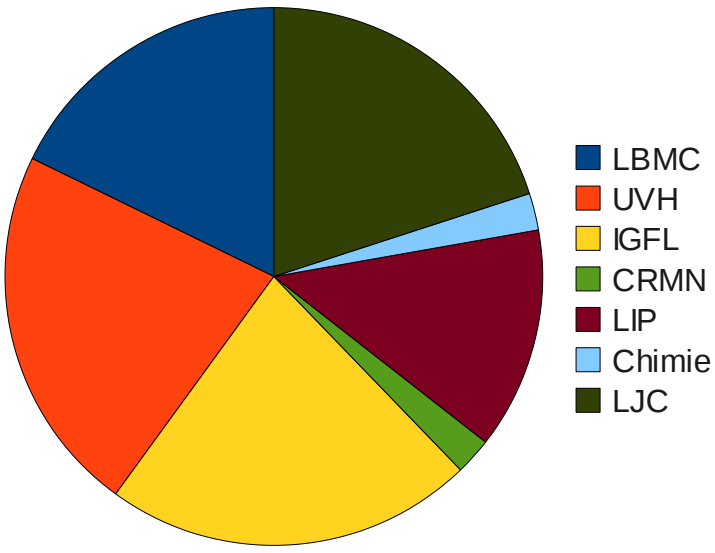


Illustration 11: Distribution des réponses des laboratoires

8.2.2 Sur la nature des données

Remarques générales :

- les données entrées et sorties restent encore essentiellement des images et des vidéos ;
- les extractions sous forme de fichiers binaires pour tableur représentent maintenant une part importante des fichiers traités ;
- les bases de données, en entrée et en sortie, sont assez rares, mais peuvent représenter de gros volumes (bases MySQL de plusieurs centaines de Go pour des traitements au LIP) ;
- la nature des données est donc très disparate mais la part non négligeable de bases de données va exiger une attention particulière sur les solutions à proposer .

Spécifications fonctionnelles :

- une présence de base de données exigeant que les fichiers des bases de données ne soient pas accessibles comme de simples documents mais comme un espace équivalent à un volume de stockage interne à la machine.

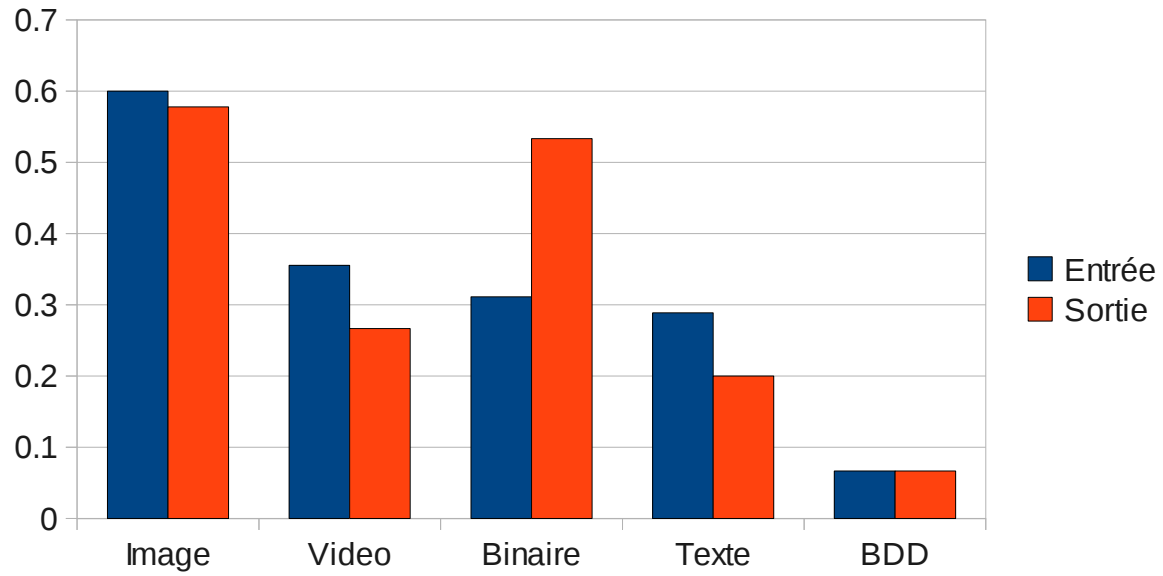


Illustration 12: Nature des données en entrée et sortie des traitements

8.2.3 Sur la nature des équipements de traitement

Remarques générales :

- près de 9/10 des sondés utilisent leur poste de travail ou leur station de travail pour les traitements ;
- les unités de calcul dédiées (clusters de laboratoire ou clusters du PSMN) représentent le reste ;
- une majorité utilise leur poste de travail pour le traitement pour une raison technique : une clé électronique est nécessaire pour traiter les données brutes : dans l'UVH, il n'existe pas de plate-forme dédiée au traitement de telle ou telle plate-forme. Le logiciel de traitement est installé sur tous les postes mais n'est utilisable que par l'individu disposant cette clé installée sur son poste.

Spécifications fonctionnelles :

- des usages préconisant la mise en place de postes dédiés au traitement, disposant d'une clé à demeure et permettant un accès distant pour l'usage des logiciels ;
- des transferts de données vers des unités dédiées mais distantes exigeant une interconnexion réseau optimale.

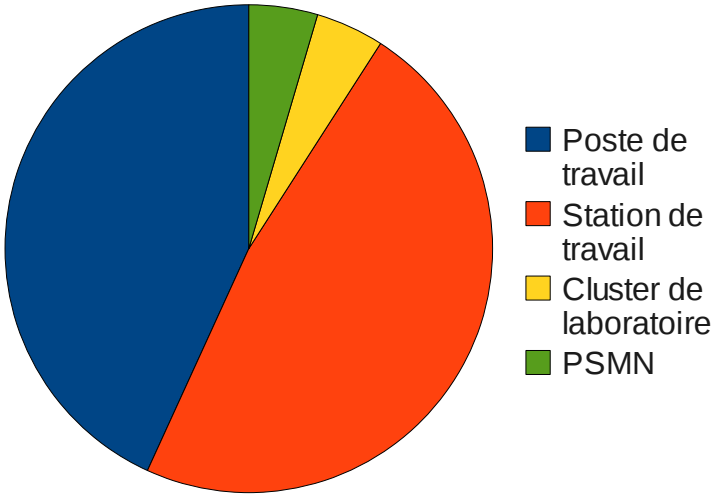


Illustration 13: Nature de l'équipement de traitement

8.2.4 Sur le « qui réalise ou exploite les traitements »

Tous, mais très majoritairement doctorant et chercheur

Remarques générales :

- toutes les catégories participent, sans exception, aux traitements
- il existe une large adéquation entre ceux qui réalisent les traitements et ceux qui les exploitent
- ce large éventail de population va exiger la mise en place d'une gestion de droits d'utilisateurs
- le fait que des prestataires externes ou des collaborateurs utilisent les plate-formes va également exiger la mise en place d'un espace ouvert vers l'extérieur de l'établissement pour que ces derniers puissent récupérer le fruit de leurs manipulations.

Spécifications fonctionnelles :

- large éventail de population exigeant la mise en place d'une gestion de droits d'utilisateurs ;
- des personnels temporaires réalisant les manipulations exigeant que leurs données restent encore accessibles à leur responsable après leur départ ;
- des prestataires externes ou des collaborateurs utilisant les plate-formes exigeant la mise en place d'un espace ouvert vers l'extérieur de l'établissement pour que ces derniers puissent récupérer le fruit de leurs manipulations.

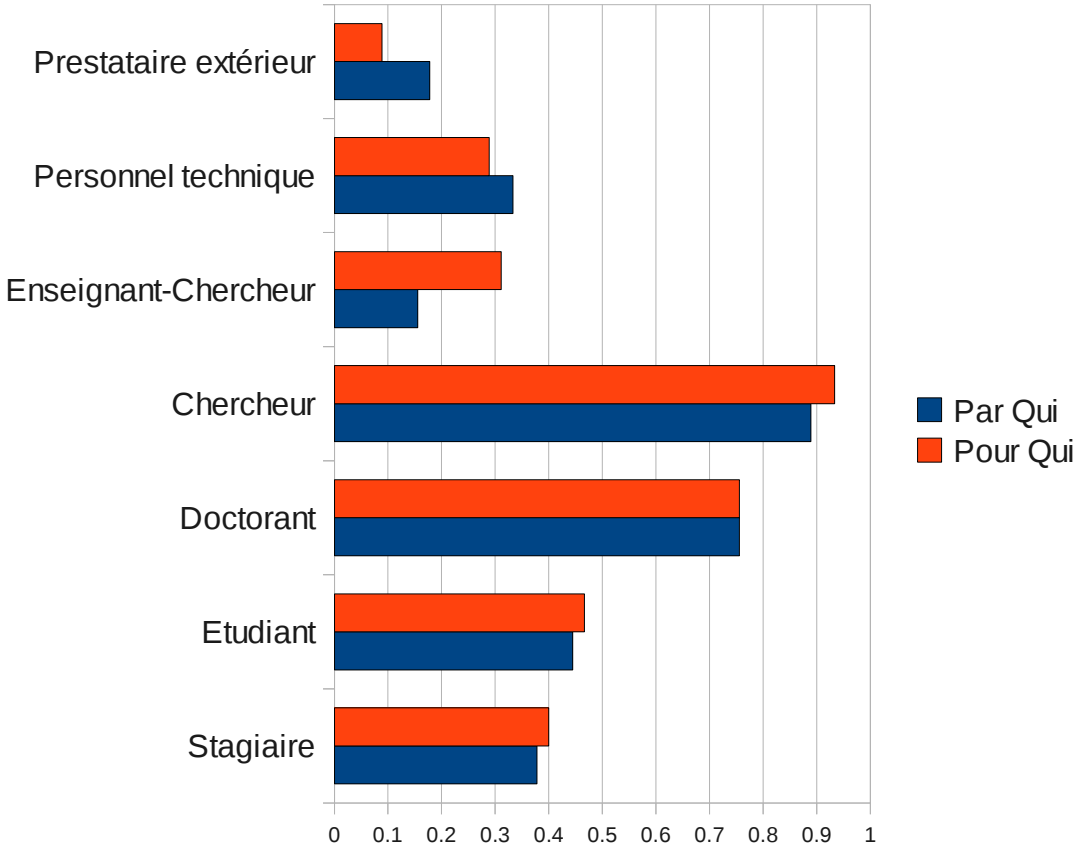


Illustration 14: Qui réalise ou exploite les traitements ?

