Perte de données Eros II

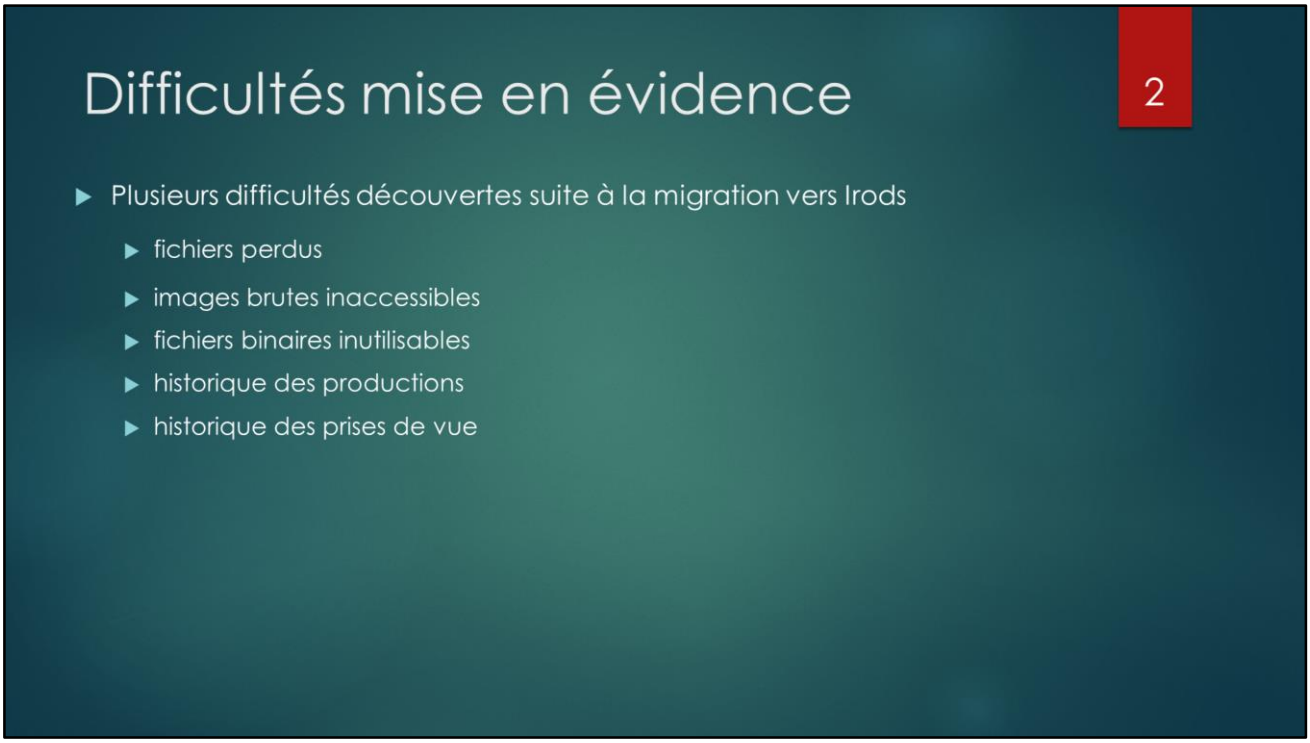
Jean-Noël Albert – LAL / IJClab

Juin 2020

Nous avons vu jusqu'ici la structure des données Eros II et leur organisation dans Irods.

Mais en 20 ans d'existence, ces données ont connu bien des vicissitudes et il y a eu de la casse.

