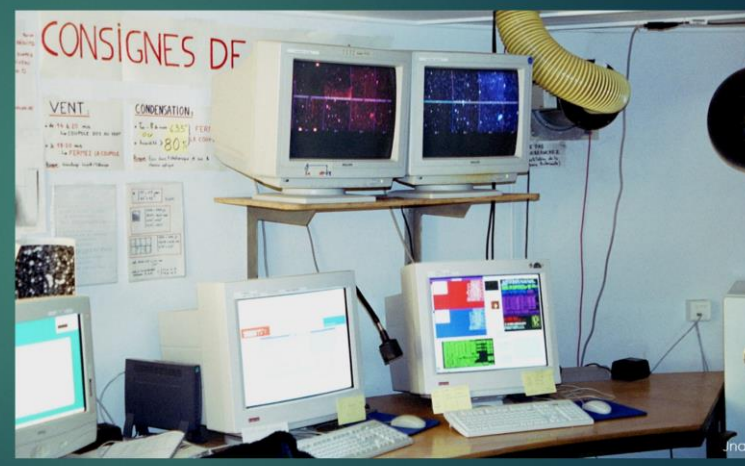
Migration des données Eros II



Jean-Noël Albert – LAL / IJClab

Juin 2020

Durant près de 7 années, de Juin 1996 à Février 2003, Eros II a réalisé plus de 120 000 observations du ciel austral en direction des Nuages de Magellan et de la Galaxie, ce qui représente de 2 millions de vue et 90 millions d'étoiles analysées.

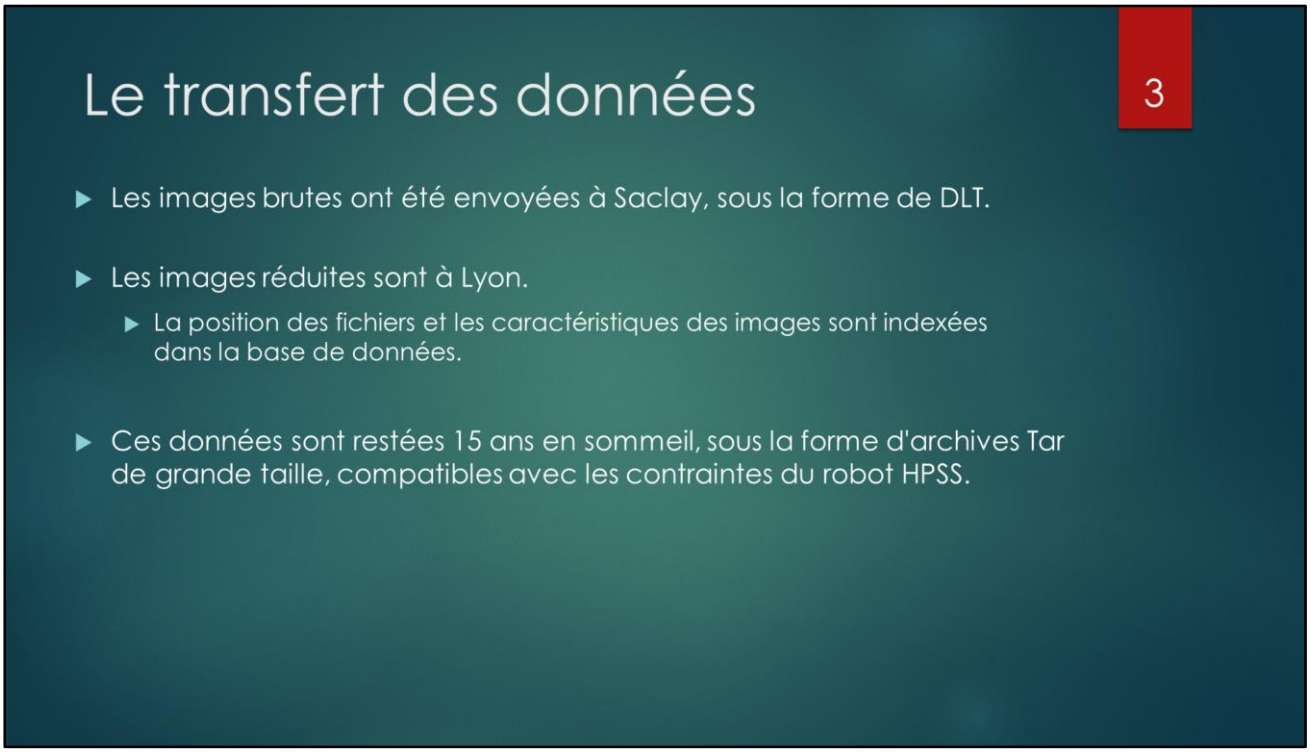
Cet ensemble de données constitue une vaste base de référence qui pourrait intéresser des expériences plus récentes. Sous l'impulsion de Jim Rich et Marc Moniez, la publication de ces observations est donc étudiée.