8.2.5 Sur la durée moyenne d'un traitement

Remarques générales :

- les traitements durent de quelques secondes à quelques jours : cela constitue plus de 5 ordres de grandeur ;
- le fait que les traitements aient des durées moyennes aussi disparates va exiger de la part des équipements de stockage une disponibilité accrue.

Spécfications fonctionnelles :

- le fait que les traitements aient des durées moyennes importantes exigeant de la part des équipements de stockage une disponibilité continue accrue, ou une reprise sur incident quasi-transparente pour le traitement.

8.2.6 Sur le nombre de traitements par semaine

Remarques générales :

- le nombre de traitements va de 1 par mois à 125 par semaine, soit 3 ordres de grandeur
- ces fréquences sont largement corrélées aux durées moyennes des expériences et imposent donc les mêmes contraintes sur la disponibilité de l'espace de stockage

8.2.7 Sur le nombre annuel de traitements

Remarques générales :

- le nombre de traitements sur l'année est très différent selon les disciplines
- plusieurs laboratoires effectuent plus de 1000 traitements par an

- de tels montants de traitements exigent, pour être correctement retrouvés, d'être finement indexés

Spécifications fonctionnelles :

- nombre de manipulations important exigeant une indexation fine et des règles d'organisation stricte dans le stockage hiérarchique des données.

8.2.8 Sur la croissance du taux de manipulation

un problème sur le questionnaire n'a pas permis de disposer d'une grande finesse dans la projection de croissance. Cependant, une augmentation est citée dans 70% des cas

8.2.9 Sur la durée de conservation des données traitées

Remarques générales :

- la conservation des données pour quelques jours est largement inexistante, pour quelques semaines anecdotique et pour quelques mois largement minoritaire ;
- la conservation au delà de plusieurs années dépasse les trois quarts. Cette longue conservation des données est également corrélée avec de grand nombre de traitements générant de gros volume de données. Il peut donc être considéré que, dans l'estimation du besoin de stockage, les données sont stockées « aussi longtemps que possible » ;
- étant donné le grand nombre de traitements réalisés, leur durée de conservation, une indexation associée à une gestion des droits d'accès, sera indispensable pour que ces données soient correctement accessibles dans le temps.

Spécifications fonctionnelles :

- une durée de conservation des résultats de traitements exigeant une gestion des données et de leur accès largement supérieure à la présence de ceux les réalisant.

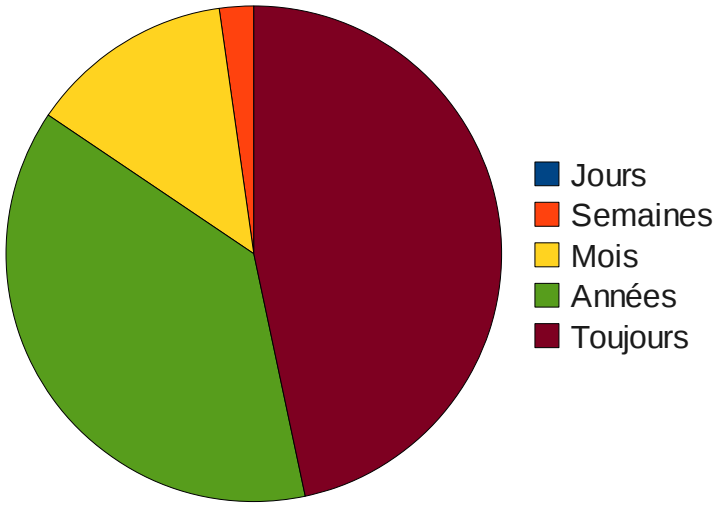


Illustration 15: Conservation des données traitées

8.2.10 Sur le stockage des données brutes

Remarques générales :

- plus de deux tiers sont stockées localement avant analyse, un quart sur support amovible ;
- presque aucune donnée n'est récupérée sur le dossier distant
- l'immense majorité du stockage local illustre le manque d'espace de stockage et la vulnérabilité de ces ressources sur des postes pas nécessairement sauvegardés.

Spécifications fonctionnelles :

- des pratiques sur l'usage de support amovible ou locaux exigeant une mise en place urgente de volumes de stockage adaptés.

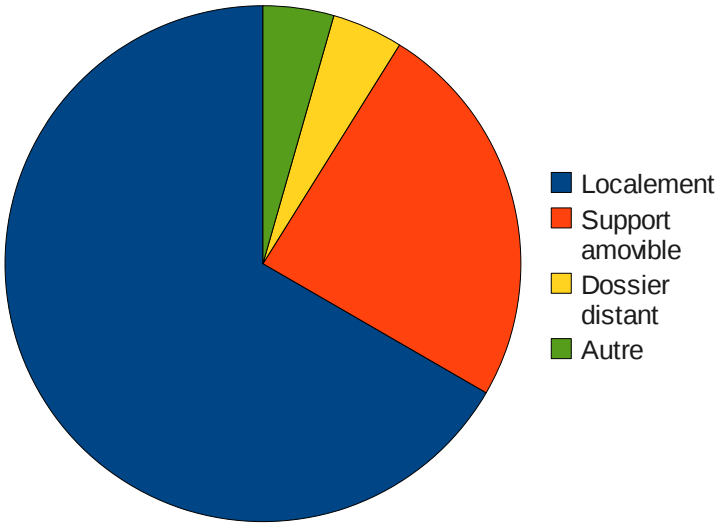


Illustration 16: Nature du stockage en entrée du traitement

8.2.11 Sur le stockage des données traitées

Remarques générales :

- plus de la moitié des données traitées sont conservées localement, un petit sixième sur support amovible ;
- près d'un quart des données est stocké sur le disque distant : visiblement, la définition du terme « disque distant » n'a pas été correctement comprise : elle a été comprise comme étant un répertoire distant, ou un dossier distant ;
- la majorité du stockage local illustre le manque d'espace de stockage et la vulnérabilité de ces ressources sur des postes pas nécessairement sauvegardés.

Spécifications fonctionnelles :

- des pratiques sur l'usage de support amovible ou locaux exigeant une mise en place urgente de volumes de stockage adaptés.

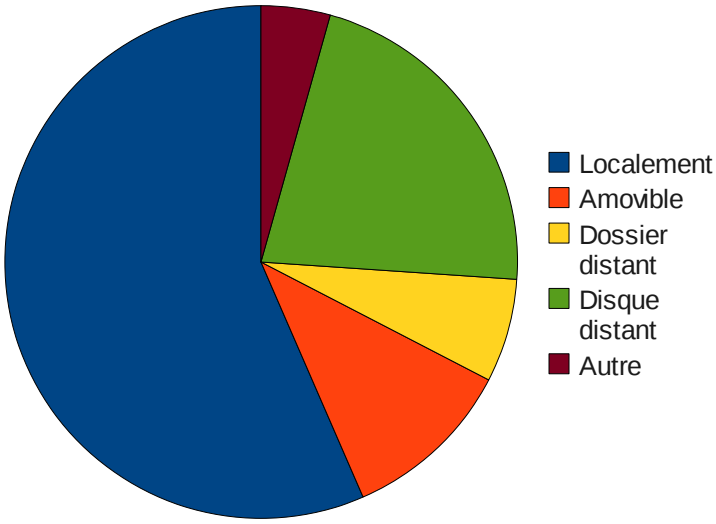


Illustration 17: Nature du stockage en sortie du traitement

8.2.12 Sur le volume moyen d'un traitement

Remarques générales :

- les traitements ont des volumes de données très divers, mais ceux utilisant ou générant des volumes de quelques centaines de Mo à plusieurs dizaines de Go sont majoritaires ;
- déplacer ces volumes précédents exige de plusieurs minutes à plusieurs dizaines de minutes sur un réseau à 100 MB/s.

Spécifications fonctionnelles :

- des volumes de données exigeant une modification des composants intervenants dans le transfert (disque « local », interconnexion réseau, volume « destination »).

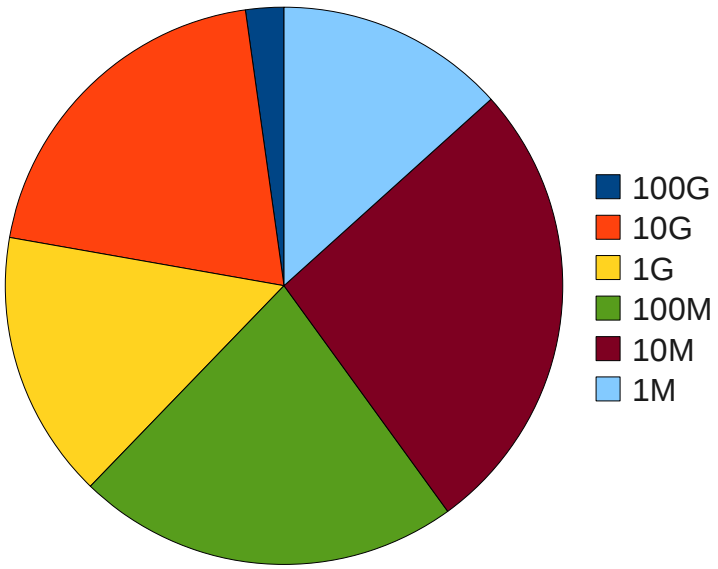


Illustration 18: Volume moyen d'un traitement

8.2.13 Sur le sous-dimensionnement du réseau pour transférer les résultats

Remarques générales :

- 40% des sondés trouvent le réseau sous-dimensionné : cela confirme le fait que le réseau est un facteur limitant dans le transfert des données
- la limitation du réseau est donc un des verrous à lever auprès des plateformes expérimentales.

Spécifications fonctionnelles :

- un jugement « insuffisant » du réseau informatique exigeant une amélioration de la liaison entre les postes de traitements et les autres équipements fournissant ou recevant ces données.