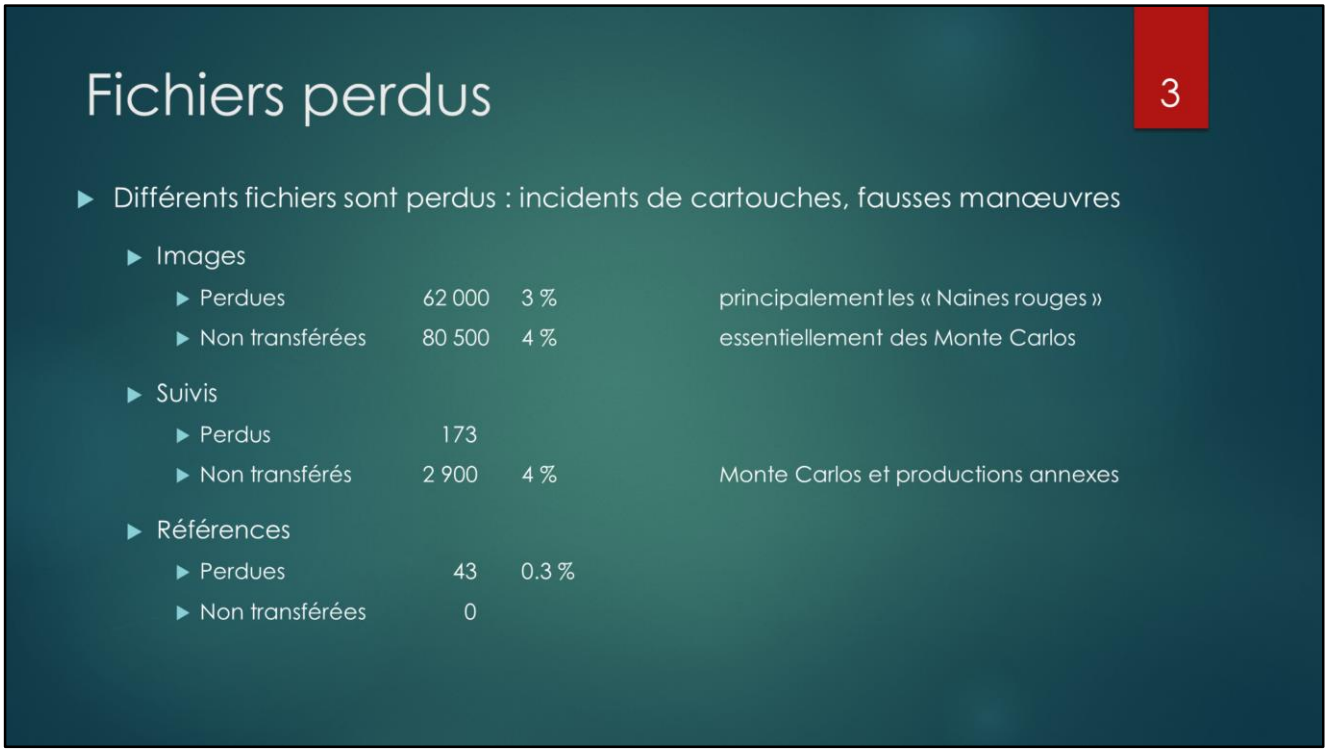
Dans cette présentation, nous allons aborder les sujets qui fâchent, à savoir les fichiers perdus ou inutilisables.



La migration des données vers Irods a été l'occasion de procéder à de profondes vérifications, ne serait-ce que pour s'assurer du succès des transferts. Ces vérifications ont mis en lumière plusieurs problèmes, parfois forts anciens.

Les principaux soucis concernent la perte de fichiers et l'impossibilité d'accéder aux fichiers binaires des catalogues d'étoiles et des mesures.

Et dans une moindre mesure, l'historique de l'expérience.



Après la mise en service du HPPS, une série d'incidents à entrainer la casse de cartouches. Certains fichiers ont pu être récupérés, d'autres non.

Il est également possible que des erreurs soient survenues durant les différentes réorganisations.

Toujours est-il qu'il manque des fichiers.

Il y a 2.2 millions de fichiers référencés dans la base de données. Il n'y a que 1.97 millions de fichiers dans Irods. Cela représente 230 mille fichiers absents après transfert.

MAIS une partie de ces fichiers n'ont pas été transférés parce qu'il ne semblait pas présenter d'intérêt – ceci inclue les images des programmes de simulation et les suivis de ces mêmes simulations ainsi que des productions autres que la production officielle P5.

Soit au total « seulement » 145 mille fichiers effectivement perdus… Ce qui fait tout de même 6 % des données.