Cette présentation fait le point sur la situation des données Eros à Lyon.

Les données Eros 2

2

- Les données Eros 2, c'est :
  - 2 millions d'images représentant 14 To
  - 70 000 fichiers binaires de mesures pour 6 To
  - 90 millions de courbes de lumière ASCII, correspondant aux étoiles suivies, soit 500 Go sous forme compressée.



Les images « brutes » issues directement des CCD ont été expédiées à Saclay sur des DLT où elles demeurent. Les images « réduites », résultant de la calibration des images brutes, sont au Centre de calcul de l'IN2P3, à Lyon.

A la réception des images réduites, celles-ci étaient recopiées sur des cartouches 3490 et l'emplacement et les principales caractéristiques des images étaient répertoriées dans une base de données Oracle du Centre.

Avec le passage au robot de cartouches HPSS et à des médias plus performants et de plus grandes capacités, il a été nécessaire de procéder à des réorganisations. La principale, effectuée après l'arrêt de l'expérience, a consisté à regrouper les fichiers sous la forme d'archives Tar de grandes tailles, plus adaptés aux contraintes des *streamers* du robot. Cette réorganisation a permis de préserver les données Eros II durant les 15 années où l'expérience est restée en sommeil.

Pour publier les données Eros, il faut pouvoir les accéder dans de bonnes conditions, ce que ne permettait pas l'organisation sous la forme d'archives Tar. Le choix fait par l'expérience et le Centre est d'utiliser Irods pour conserver les fichiers et d'éviter le recours à des archives Tar.

La migration s'est déroulée en deux temps. Tout d'abord, les équipes du CC ont transféré globalement les archive Tar vers Irods. Nous avons ensuite extrait les fichiers des archives, nous les avons recopiés individuellement dans des arborescences Irods et nous avons actualisé leur emplacement dans la base de données.

Les opérations se sont déroulées pour l'essentiel en batch, sous le contrôle d'applications Java développées spécifiquement, utilisant à la fois l'accès intégré de la librairie Irods (Jargon NIO) et la connexion à la base de données afin de piloter les transferts et actualiser les informations sur les fichiers transférés.

Différentes opérations annexes ont été entreprises, comme la sauvegarde des images calibrées astronomiquement après « normalisation » de leur nom – afin de pouvoir les intégrer aux standards Eros et à la base de données.

L'architecture des répertoires dans Irods est basée sur l'organisation naturelle des données Eros : tout d'abord le type de donnée, puis en second niveau le programme scientifique, au troisième niveau le champ.

Les niveaux suivants dépendent de la nature des données et du nombre de fichiers, le souci étant de ne pas avoir un nombre trop élevé d'entrées par répertoire.

Les données Eros II sont dans une arborescence "**data/eros2/**", ce qui laisse la possibilité d'intégrer d'autres expériences et typiquement les données d'Eros 1 Plaques et d'Eros 1 CCD …

Pour les images, compte tenu du grand nombre de fichiers, l'arborescence est composée de trois niveaux :

- le programme scientifique – ici le LMC, code "lm"
- le champ – ici le champ 003
- la caméra et le CCD – ici la caméra 0, correspondant au filtre rouge, et son CCD 1

Il n'y a pas de décomposition explicite en termes de couleur puisque la caméra 0 est toujours associée à la couleur rouge et la caméra 1 à la couleur bleue. Ce choix de regrouper la caméra et le CCD fut fait par « les pères fondateurs », lors du « grand sommeil ». Il a été conservé pour Irods. Il présente l'intérêt d'éviter un niveau un peu artificiel *caméra/ccd*, où il n'y aurait que deux entrées dans le répertoire *caméra* et seulement huit dans le répertoires *ccd*…