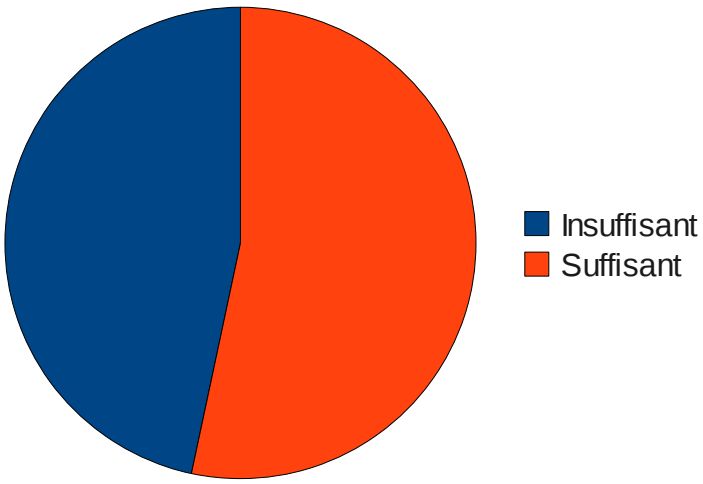


Illustration 19: Adaptation du réseau au transfert de données

8.2.14 Sur les contraintes techniques d'exploitation

Remarques générales :

- près d'un tiers des utilisateurs sondés ne peut utiliser un répertoire distant
- les mêmes conclusions que celles portées sur les plateformes expérimentales peuvent être reprises.
- le fait de ne pas pouvoir utiliser un répertoire distant pour les sorties de ses expériences a plusieurs origines :
  - stockage du serveur distant insuffisant,
  - vitesse d'écriture sur le serveur insuffisante (protocole utilisé ou matériel),
  - vitesse du réseau d'interconnexion,
  - contrainte de la plate-forme dans un réseau dédié ne communiquant pas directement avec le serveur ;
- ces contraintes militent pour une étude approfondie de solutions permettant de se libérer de ces contraintes.

Spécifications fonctionnelles :

- des contraintes exigeant pour les plate-formes de traitement citées une étude de la chaîne de transmission.

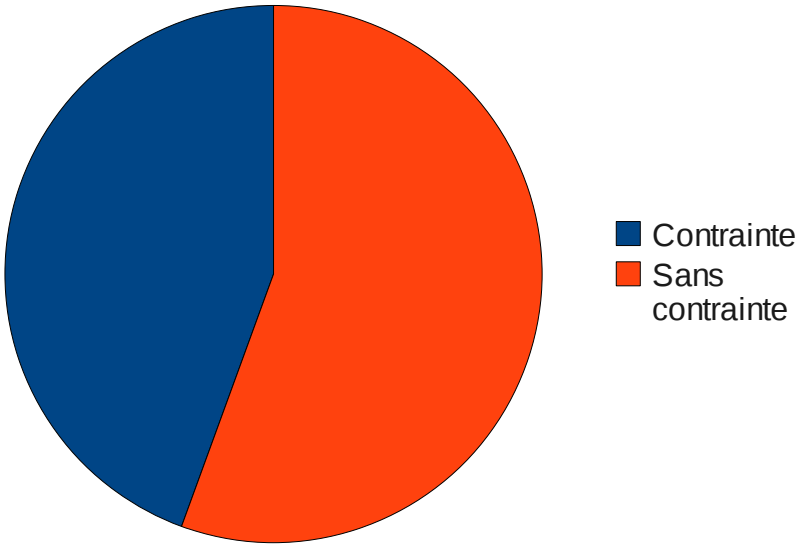


Illustration 20: Contraintes d'exploitation des données de traitements

8.2.15 Sur la nature du logiciel

Le questionnaire était entaché d'une incohérence dans les réponses. Il n'a pas été possible d'établir une statistique cohérente à partir des réponses données.

9 Analyse quantitative

9.1 Avertissement : aménagement des données « incohérentes »

Lorsque que les réponses aux questionnaires étaient incomplètes ou donnaient des estimations trop pharaoniques, les personnes ayant répondu aux questionnaires ont été contactées pour un complément d'information. Cela a été le cas pour le CRMN, pour le LBMC, pour l'IGFL : des gros besoins ont été ainsi confirmés.

Dans 2 cas, à l'IGFL et au LBMC, les personnes ont été contactées par courriel et par téléphone, mais n'ont pas jugé « utile » de répondre à cette demande de clarification exprimée par les messages laissés. Par précaution, leurs réponses quantitatives auraient pu être simplement évincées (agir plus par la « médiane » que par la « moyenne »). Cependant, pour ne pas risquer de laisser leurs besoins orphelins d'une réponse technique, ces derniers ont été revus à la baisse : ils ont été estimés comme étant identique au maximum exprimé dans le même laboratoire.

9.2 Méthodologie

Les questionnaires demandaient :

- le nombre d'expériences par semaine ;
- la croissance estimée du nombre d'expériences à 3 ans ;
- le volume moyen d'une expérience.

De manière à se placer dans le pire des scénarii, il est considéré que :

- les expériences « tournent » 50 semaines par an
- la croissance estimée à 3 ans considère le stockage de cette année ajouté à celui des 3 années suivantes, soit 4 années de stockage
- la croissance est considérée comme linéaire sur les 3 prochaines années
- le volume moyen considéré pour chaque expérience est borné par ce qui suit :
  - quelques méga-octets : 5 Mo

- quelques dizaines de méga-octets : 50 Mo
- quelques centaines de méga-octets : 500 Mo
- quelques giga-octets : 5 Go
- quelques dizaines de giga-octets : 50 Go

Tout d'abord, pour chaque réponse, un nombre d'expériences ou de traitements à l'année a été estimée.

Puis, une projection du nombre d'expériences ou de traitements pour les 3 prochaines années a été faite.

Ensuite, le volume nécessaire pour chaque plate-forme sur les expériences ou traitements pour chaque année a été calculé, puis cumulé.

Enfin, chaque réponse a été groupée par laboratoire pour établir un total, par année, puis cumulé.

Les expressions de besoins pour les plateformes expérimentales et les plateformes de traitements seront représentées par un tableau puis trois graphiques, représentant :

- la répartition de l'expression des besoins par laboratoire, sur les 4 années
- la répartition cumulée pour les 4 années par laboratoire
- le cumul de tous les laboratoires pour chaque année et sa progression

9.3 Sur les besoins de stockage dans les plates-formes expérimentales

	Année 0	Année 1	Année 2	Année 3	4 années
LBMC	29655.5	41572	53485	65401	**190113.5**
UVH	584.5	753	919	1086	**3342.5**
IGFL	26675	35258	43842	52425	**158200**
CRMN	1000	1333	1667	2000	**6000**
Physique	29125	38808	48492	58175	**174600**
LST	200	267	333	400	**1200**
Chimie	2.5	3	3	3	**11.5**
LJC	10438.5	13896	17353	20811	**62498.5**
Total	**97681**	**131890**	**166094**	**200301**	**595966**

Table 1: Besoins de stockage pour les expériences (en Go)

Ainsi, la capacité totale de stockage estimée à 3 ans (pour 4 années d'exploitation) dépasse le demi péta-octet, soit près de 600 disques d'une capacité de 1 To.

La grosse capacité demandée par le LBMC a bien été confirmée.

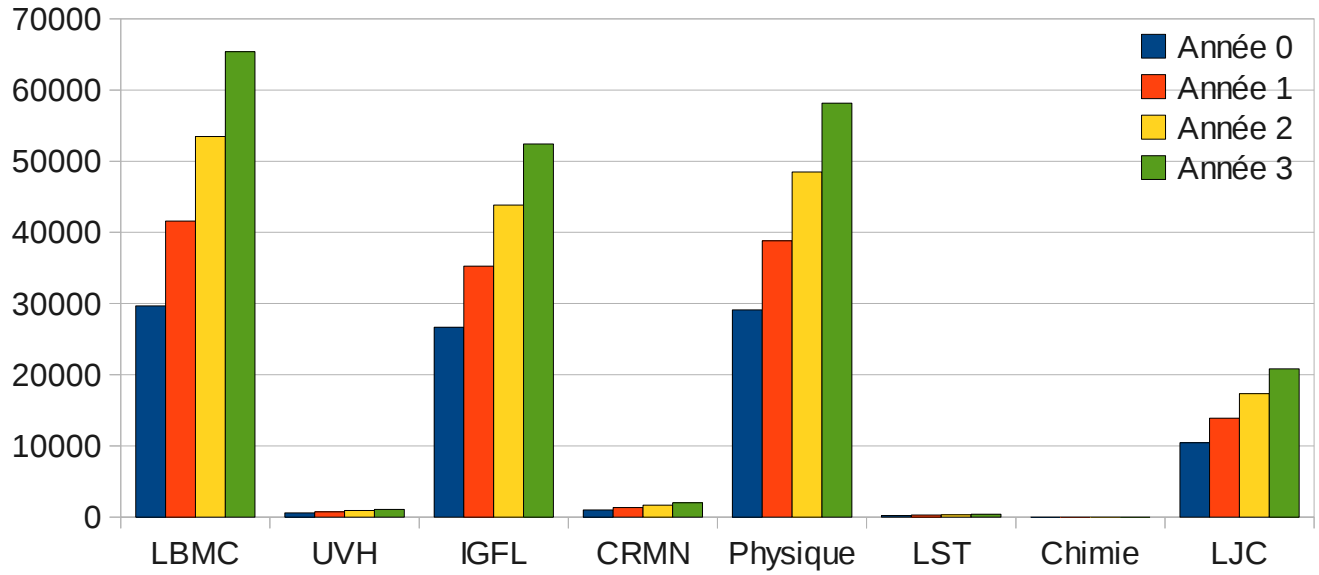


Illustration 21: Besoins des laboratoires sur 4 années (en Go)

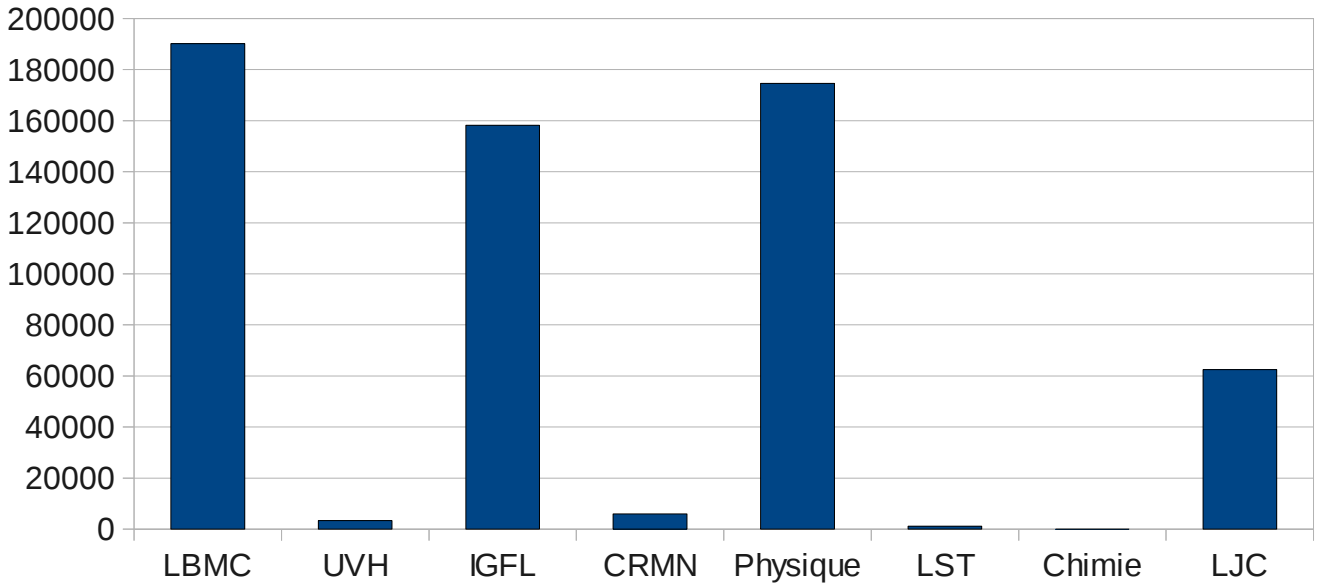


Illustration 22: Besoins cumulés des laboratoires pour 4 années (en Go)