Lorsqu'on étudie la base de données, on trouve les références des fichiers enregistrés sur 3490 ou autres cartouches, et les fichiers réorganisés dans HPSS par répertoire et *« cost »*. Et il manque *déjà* des éléments – donc des disparitions avaient déjà eu lieu *avant* la mise en containers Tars. On ne peut donc pas réellement savoir où les pertes se sont produites.

Cependant, les multiples pointages réalisés après la migration montrent que tous les fichiers valides des archives Tar sont dans Irods, quelques dizaines de fichiers de ces archives ayant des tailles incorrectes.

Alors que la plupart des pertes de fichiers pourraient s'expliquer par des casses ou des erreurs de transfert, de telles explications ne semblent pas tenir pour ce qui concerne le programme « Naines rouges » dont ***toutes*** les images ont disparu. Et on sait que ces images ont existé car elles sont toujours référencées dans la base de données.

Ceci représente tout de même 60 000 images pour 1/2 To, soit près de 3 % des 2 millions d'images Eros II.

D'un autre côté, il semble que ce programme n'ait jamais été analysé : il n'existe ni images composées, ni références, ni suivis – Faible consolation…

Il n'y a pas d'explication à ce stade.

Les images brutes, directement issues des CCD, ont été transférées à Saclay, *via* des DLT, alors que les images réduites, résultant de la calibration des images brutes, et sur lesquelles les mesures sont réalisées, allaient au Centre de calcul de l'IN2P3. L'idée initiale était peut-être d'éviter de regrouper toutes les ressources de l'expérience sur un seul site.

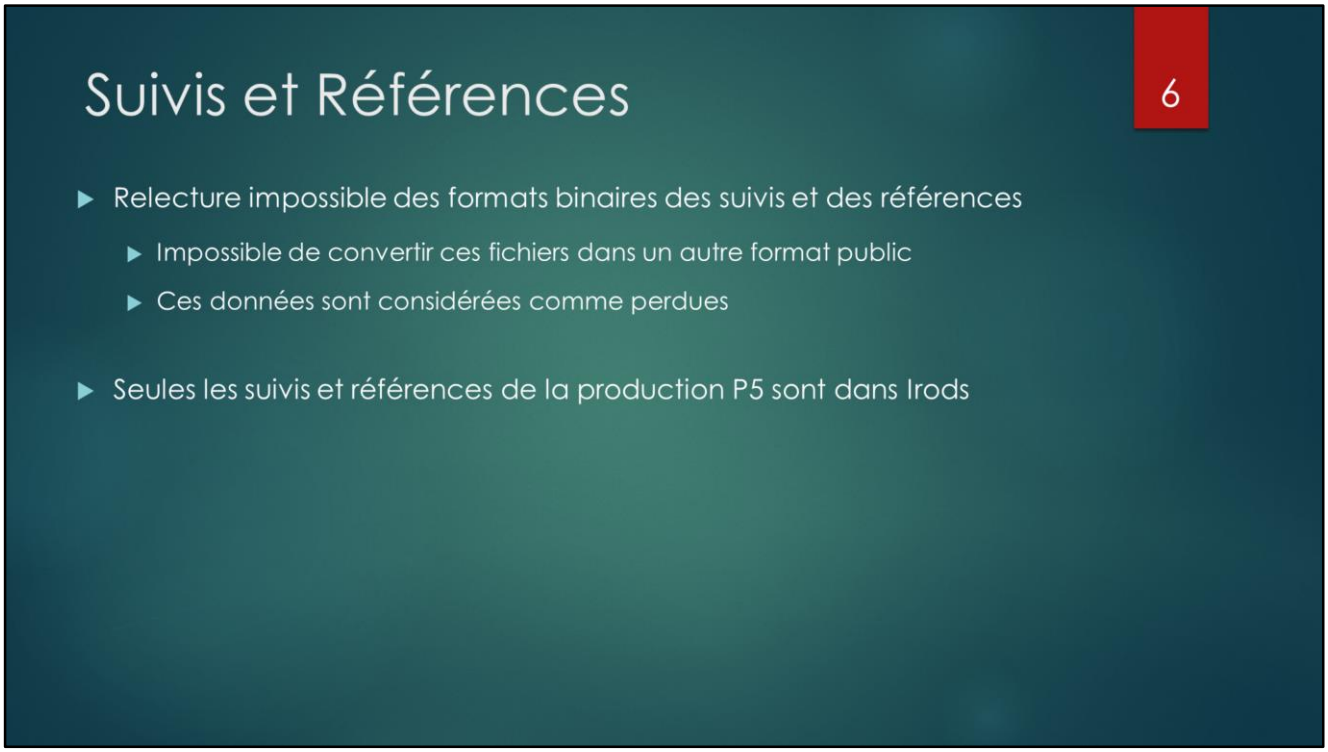
A Lyon, les images ont été recopiées de média en média, au fur et à mesure des évolutions technologiques, ce qui a permis de les préserver jusqu'à aujourd'hui.