Toutes les images de la même caméra et du même CCD sont regroupées dans le même répertoire, quel que soit leur nature :

- images réduites – les plus nombreuses bien sûr
- images composées, construites par quart de CCD
- et parfois, mais rarement, des images brutes

Le décodage du nom de l'image se fait comme suit :

lm 003 0 1 k r x 6a31 50.fits
lm 003 0 1 t r r cj31 81.fits

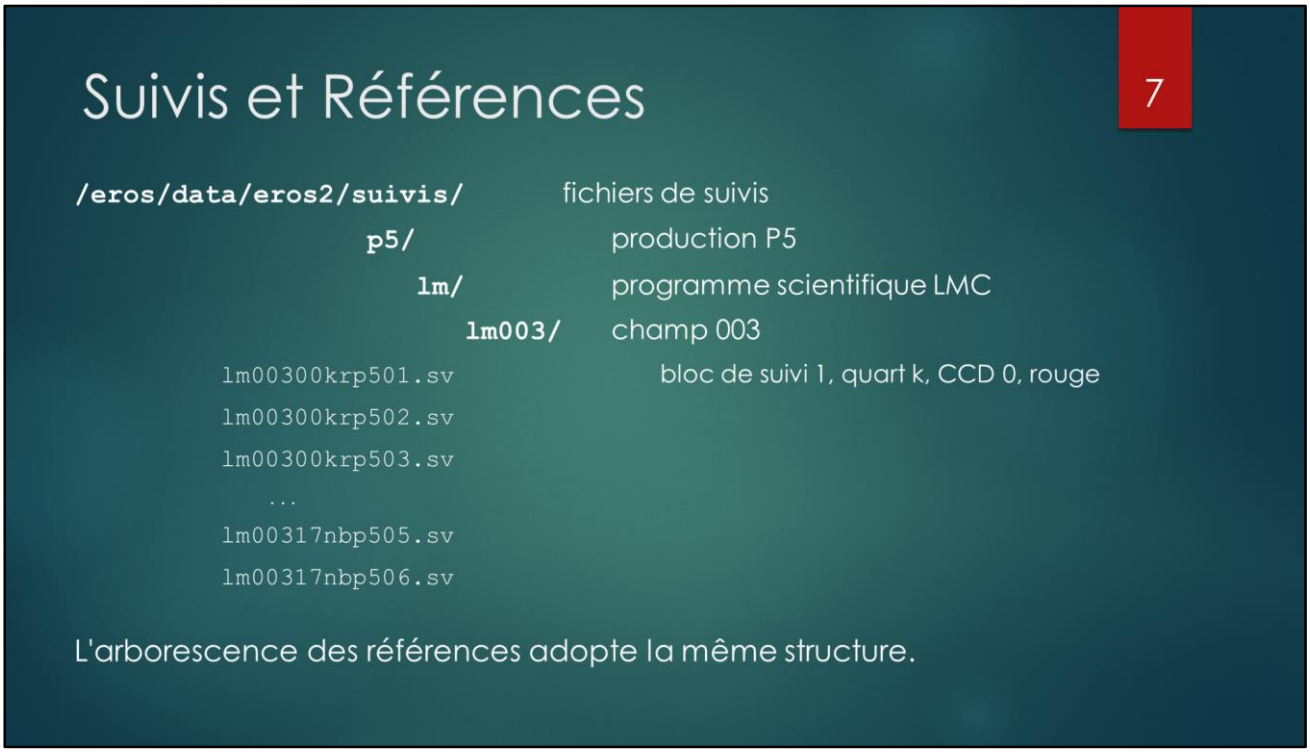
- le programme scientifique, codé sur 2 lettres : ici "lm" pour le LMC
- le nom du champ, codé sur 3 alphanumériques – mais généralement des chiffres : ici le champ "003"
- le code camera : "0"
- le code CCD : "1"

- le code de découpage : **t**: image complète, **k**: quart de CCD "k", **l**: quart "l", …
- le filtre de couleur : **r**: rouge, **b**: bleu
- le code de traitement : **r**: réduite, **b**: brute, **c**, **x**: composée "c" ou "x", **w**: composé callée
- la date, sur 4 alphanumériques
- un numéro de prise de vue

Bien que dans le cas d'Eros II le filtre de couleur et la caméra soient liés, les deux informations figurent dans les noms des images, mais aussi dans ceux des références et des suivis, car lors du choix de nommage des images la possibilité d'utiliser d'autres filtres devait être préservée.