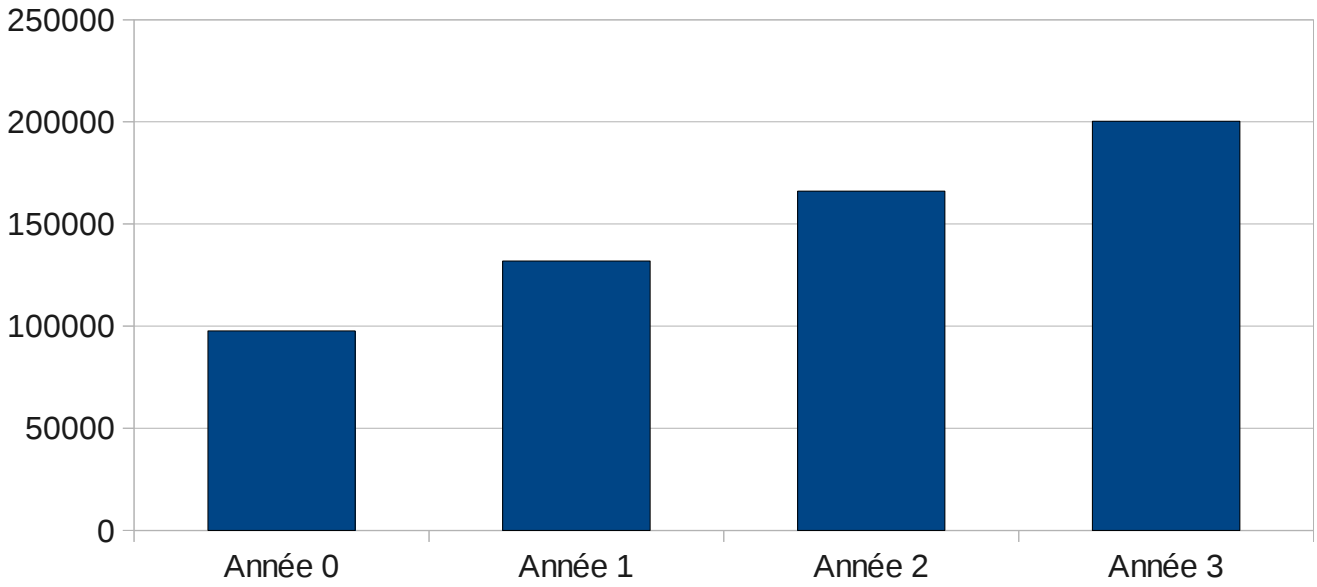


Illustration 23: Expériences : besoins cumulés pour chaque année (en Go)

9.4 Sur les besoins de stockage dans les plates-formes de traitements

Laboratoire	Année 0	Année 1	Année 2	Année 3	4 années
LBMC	50980	93007	134938	176960	**455885**
UVH	2073.25	2755	3443	4124	**12395.25**
IGFL	21443	29378	37203	45136	**133160**
CRMN	375	500	625	750	**2250**
LIP	10150	23750	37150	50750	**121800**
Chimie	2500	2500	2500	2500	**10000**
LJC	2560.75	3395	4238	5070	**15263.75**
Total	**90082**	**155285**	**220097**	**285290**	**750754**

Table 2: Besoins de stockage pour les traitements (en Go)

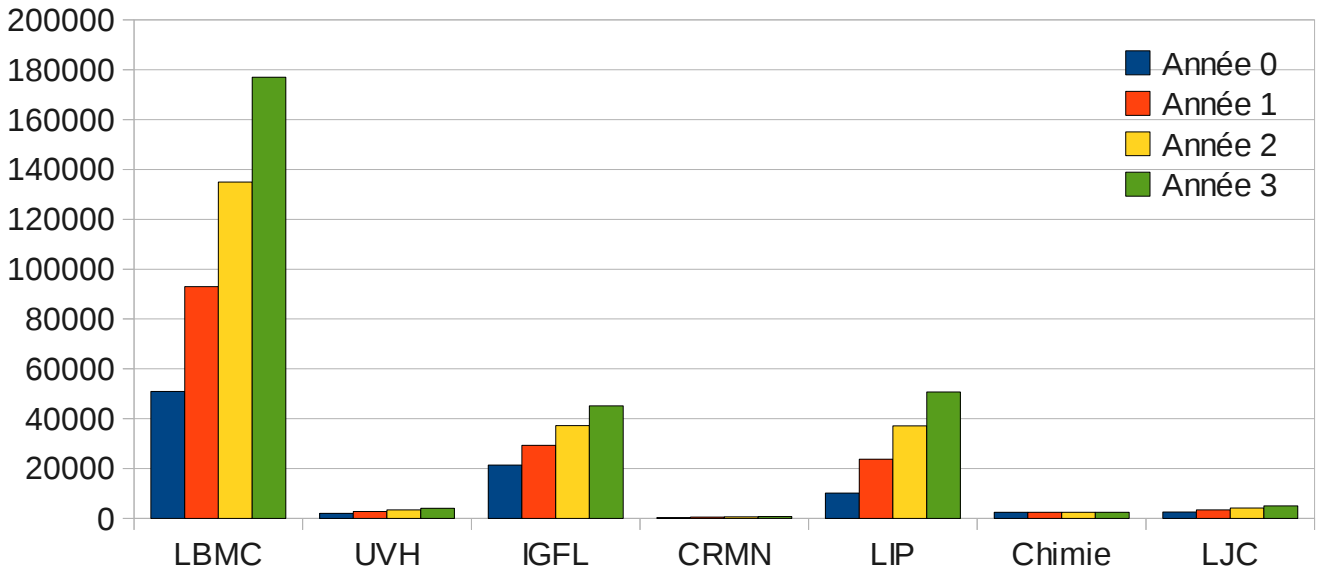


Illustration 24: Traitements : besoins des laboratoires sur 4 années (en Go)

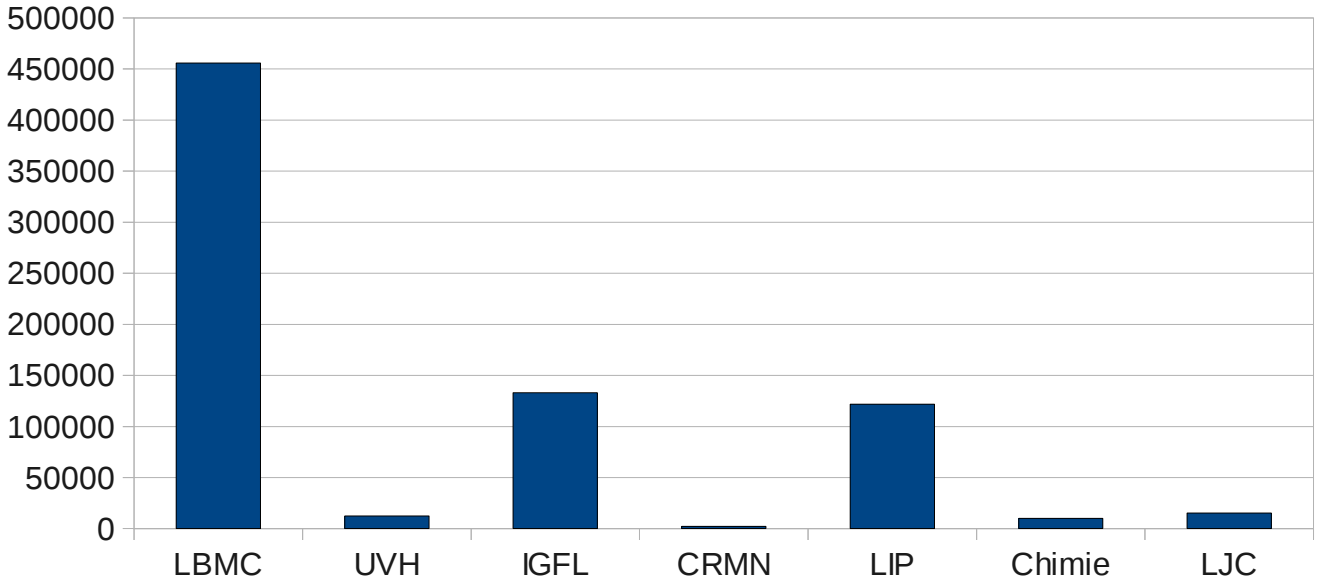


Illustration 25: Traitements : besoins cumulés des laboratoires pour 4 années (en Go)

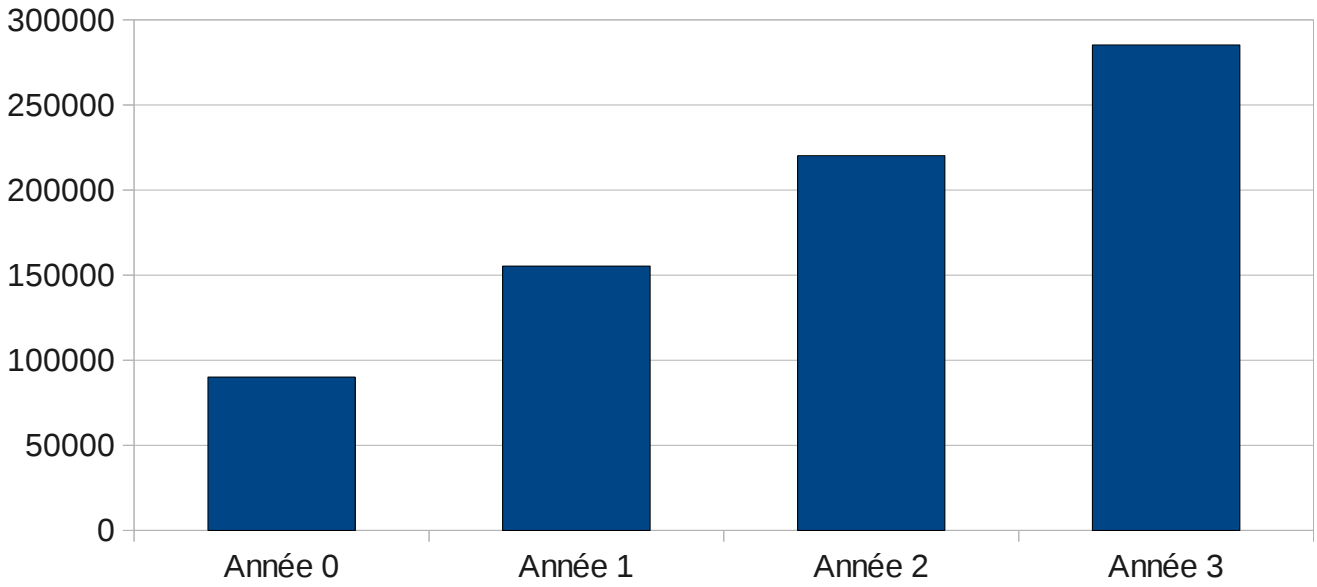


Illustration 26: Traitements : besoins cumulés pour chaque année en Go

Ainsi, la capacité totale de stockage estimée à 3 ans (pour 4 années d'exploitation) approche les trois-quart de péta-octet, soit près de 750 disques d'une capacité de 1 To.