Il ne semble pas que ce soit le cas des images brutes. J'ignore même si les DLT ont été préservées, et si c'est le cas, dans quelle condition elles auraient été conservées. Mais il ne semble guère y avoir les moyens, encore moins la volonté, de les sauver et de les ramener au Centre.

A ce stade, ces données sont donc malheureusement considérées comme perdues.

Il n'y a pas de moyens pour connaitre avec précision le nombre et la nature des fichiers perdus, mais on peut supposer que chaque image réduite arrivée à Lyon avait son pendant en termes d'image brute, et que chaque DLT réduite enregistrée à Lyon correspondait à une DLT brute.

Sur cette base, on peut estimer à 1 675 le nombre de DLT brutes à Saclay, soit 1.9 millions d'images brutes, pour environ 15 To. Ceci représente donc la perte de 45 % de l'ensemble des données Eros II.



Les fichiers de suivis et de références sont dans un format binaire propre à l'expérience.

La (ou les ?) librairie permettant de les accéder pour les convertir *"as it"* dans un autre format n'est pas accessible, et le format interne n'étant pas publié, il est impossible d'envisager ne serait-ce qu'une simple transcription – contrairement à ce qui peut être fait pour des formats grands publics, comme Jpeg, Wave, MP3, …

En l'état, les suivis et références peuvent être classés comme perdus. Ceci représente tout de même 70 mille suivis et 15 mille références pour 7 To.

Par ailleurs, à ce stade, seuls les fichiers de la production P5 ont été sauvés dans Irods.

La migration des catalogues ASCII et des courbes de lumière a été faite à partir des archives trouvées dans le catalogue HPPS. Mais comme aucune de ces données n'était répertoriée dans la base de données, il est impossible de réellement contrôler la migration. Le seul point vérifiable est que l'ensemble des archives Tar et leur contenu a bien été recopié dans Irods.