La date est codée sur 4 alphanumériques selon une logique subtile – qui n'a rien à envier aux dates astronomiques !

Un outil ***ObjectName***, présenté par la suite, permet de décoder les noms des données Eros.



Les arborescences des fichiers de suivi et des fichiers de références (les catalogues binaires des étoiles) suivent une logique similaire à celle adoptée pour les images, mais avec des différences liées à leur nature et au nombre beaucoup plus réduit de fichiers :

- un niveau supplémentaire est ajouté au sommet de la hiérarchie afin de distinguer les différentes productions
- il n'y a pas de sous-niveau correspondant aux CCD ni aux caméras, le nombre de fichiers par champ étant relativement limités (400 environ)

A cette heure, le choix de l'expérience a été de ne transférer que la production P5, mais de l'espace a été réservé pour le transfert éventuel des autres productions.

Le décodage du nom d'un suivi se fait comme suit :

lm 003 0 0 k r p5 01.sv

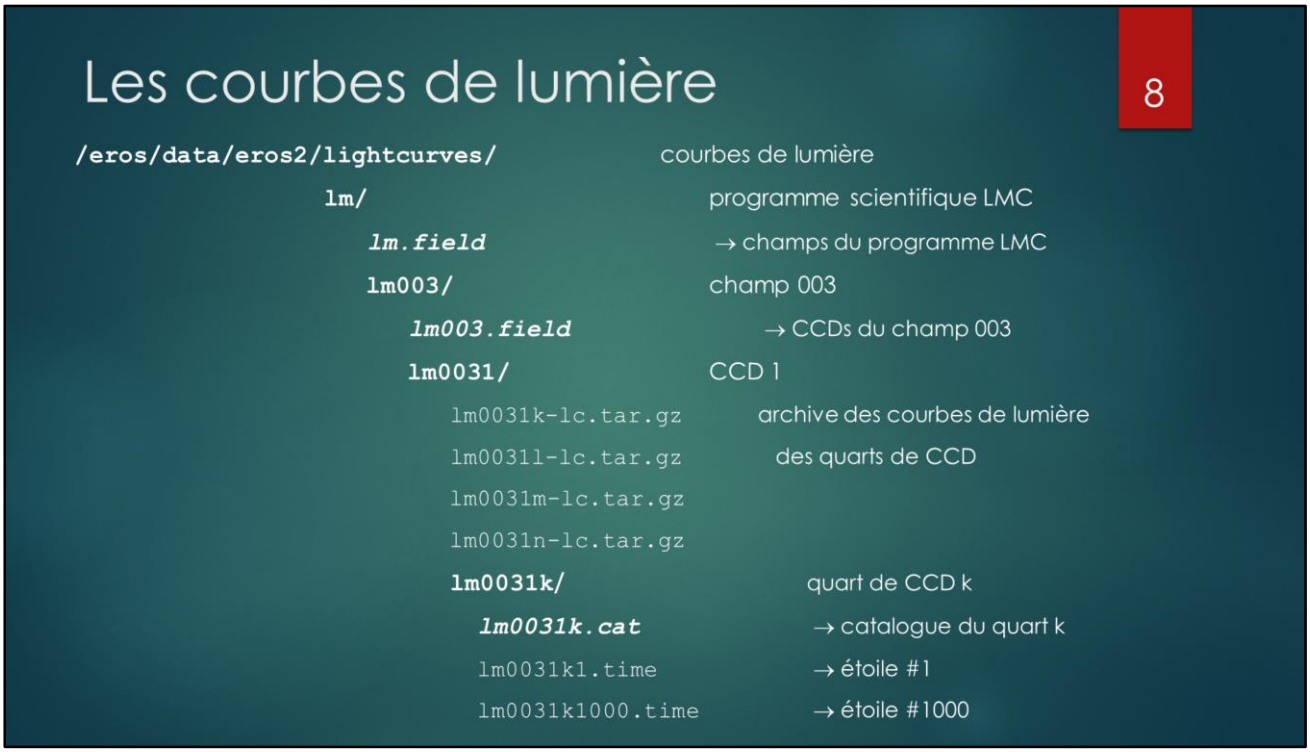
- programme scientifique : ici "lm" pour le LMC
- le code du champ : ici le champ "003"
- le numéro de la caméra : ici la caméra "0"
- le code du CCD : ici le CCD "0"
- le code découpage de l'image : ici le quart de CCD "k"
- l'indication de la couleur : ici "r" pour rouge
- **l'indication de la production** : ici la production "P5"
- **le numéro de bloc du suivi** : ici le premier bloc, ie. "1"

L'arborescence des références est similaire à celle des suivis.