9.5 Sur les besoins de stockage dans les plates-formes d'exploitation

Le nombre de questionnaire validés, inférieur à la dizaine, rend l'exploitation difficile des résultats.

Toutefois, si les données présentées se généralisaient, les volumes engagés seraient infiniment inférieurs aux volumes présentés ci-dessus.

9.6 Sur les besoins de stockage pour la valorisation

Dans le schéma synoptique 1 page 11, il avait été précisé l'action de « valorisation », cœur de métier du chercheur. Cette action de valorisation exige un volume conséquent, se basant sur toute l'historique numérique que le chercheur utilise au quotidien (archives de messagerie électronique, articles en préparation ou parus, etc...)

Certains laboratoires ont estimé cet espace « personnel » de plusieurs Go à quelques dizaines de Go. D'autres offrent ce service là à hauteur de 20 Go.

Il est donc raisonnable d'estimer le volume nécessaire en se basant sur le nombre de personnes du laboratoire et ce montant de 20 Go.

L'estimation du nombre de personnes dans les laboratoires a été menée par deux canaux : via l'annuaire directement connecté sur le système d'information global de l'établissement (par les extractions permettant de générer un annuaire imprimable) et via SIRE, le Système d'Information de la Recherche. Les estimations étaient comparables, à quelques unités près (avec des erreurs de 10% à 20%). Pour ne pas risquer de sous estimer le nombre de personnes dans les laboratoires, la population maximale pour chaque laboratoire issue des deux sources a été utilisée.

Il a été également choisi arbitrairement un taux de croissance de ces données de 2 sur 3 années.

Laboratoire	Année 0	Année 1	Année 2	Année 3	4 années
LBMC	2300	3066.67	3833.33	4600	13800
UVH	1360	1813.33	2266.67	2720	8160
IGFL	1780	2373.33	2966.67	3560	10680
RDP	1260	1680	2100	2520	7560
CRMN	560	746.67	933.33	1120	3360
Physique	2040	2720	3400	4080	12240
LST	680	906.67	1133.33	1360	4080
LIP	1720	2293.33	2866.67	3440	10320
Chimie	1560	2080	2600	3120	9360
LJC	680	906.67	1133.33	1360	4080
CRAL	360	480	600	720	2160
UMPA	1020	1360	1700	2040	6120
IXXI	680	906.67	1133.33	1360	4080
Total	16000	21333.33	26666.67	32000	96000

Table 3: Besoins de stockage courant pour les laboratoires (en Go)

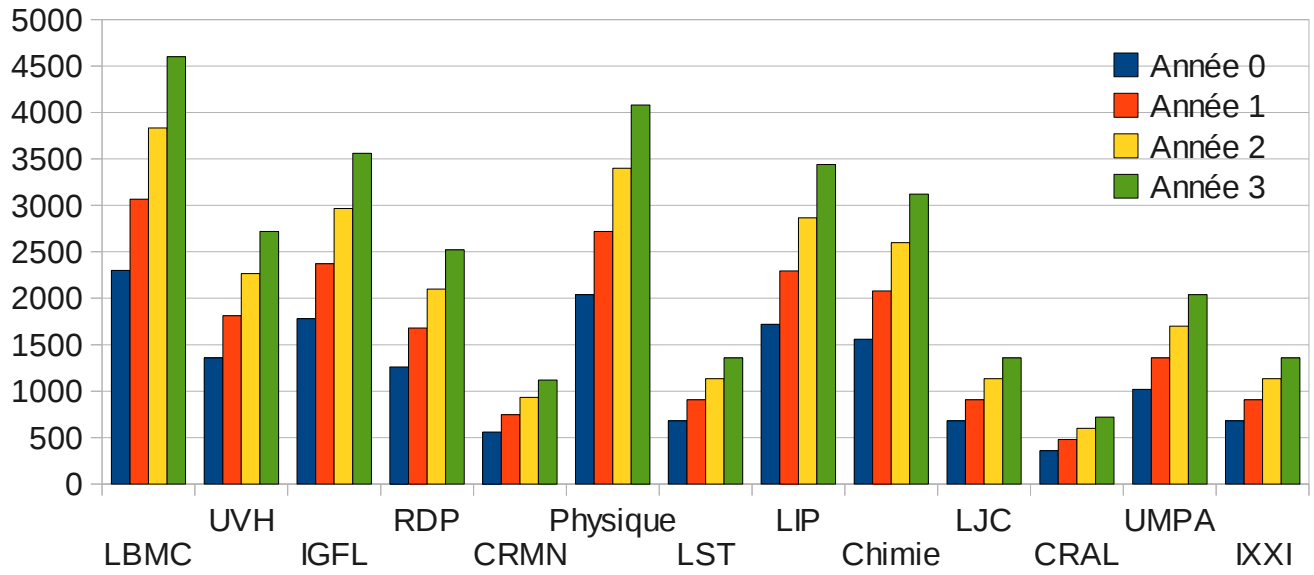


Illustration 27: Stockage courant dans les laboratoires pour 4 années (en Go)

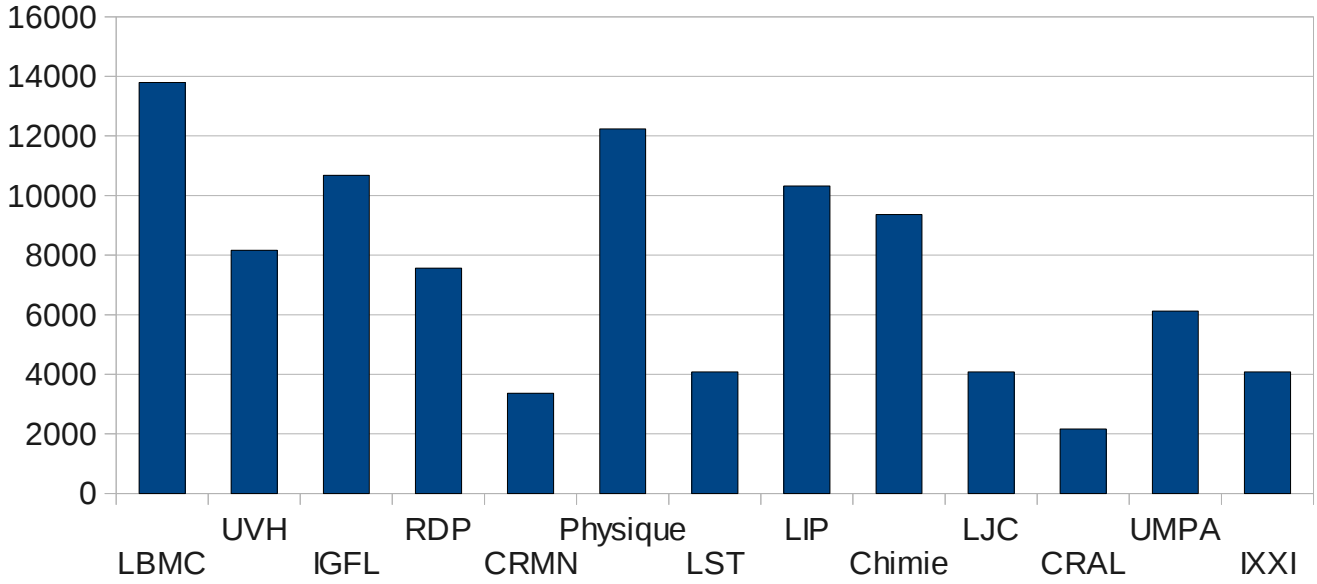


Illustration 28: Stockage courant cumulé pour les laboratoires pour 4 années (en Go)

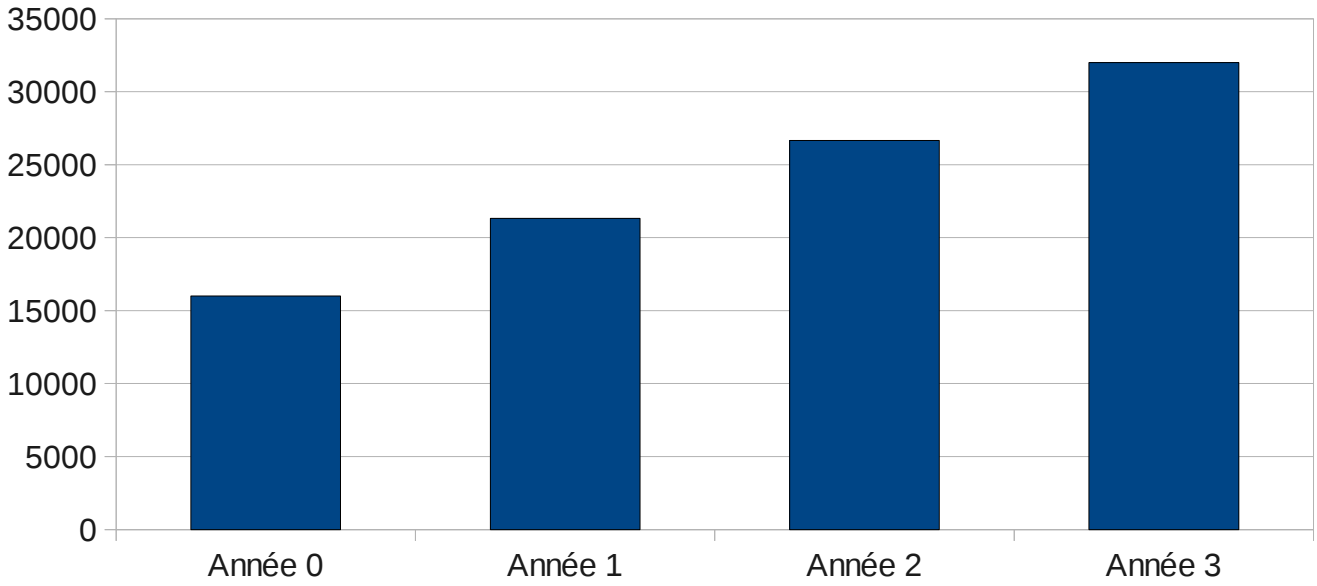


Illustration 29: Stockage courant cumulé par année (en Go)