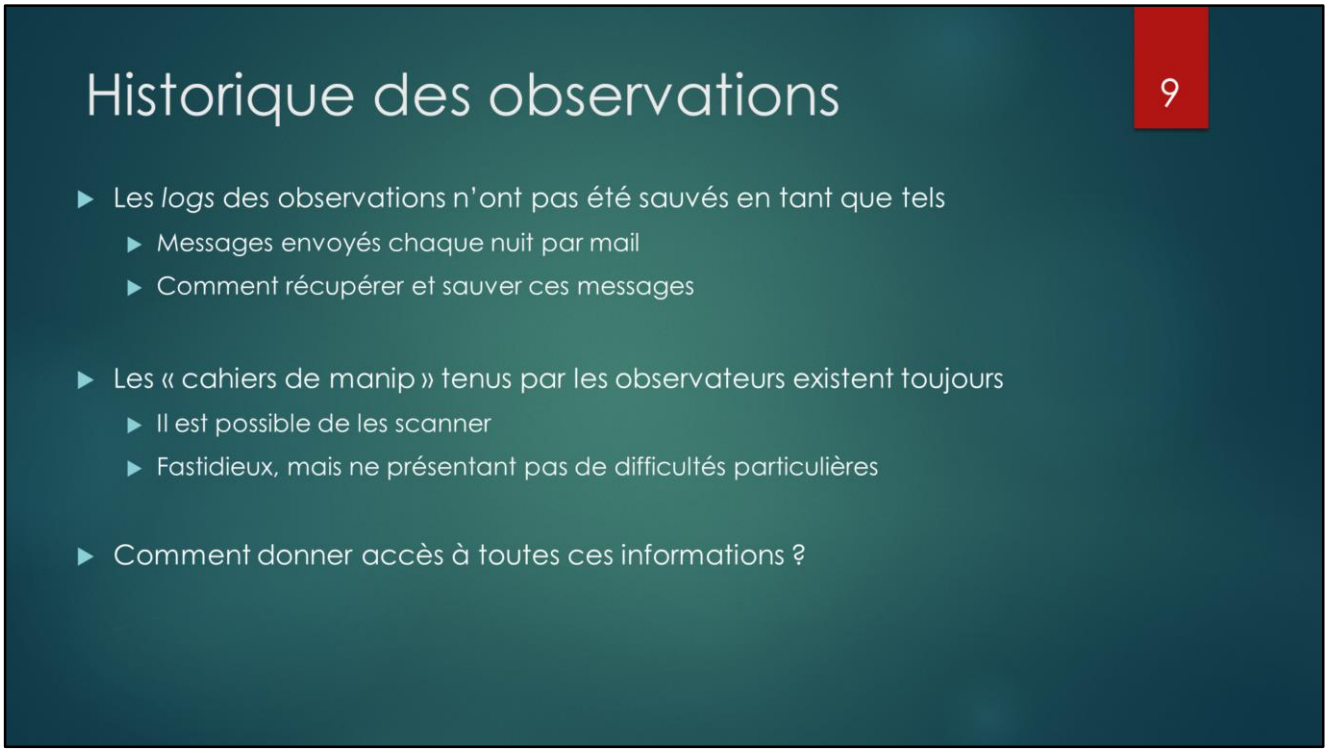
Dernière minute : en écrivant ces notes, il semble possible de procéder à quelques vérifications plus poussées :

- *contrôler que toutes les courbes de lumière trouvées sont référencées dans des catalogues*
  - *ceci vérifiera que les catalogues ont un risque mineur d'avoir été corrompus*
- *contrôler que toutes les courbes de lumière référencées dans les catalogues existent*
  - *ceci vérifiera qu'aucune courbe de lumière n'a été perdue*
- *procéder aux mêmes vérifications croisées pour les fichiers "fields" et les catalogues*
- *vérifier que tous les CCD et quarts de CCD référencés dans la base de données pour les (7) programmes traités disposent bien de fichiers "fields" et "catalogues" complets.*
  - *ceci permettra d'identifier des champs, des CCD ou des quarts de CCD non traités ou perdus.*

Les *logs* des productions n'ont pas été archivés – ou en tout cas ne sont plus visibles – à l'exception peut-être de la partie LMC de la production (P5 ?) archivée par Patrick Tisserand dans son espace HPSS. Il est peut-être possible de récupérer et de sauver ces logs.

Mais s'agit-il d'une production « officielle » ? Il existe en effet des différences entre les chiffres présentés dans la thèse et ceux de la base de données. Dans ces conditions, faut-il conserver ces logs dans l'espace officiel ?