Le décodage du nom d'une référence se fait comme pour les suivis, mais sans numéro de bloc.

lm 003 0 0 k r p5.ref

- programme scientifique : ici "lm" pour le LMC
- le code du champ : ici le champ "003"
- le numéro de la caméra : ici la caméra "0"
- le code du CCD : ici le CCD "0"
- le code découpage de l'image : ici le quart de CCD "k"
- l'indication de la couleur : ici "r" pour rouge
- **l'indication de la production** : ici la production "P5"



L'arborescence des courbes de lumière est similaire à celles des images, si ce n'est qu'il n'y a pas de distinction en termes de caméra et de couleur, puisque les courbes de lumière regroupent les mesures des deux caméras et donc des deux couleurs. Mais à l'inverse, compte tenu du nombre très élevé de fichiers, un niveau supplémentaire est prévu pour séparer les courbes en termes de quarts de CCD. Une fois encore, cette organisation suit les choix faits lors de la création des archives Tar – simplifiés du fait de l'utilisation d'Irods.

Les fichiers de description des champs des programmes scientifiques, des CCD et quarts de CCD des champs et les catalogues des étoiles des quarts de CCD sont inclus dans l'arborescence.

Une autre différence importante par rapport aux autres arborescences est que les courbes de lumière sont regroupées dans des archives Tar, par quart de CCD, mais que les fichiers ASCII sont indexés individuellement dans le catalogue Irods. Ce choix résulte du nombre très important de fichiers (près de 90 millions, soit plus de ***15 mille*** courbes par quarts de CCD denses, comme ceux du LMC) et de la taille réduite de chaque fichier – de 15 à 20 Ko. Le regroupement par archive Tar limite la pollution de l'espace de stockage par un trop grand nombre de petits fichiers alors que leur indexation individuelle dans le catalogue Irods permet de les accéder directement.

La seule conséquence à cette organisation se fait sentir lors d'un premier accès à une courbe de lumière. Le système de stockage doit récupérer l'archive afin de pouvoir en extraire le fichier, ce qui peut entrainer un délai. Mais le système de cache d'Irods fait que l'archive reste un certain temps sur les disques internes : les accès suivants sont donc bien plus rapides.