10 Cumul de tous les besoins en stockage

Laboratoire	Année 0	Année 1	Année 2	Année 3	4 années
LBMC	82935.5	137645.67	192256.33	246961	659798.5
UVH	4017.75	5321.33	6628.67	7930	23897.75
IGFL	49898	67009.33	84011.67	101121	302040
RDP	1260	1680	2100	2520	7560
CRMN	1935	2579.67	3225.33	3870	11610
Physique	31165	41528	51892	62255	186840
LST	880	1173.67	1466.33	1760	5280
LIP	11870	26043.33	40016.67	54190	132120
Chimie	4062.5	4583	5103	5623	19371.5
LJC	13679.25	18197.67	22724.33	27241	81842.25
CRAL	360	480	600	720	2160
UMPA	1020	1360	1700	2040	6120
IXXI	680	906.67	1133.33	1360	4080
Total	**203763**	**308508.33**	**412857.67**	**517591**	**1442720**

Table 4: Besoins en stockage global pour les laboratoires, pour 4 années (en Go)

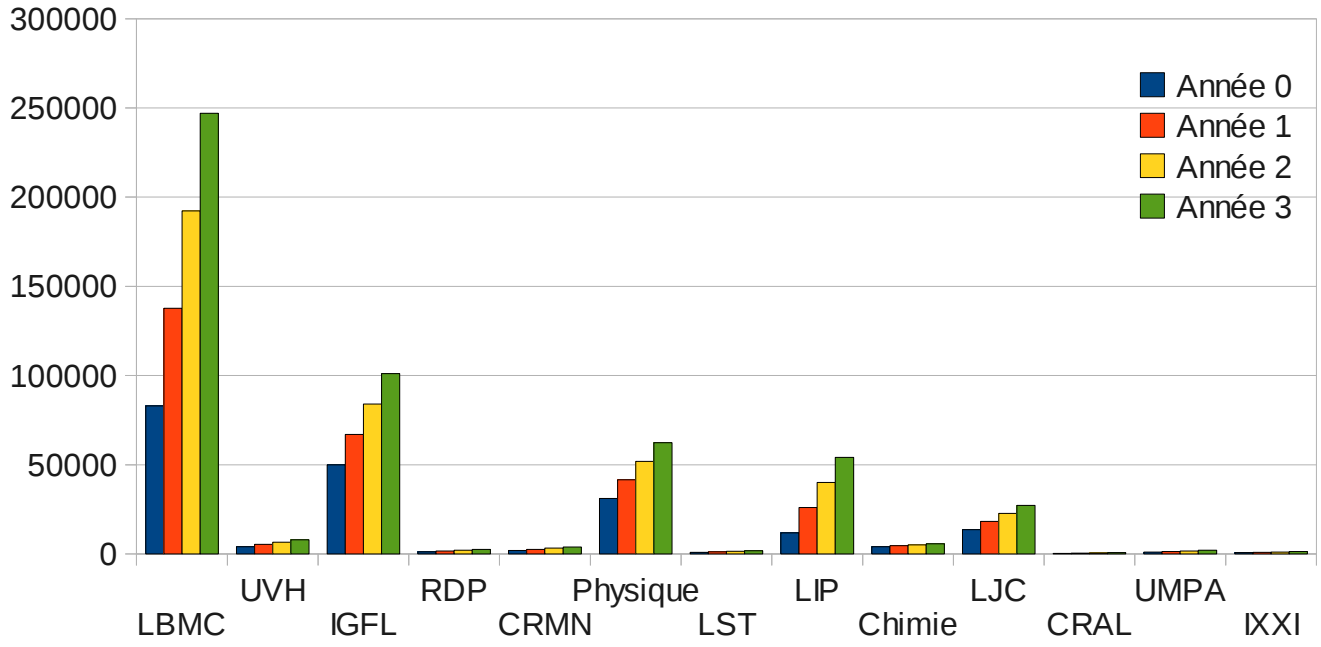


Illustration 30: Besoins en stockage global par laboratoire, pour 4 années (en Go)

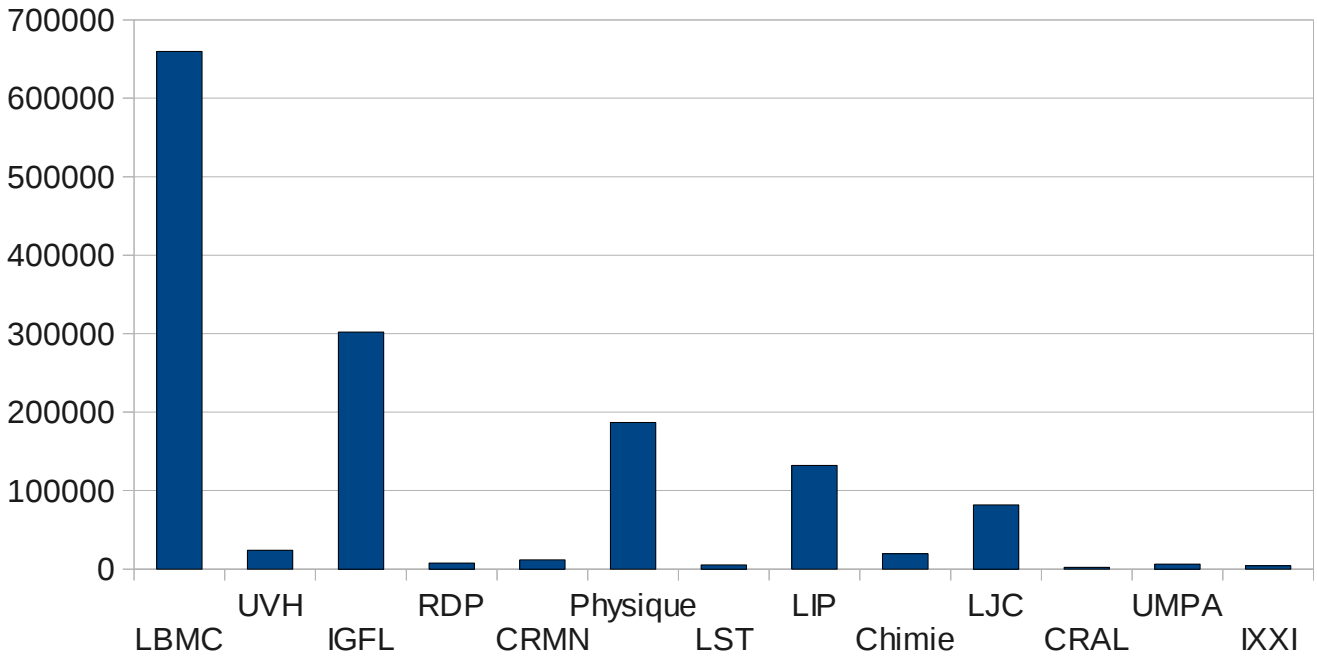


Illustration 31: Besoins cumulés en stockage global par laboratoire pour 4 années (en Go)

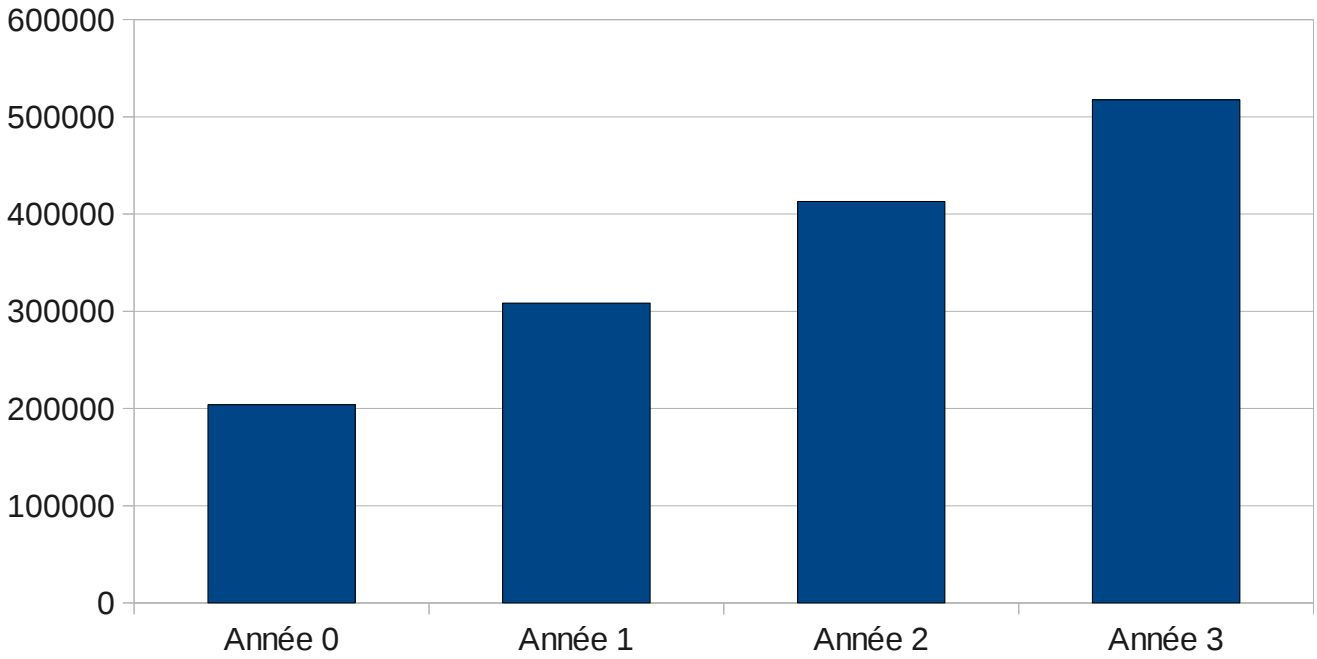


Illustration 32: Besoins en stockage global, par année (en Go)

11 Des spécifications fonctionnelles aux spécifications techniques

11.1 Vers la clôture du triptyque de l'étude

Cette étude a été menée suivant le triptyque traditionnel de gestion de projet : « où en est-on ? », « où va-t-on ? », « comment y-va-t-on ? ».