⇒ *A (re)vérifier…*

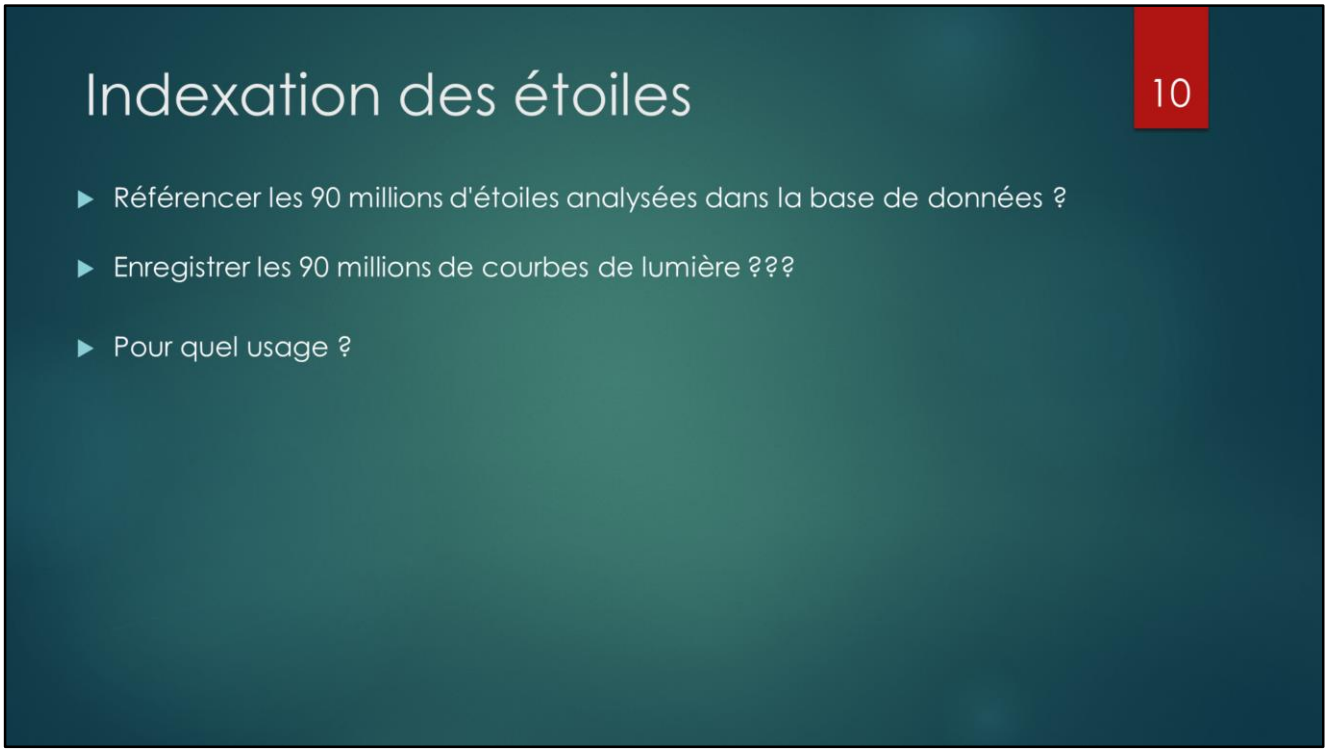


Les *logs* des observations ne semblent pas avoir été explicitement conservés.

Des messages étaient envoyés chaque nuit depuis La Silla sous la forme de *logbooks*. La question est de savoir s'il est possible de récupérer et de sauver ces messages.

Les « cahiers de manip » tenus par les observateurs sont disponibles. Marc Moniez les a conservés. Il serait donc possible de les scanner afin de les préserver. Travail fastidieux, mais ne présentant pas de difficultés particulières. La disponibilité d'un scanner A3 faciliterait toutefois l'opération.