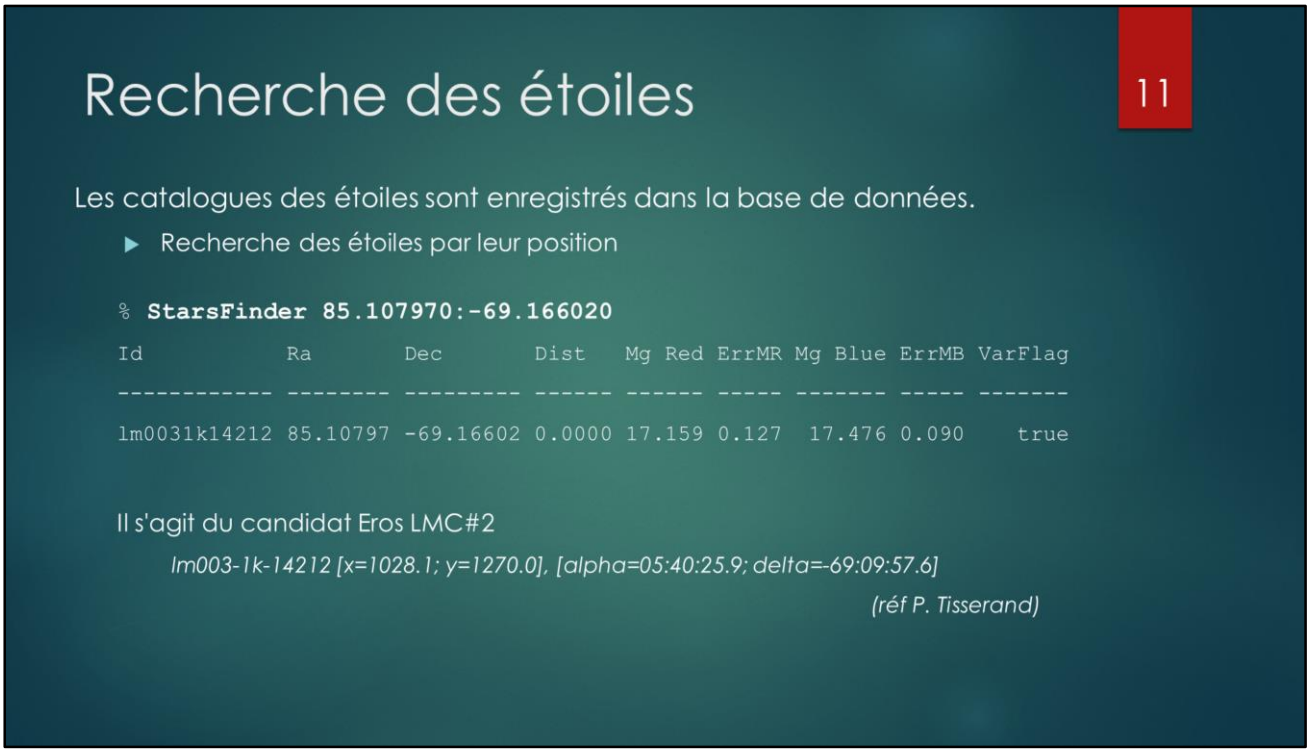
L'en-têtes FITS contienne des informations utiles sur l'image, son codage et les conditions d'observation. Lors de la migration, les en-têtes ont été extraites et sont conservées dans une arborescence symétrique à celles des images.

Du fait de la petite taille de ces fichiers – 5 à 10 Ko – leur accès est bien plus rapide que s'il fallait transférer l'ensemble des images (8 Mo).

L'exemple montre l'utilisation de la commande Irods ***iget*** pour afficher un en-tête FITS à l'écran.

Les en-têtes FITS sont également conservés dans la base de données sous la forme de BLOB (*Binary Large Object*) compressés. Ceci permet de les