Les deux premières questions ont été exprimées par les deux enquêtes sous forme de questionnaires et leur analyse. De plus, une sommaire modélisation, à partir de conversations et de courriels, a permis d'établir les besoins courants de stockage informatique pour l'activité quotidienne des personnes des laboratoires.

Il convient maintenant d'agréger les besoins, tant qualitatifs que quantitatifs, pour en sortir une liste de spécifications fonctionnelles destinées à être, ensuite, déclinées en spécifications techniques purement informatiques.

11.2 Spécifications fonctionnelles

11.2.1 Pour le stockage

Pour le « front office » (le « salon », une affaire d'utilisateurs et leur appropriation des outils) :

- une gestion fine de l'accès aux données :
  - droits des utilisateurs pour les accès en écriture à partir des postes de manipulation,
  - droits des utilisateurs pour les accès en lecture à partir des postes de traitement,
  - droits pour les responsables de ces utilisateurs d'accéder à leurs données après leur départ ;
- une accessibilité des données dépassant le cadre du laboratoire :
  - espace accessible de l'extérieur de manière simple et sécurisée ;
- une indexation indispensable des expériences et des traitements pour assurer la pérennité des données ;
- une mise en place de plateformes de traitements dédiées et partagées entre les utilisateurs ;
- une abstraction des volumes de stockage, la plus transparente possible des « boîtes noires » utilisées ;
- une mise à disposition rapide pour faire face aux demandes urgentes et limiter le risque de perte par casse ;

Pour le « back-office » (la « cuisine », une affaire d'informaticiens et la mise à disposition des outils) :

- une amélioration des conditions de transfert des données entre plateformes (de l'expérience à son traitement) notamment par la généralisation d'une connectivité haut débit (GE minimum) ;
- une disponibilité accrue des dispositifs de stockage (pour les expériences durant le plus longtemps) ;
- une souscription la plus large possible du contrat de maintenance ;
- des procédures simplifiées pour la mise à disposition ou l'extension d'un volume de stockage ;
- des procédures simplifiées pour la restauration d'un volume de stockage ;
- une « scalabilité » de la solution de stockage pour l'étendre chaque année en fonction des besoins.

11.2.2 Pour la sauvegarde

Cette opération est purement informatique : elle vise à sauvegarder les données (et permettre leur restauration) en cas de destruction d'un élément (ou plusieurs) du stockage.

Dans le cas idéal, elle exige :

- une séparation physique du stockage primaire ;
- une représentation la plus synchrone possible des données originelles.

11.2.3 Pour l'archivage

Cette opération, pour être la plus efficace possible, se base sur les informations mentionnées par les utilisateurs dans l'indexation de leurs données.

Deux approches sont possibles :

- un archivage basé sur le stockage originel ;
- un archivage basé sur la sauvegarde.

Dans les deux cas, les archives peuvent prendre la forme :

- d'une série d'instantanés pris suivant une politique pré-établie : elle est alors intégrée à la solution de stockage ;
- une copie complète sur un support tierce, archivée physiquement : elle utilise généralement une technologie à base de supports à accès séquentiels, comme les bandes.

11.3 Éléments de spécifications techniques

11.3.1 Proposition : une solution technique centralisée, mais distribuée

Étant donné les volumes, leur gestion, la gestion de leur accès, une solution de stockage centralisée s'impose d'elle même : elle demeure la seule à pouvoir s'adapter de la manière la plus flexible aux demandes des laboratoires.

Une solution de stockage centralisée existe déjà au Service Informatique : sur elle s'est déjà greffé un certain nombre d'espace critique comme la messagerie électronique, des sauvegardes et des espaces de stockage pour quelques laboratoires.

La solution proposée consiste à utiliser une technologie comparable à celle déjà déployée par le SI et déjà en œuvre pour plusieurs laboratoires (pour des volumes beaucoup plus modestes).

11.3.2 Stockage, sauvegarde et archivage : quelques simulations réalistes

L'expression de besoins de stockage est évaluée à 200 To à l'instant, 500 To en 2011, 1 Po fin 2012 et 1,5 Po cumulés sur 4 années.

Si toutes les données sont sauvegardées (dupliquées à au moins un endroit), cette capacité doit être doublée.

Si l'archivage est réalisé sur les unités de stockage, il est raisonnable de rajouter 20% au montant du stockage.

Quatre simulations de solutions de stockage ont été réalisées :

- la première se basant sur les éléments d'infrastructure déjà présents sur le site
- la seconde utilisant la base des équipements précédents mais mis à jour (disques durs étendus de 1 à 2 To)
- la troisième issue d'un élément de NAS proposé par Transtec : un Transtec 4300L NAS4324L-A
- la quatrième à partir d'une proposition de NAS proposé par RackServers.com : un RS4-5450-B

Les deux dernières simulations se basent sur des équipements « nus », donc dépourvus de logiciels spécifiques d'infrastructure de stockage. Ces fonctionnalités peuvent être rajoutées à l'aide de composants tierces, Open Source ou propriétaires. Ces machines, qui sont plus des serveurs gavés de disques durs, ont vus leur prix estimés suivant les mêmes spécifications (mémoire vive, processeur, interface réseau très haut débit, maintenance 3 ans avec une GTI 4 heures).

11.3.3 Présentation de la simulation : des écarts significatifs

A partir des spécifications techniques des équipements, des besoins en stockage, les estimations suivantes sont réalisées :

- l'occupation en terme de surface (par le nombre de baies) ;
- la puissance électrique nécessaire (par le nombre de KVA consommés par les équipements) ;
- les besoins en climatisation (par le nombre de BTU spécifiés ou estimés par le constructeur ;

- le coût (par les montants présentés ou estimés des équipements à partir d'une sommaire simulation en ligne).

Equipements	Dell Sumo 1T	Dell Sumo 2T	Transtec 4300L	RackServer 5450
Année 0				
Stockage (en Go)	512271.33	512271.33	512271.33	512271.33
Archivage Stockage (en Go)	102454.27	102454.27	102454.27	102454.27
Total Stockage (en Go)	614725.6	614725.6	614725.6	614725.6
Nombre unités	6	3	6	3
Nombre baies 42U nécessaires	1	1	1	1
Consommation totale en KVA	7.2	3.6	5.4	6.54
Besoins Climatisation KBTU/h	21	10.5	18.6	22.2
Prix en k€ (unités disque)	360	270	96	72
Pour les quatre années				
Stockage (en Go)	1442720	1442720	1442720	1442720
Archivage Stockage (en Go)	288544	288544	288544	288544
Total Stockage (en Go)	1731264	1731264	1731264	1731264
Nombre d'unités	40	20	40	20
Nombre baies 42U nécessaires	5	3	5	3
Consommation totale en KVA	48	24	36	43.6
Besoins Climatisation KBTU/h	140	70	124	148
Prix en k€ (unités disque)	2400	1800	640	480

Table 5: Estimation des coûts, stockage uniquement (année 0 et 4 années cumulées)

Les montants présentés dans le tableau 5, page 40 ne comprennent que le stockage et l'archivage local au stockage.

La différence sur le nombre de baies vient essentiellement de l'intégration des disques durs sur certaines baies : elle varie du simple au double pour une même occupation.

La consommation électrique et les besoins en climatisation varie également suivant les mêmes règles. Seule la future baie Sumo Dell équipée de disques de 2 To se situe en dessous.

Le principal écart se situe sur le coût : l'utilisation d'équipements «spécialisés» comme les Dell Sumo présente un budget très supérieur à celui que nous pouvons envisager en utilisant des équipements, certes « génériques », mais demandant un travail d'intégration logiciel non négligeable. Le principal écueil viendra de l'agrégation de volumes sur des serveurs différents : les solutions existent, notamment issues du monde du calcul scientifique.

De plus, si nous considérons une sauvegarde équivalente au stockage, tous les montants précédents doivent simplement être doublés. Le tableau 6, page 41 présente, pour les quatre années d'exploitations, les investissements qu'il sera nécessaire de faire pour adapter l'infrastructure à la demande.

Investissements (en k€)	Dell Sumo 1T	Dell Sumo 2T	Transtec 4300L	RackServer 5450
Année 0	720	540	192	144
Année 1	960	720	256	192
Année 2	1440	1080	384	288
Année 3	1680	1260	448	336

Table 6: Investissements annuels pour les unités de stockage ET de sauvegarde identiques

12 Retour à la lettre de mission

La lettre de mission du 20 décembre 2009 demandait de mener une étude sur les moyens de stockage du site Monod en général et des laboratoires de biologie en répondant aux questions suivantes :

- les besoins en sauvegarde et de stockage des laboratoires de Biologie
- les besoins en sauvegarde et de stockage des autres laboratoires de l'école sur le site Jacques Monod
- les conséquences, en terme de « froid », que ces besoins vont créer ;
- les conséquences, en terme de locaux, que ces besoins vont créer.

12.1 Besoins de stockage/sauvegarde/archivage des laboratoires de biologie

Sur la première question, les besoins de stockage et de sauvegarde ont été estimés par les laboratoires de biologie, à l'exception du laboratoire de Reproduction et Développement des Plantes.

Leurs demandes ont été analysées et confirmées, à l'exception de 2 d'entre elles. Pour ne pas les omettre, elles ont tout de même été intégrées sur la base des besoins maximum exprimés dans leur laboratoire respectif.

Le besoin en sauvegarde n'est pas apparu directement dans les enquêtes, mais a été estimé compte-tenu de la durée de conservation des données généralement supérieur à l'année.

Pour les 4 laboratoires de biologie (LBMC, IGFL, UVH, RDP), l'expression de leurs besoins en stockage se chiffre à **138 To** pour la première année. Sur 4 années, leurs besoins s'estime à **993 To, soit près de 1 Po.**

Si est admise l'idée d'un archivage intégré au stockage et à la sauvegarde, et une sauvegarde identique au stockage, les besoins passent à **276 To pour la première année** et près de **2 Po pour 4 années d'exploitation**.

12.2 Besoins de stockage/sauvegarde/archivage des autres laboratoires

Sur la seconde question, la majorité des autres laboratoires du site Monod a répondu à l'enquête, de manière très hétérogène. Des laboratoires ne se sont cependant pas exprimés : l'UMPA et le CRAL.

D'autres se sont exprimés de manière très timide, comme le laboratoire de chimie et le LST.

Le LJC s'est particulièrement illustré en exprimant largement ses besoins : la motivation de son informaticien de laboratoire a certainement été la clé de cette mobilisation.