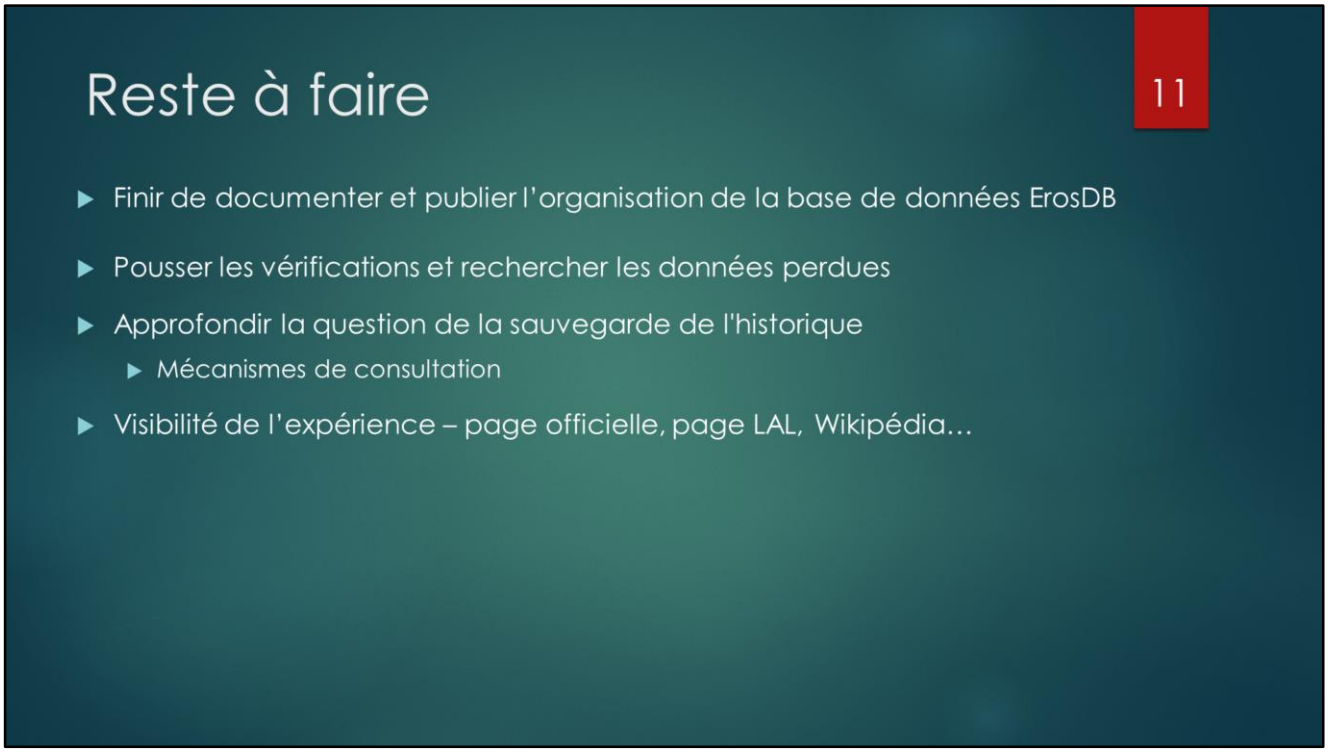
Un point à préciser est de savoir comment présenter ces informations si celles-ci pouvaient être récupérer.



Est-il envisageable d'enregistrer dans la base de données les références des 90 millions d'étoiles analysées ? Une tentative avait été faite il y a des années sans aboutir, faute de temps et de persévérance…

Est-il envisageable d'enregistrer les courbes de lumière de ces 90 millions d'étoiles ?
Pour quel usage ?

Après tout, le mécanisme de recherche grâce aux catalogues embarqués dans la base de données fonctionne bien. Il est implémenté sous la forme d'une application Java, mais, connaissant la structure de la base de données, la logique peut être implémentée dans n'importe quel langage.



La base de données Oracle, nommée ErosDB, est un élément central de l'organisation des données Eros. Elle est partiellement documentée dans le site ErosDB II mais il faut poursuivre la présentation des différentes tables et finir de les publier.

Il faut poursuivre la vérification des données, en particulier dans les espaces HPSS et anciennement AFS, encore inexplorés.

Il faut enfin approfondir la question de la sauvegarde de l'historique des observations, et plus particulièrement de la manière dont ces informations peuvent être présentées.

Il existe une page Eros au LAL référencée dans Google et Bing, mais la page a disparu (https://www.lal.in2p3.fr/EROS/).

Il n'y a aucune page « Eros, Expérience de Recherche d'Objets Sombres » dans Wikipédia, ni FR, ni EN – pourtant, de mémoire, il me semble qu'une telle page ait existé – au moins dans Wikipédia.fr…

La page officielle http://eros.in2p3.fr/ a *beaucoup* vieilli…

12

Jean-Noël Albert – LAL / IJClab

- Le blog Eros / Anastasis : https://groups.lal.in2p3.fr/erosanastasis/
- Le site ErosDb II : http://eros.lal.in2p3.fr/ErosDB/
- Le site historique Eros : http://eros.in2p3.fr/