accéder directement sans même avoir à activer Irods et surtout sans connaitre l'emplacement du fichier dans les répertoires et sous-répertoires grâce à l'application ***FitsHeader***.

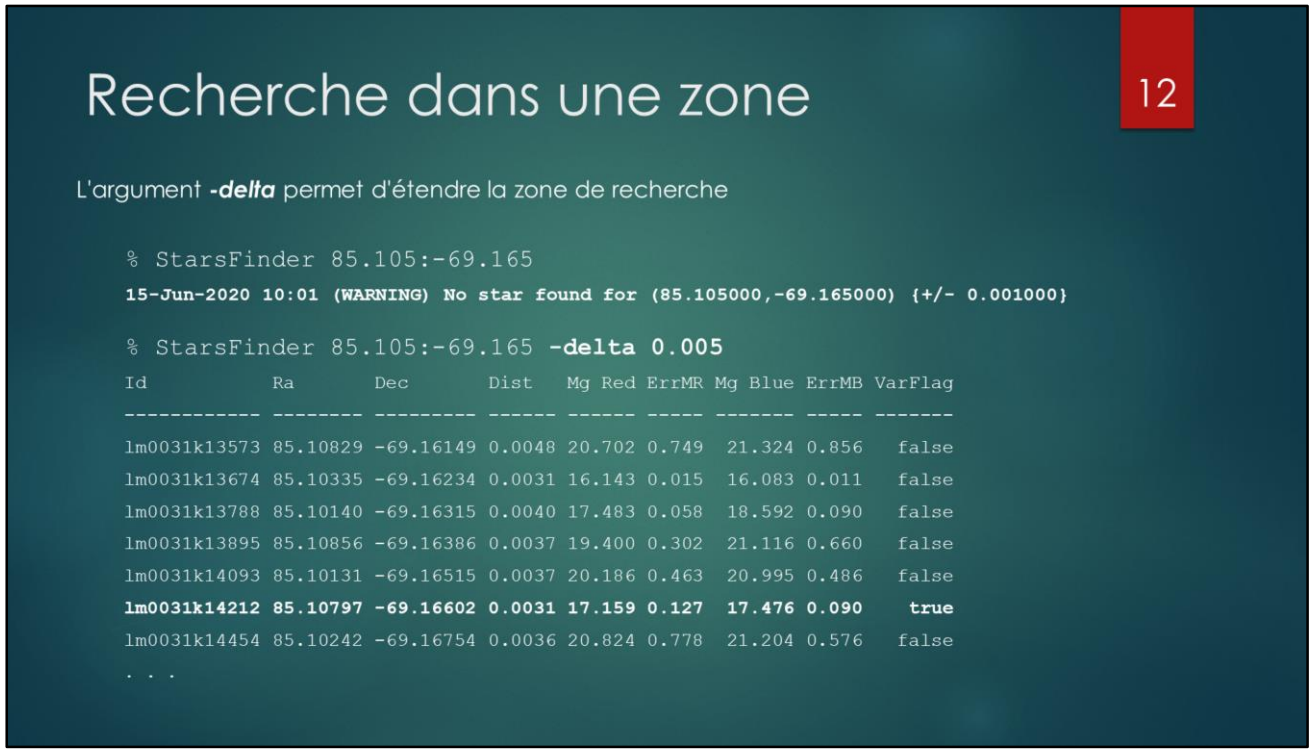
Par ailleurs, *FitsHeader* assure la mise en forme de l'en-tête de manière à s'adapter au terminal. En effet, un en-tête FITS est une suite d'enregistrement de 80 caractères *sans terminateur de ligne* – images des anciennes cartes perforées IBM. Il faut donc caller son terminal sur 80 colonnes faute de quoi le résultat est assez "hasardeux". *FitsHeader* évite ce désagrément. Ce n'est qu'un petit détail, mais qui s'avère bien pratique à l'usage.