Pour les 8 autres laboratoires du site Monod, l'expression de leurs besoins en stockage se chiffre à 66 To pour la première année. Sur 4 années, cela atteint les 450 To.

Si est admise l'idée d'un archivage intégré au stockage et à la sauvegarde, et une sauvegarde identique au stockage, les besoins passent à **132 To pour la première année** et près de **900 To pour 4 années d'exploitation**.

Ces besoins sont inférieurs à ceux exprimés par les laboratoires de biologie.

12.3 Les conséquences en terme de « froid » que ces besoins vont créer

Sur la troisième question, à partir :

- des besoins de stockage ;

- de la sauvegarde (identique aux besoins de stockage) ;
- de la structure d'archivage (équivalente à 20% du stockage et de la sauvegarde) ;

et de quatre équipements informatiques remplissant ces fonctions de stockage, il a été estimé, au maximum :

- une consommation électrique : **de 14,4 kVA pour la première année à 96 kVA pour les 4 cumulées ;**
- les besoins en climatisation : **de 44,4 kBTU/h la première année à 296 kBTU/h pour les 4 cumulées.**

12.4 Les conséquences en terme d'espace que ces besoins vont créer

Quant à la quatrième question, elle a été estimé en évaluant le nombre de baies nécessaires pour accueillir tous les équipements de stockage nécessaires.

S'il est considéré que la sauvegarde est identique au stockage et que l'archivage s'y intègre à hauteur de 20% supplémentaire, une baie serait suffisante la première année pour le stockage et l'archivage, il en faudrait entre 3 et 5 selon la solution retenue sur 4 années cumulées. Pour la partie sauvegarde, il faudrait une baie la première année et de 2 à 4 supplémentaires sur 4 années.

A noter que les estimations du nombre de baies l'ont été en estimant un intégration maximale des équipements (les baies sont complètement remplies). Il est indispensable qu'un équipement de refroidissement équivalent à celui utilisé pour un cluster de calcul soit utilisé dans ce cas. Pour finir, la résistance structurelle de la baie risque d'être largement sollicitée

13 Conclusion

L'expression des besoins a donc été menée pour les laboratoires de l'Ecole Normale Supérieure de Lyon, site Monod.

L'analyse des besoins, actuels et sur une projection de 3 années, a permis de dégager un volume de stockage nécessaire.

Il convient maintenant, à partir de ces volumes, de s'entendre sur une stratégie de sauvegarde et d'archivage pour ces volumes très conséquents. A titre d'exemple, une stratégie a été proposée pour simuler, de la manière la plus fidèle possible, les besoins en climatisation et en espace.

Une fois cette stratégie établie, l'ensemble des spécifications fonctionnelles et les quelques spécifications techniques serviront à compléter le cahier des charges dans l'achat des équipements nécessaires.

Pour finir, cette étude a également permis de mettre en lumière plusieurs aspects : liés à la sécurité informatique et la conservation du patrimoine scientifique d'un côté et illustrant les besoins connexes en informatique des laboratoires, notamment dans la mise à disposition de compétences informatique pour la valorisation des travaux scientifiques.

14 Annexes

14.1 Questionnaire « stockage pour les plates-formes expérimentales »

1. Quel est votre nom ?
2. Quel est le laboratoire destinataire des expériences ?
3. Quel est le nom de la plate-forme expérimentale ?
4. Quel est le type d'équipement d'acquisition (caméra, ...) ?
5. Quel est le type de données en sortie de l'équipement (image, vidéo, ...) ?
6. Qui réalisent les expériences ?
    1. Stagiaire
    2. Étudiant
    3. Doctorant
    4. Chercheur
    5. Enseignant-Chercheur
    6. Personnel Technique
    7. Prestataire externe
    8. Autre :
7. Qui exploitent des données brutes de la plate-forme ?
    1. Stagiaire
    2. Étudiant
    3. Doctorant
    4. Chercheur
    5. Enseignant-Chercheur
    6. Personnel Technique
    7. Prestataire externe
    8. Autre :
8. Quelle est la durée moyenne d'utilisation de la plate-forme pour une expérience ?
9. Quel est le nombre d'expériences réalisées par semaine ?
10. Par combien comptez-vous multiplier votre taux d'utilisation de la plate-forme sur les 3 prochaines années ?
11. Combien de temps les données doivent-elles être conservées ?
12. Où se situent les protocoles ou les fichiers de configuration des équipements ?
13. Où se situe l'équipement ?
14. Où sont stockés les résultats numériques des expériences ?
15. Quel est le volume de données moyen généré par une expérience ?
16. Le réseau actuel (100 Mb/s) est-il sous-dimensionné pour transférer les données ailleurs (dépasse le 1/4 d'heure) ?
17. Existe-t-il des contraintes sur l'acquisition ?
18. Quel type de logiciel utilisez-vous pour les acquisitions ?

14.2 Questionnaire « stockage pour les plates-formes de traitement »

1. Quel est votre nom ?
2. Quel est le laboratoire destinataire des traitements ?
3. Quel est le nom de la plate-forme de traitement ?
4. Quel est le type de données en entrée du traitement (image, vidéo, ...) ?
5. Quel est le type d'équipement ?
6. Quel est le type de données en sortie du traitement (image, vidéo, ...) ?
7. Qui réalise l'intégration du traitement
    1. Stagiaire
    2. Étudiant
    3. Doctorant
    4. Chercheur
    5. Enseignant-Chercheur
    6. Personnel Technique
    7. Prestataire externe
    8. Autre :
8. Qui exploitent des données traitées de la plate-forme ?
    1. Stagiaire
    2. Étudiant
    3. Doctorant
    4. Chercheur
    5. Enseignant-Chercheur
    6. Personnel Technique
    7. Prestataire externe
    8. Autre :
9. Quelle est la durée moyenne d'utilisation de la plate-forme pour un traitement ?
10. Quel est le nombre de traitements réalisé par semaine ?
11. Par combien comptez-vous multiplier votre taux d'utilisation de la plate-forme sur les 3 prochaines années ?
12. Combien de temps les données doivent-elles être conservées ?
13. Où se situent les données brutes et le logiciel de traitement ?
14. Où se situe l'équipement ?
15. Où sont stockés les résultats issus du traitement ?
16. Quel est le volume de données moyen généré par une expérience ?
17. Le réseau actuel (100 Mb/s) est-il sous-dimensionné pour transférer les données ailleurs (dépasse le 1/4 d'heure) ?
18. Existe-t-il des contraintes sur l'acquisition ?
19. Quel type de logiciel utilisez-vous pour les traitements ?

14.3 Questionnaire "stockage pour l'exploitation des résultats"

1. Quel est votre nom ?

2. Quel est le laboratoire à l'origine de l'exploitation ?
3. Quel nom donnez-vous à l'exploitation de ces résultats ?
4. Quel est le type de données utilisées ?
5. Quel est le type d'exploitation ?
6. Quelle personne propose l'exploitation ?
    1. Stagiaire
    2. Étudiant
    3. Doctorant
    4. Chercheur
    5. Enseignant-Chercheur
    6. Personnel Technique
    7. Prestataire externe
    8. Autre :
7. A qui est destiné cette exploitation ?
    1. Chercheur
    2. Enseignant
    3. Étudiant
    4. Collégien ou Lycéen
    5. Enseignant du secondaire
    6. Tout public
    7. Autre :
8. Quelle est la durée moyenne d'exploitation ?
9. Combien de publications sont-elles réalisées par semaine ?
10. Par combien comptez-vous multiplier votre taux de publication sur ces 3 prochaines années ?
11. Combien de temps les données doivent-elles être conservées ?
12. Où se situent les données sources ?
13. Quelle est la nature de l'exploitation ?
14. Quel est le volume de données d'une exploitation ?
15. Le réseau actuel (100 Mb/s) est-il sous-dimensionné pour transférer les données ailleurs (dépasse le 1/4 d'heure) ?
16. Quel type de logiciel utilisez-vous pour l'exploitation ?

14.4 Requêtes et commentaires

14.4.1 Plate-forme Preci

Ces éléments sont issus d'une conversation avec Jean Sauboua.

- La plate-forme Preci ne nécessite pas une infrastructure de stockage pour les données qu'elles génèrent
- Les données, d'un volume inférieur à 1Go, concernent l'exploitation de la plate-forme, notamment sur la traçabilité des opérations réalisées dans le passé
- Le besoin s'articule plutôt autour d'un espace collaboratif comprenant un espace de dépôt de document (CMS ou Content Management System) et un agenda pour mentionner l'exploitation de la plate-forme.

- Les espaces des groupes de travail de l'ENT ont tenté d'être utilisé. Cependant, le fait que le responsable du groupe soit un CDD et que son compte soit périodiquement suspendu ont montré les limites de ce système
- Dans le cadre d'une extension possible des expérimentations sur les poissons et la construction, encore à l'étude d'un bâtiment dédié de grande taille, cette nouvelle PF aurait, dans cette hypothèse, besoin d'un système de gestion comparable à Phare
- Le souhait des utilisateurs de la plate-forme est de pouvoir disposer de cet espace collaboratif courant 2010
- Laure Bernard reprend son poste et remplace Jean Sauboua à partir de fin février sur cette thématique d'espace collaboratif

14.4.2 Unité de Virologie Humaine

Ces éléments sont issus d'une conversation avec Didier Nègre.

- Conditions particulières d'une expérience : une des plates-formes expérimentales dispose d'un logiciel ne permettant pas l'intégration d'un antivirus. Pour sécuriser l'accès à cette machine, il a été construit un réseau privé. La passerelle entre ce réseau privé et le réseau du laboratoire permet de récupérer les données de la plate-forme et la transférer, soit sur un poste de travail, soit sur un serveur. 2 manipulations sont donc nécessaires pour passer ses données de la plate-forme expérimentale vers la plate-forme de traitement.
- Contraintes du traitement : le traitement est nécessairement réalisé par un logiciel propriétaire dédié nécessitant l'utilisation d'une clé (un dongle). L'utilisation de ces logiciels est très contraignante (nécessité de disposer d'un dongle cher et fixation de la version du logiciel à une version de système d'exploitation).
- Sauvegarde des postes de travail : le coeur de métier du chercheur exige de disposer d'un espace de plusieurs dizaines de Go pour stocker son travail (cela constitue le point 3, la partie analyse du traitement)
- Nécessité d'une aide technique en informatique : une des missions des chercheurs est la diffusion de l'information scientifique sous forme de publications, revues, congrès, conférences.... . Il y a maintenant un besoin de diffuser les travaux directement sur le site Web du laboratoire, mais cela exigerait une aide d'un informaticien pour la mise en place du site et son maintien en conditions opérationnelles.