Une autre extension de la base de données est l'enregistrement des catalogues des étoiles sous la forme de BLOB comprimés. Ceci permet de rechercher rapidement une étoile, ou les étoiles d'une zone, à partir de la position dans le ciel.

Remarque: Il faudra adapter le programme pour qu'il accepte les coordonnées sous une forme horaire – la conversion entre les coordonnées astronomiques utilisées par Patrick Tisserand – et sans doute par l'ensemble de la communauté astrophysiciennes – et les coordonnés des catalogues ASCII étant quelque peu fastidieuse.

Il serait aussi utile que le programme accepte directement les identifiants Eros des étoiles.



Dans l'exemple précédent, les coordonnées correspondaient exactement à l'étoile cherchée, cette seule étoile était donc présentée.

Mais si l'étoile n'est pas repérée avec précision, la recherche peut ne rien donner, ce qui est le cas de l'exemple présenté.

Dans cette situation, il est possible d'étendre la zone de recherche, quitte à récupérer trop des références, comme le montre le second exemple.

Informations sur les étoiles cherchées

13

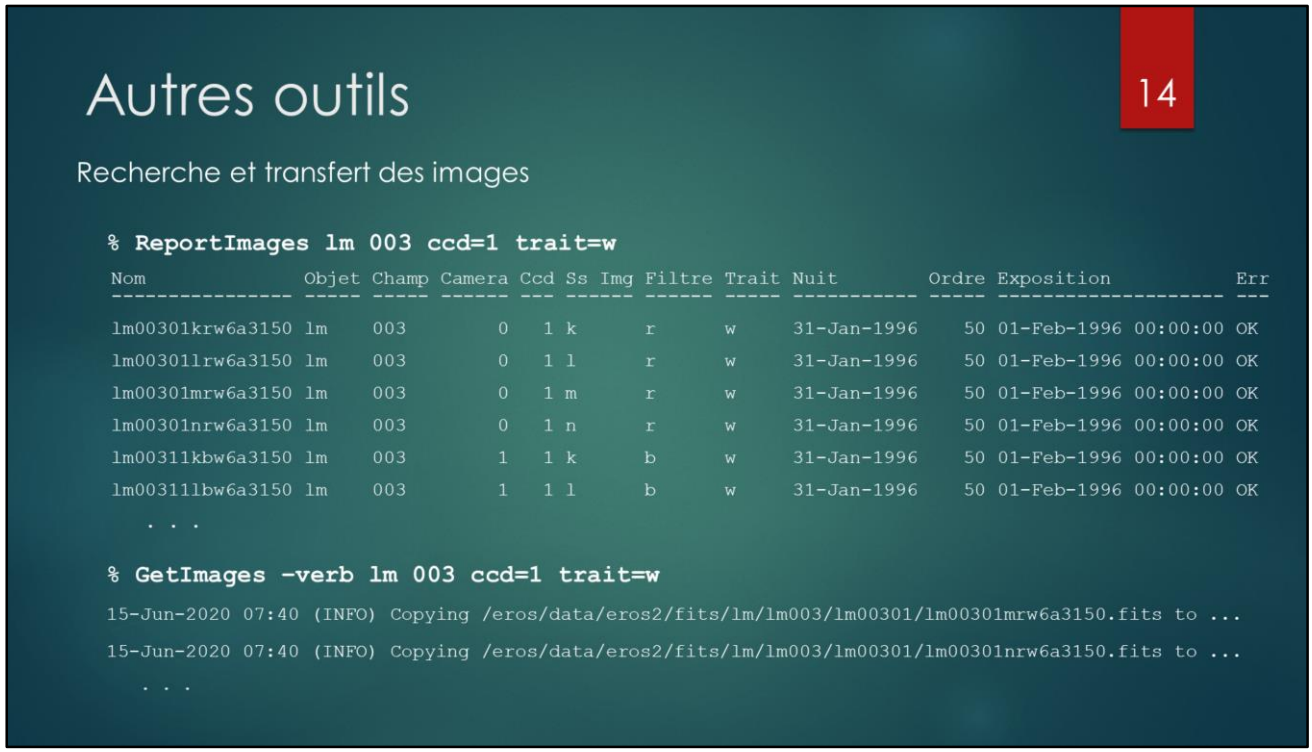
Accès à la courbe de lumière, à la liste des images de l'étoile, ...

```
% StarsFinder -verb 85.1079166:-69.166 -light -save
Id          Ra       Dec       Dist   Mg Red ErrMR Mg Blue ErrMB VarFlag
----------- -------- --------- ------ ------ ----- ------- ----- -------
lm0031k14212 85.10797 -69.16602 0.0001 17.159 0.127  17.476 0.090    true

Light curve file names in iRdos
-------------------------------
irods:/eros/data/eros2/lightcurves/lm/lm003/lm0031/lm0031k/lm0031k14212.time

12-Jun-2020 11:55 (INFO) Saving stars information to /sps/hep/eros/users/albert/85.107917_-69.166000+-0.001000.cat
12-Jun-2020 11:55 (INFO) Saving 1 light curve

% StarsFinder 85.1079166:-69.166 -delta 0.001 -images
Nom               Objet Champ Camera Ccd Sous Image Filtre Traitement Nuit        Ordre
----------------- ----- ----- ------ --- ---------- ------ ---------- ----------- -----
lm00301trr6g29119 lm    003        0   1 t          r      r          29-Jul-1996   119
lm00311tbr6g29119 lm    003        1   1 t          b      r          29-Jul-1996   119
lm00301trr6g31145 lm    003        0   1 t          r      r          31-Jul-1996   145
lm00311tbr6g31145 lm    003        1   1 t          b      r          31-Jul-1996   145
 (...)
lm00311tbrdb1226  lm    003        1   1 t          b      r          12-Feb-2003    26
lm00311tbrdb1517  lm    003        1   1 t          b      r          15-Feb-2003    17
lm00311tbrdb2349  lm    003        1   1 t          b      r          23-Feb-2003    49
lm00311tbrdb2635  lm    003        1   1 t          b      r          26-Feb-2003    35
```

StarsFinder permet aussi d'afficher la liste des courbes de lumière des étoiles identifiées, la liste des images où ces étoiles apparaissent, les catalogues où elles sont référencées, etc.

L'outil permet également de récupérer les éléments trouvés.

Attention toutefois avec les images ! Rien que pour le candidat Eros de l'exemple, il existe plus de mille références.



L'outil ***ReportImages*** permet de présenter les références des images correspondant à une sélection. La commande permet d'indiquer directement le programme, le champ, la caméra mais aussi tout autre paramètre de l'image en indiquant son nom et sa valeur, dans un mélange des conventions des commandes UNIX et des requêtes SQL.

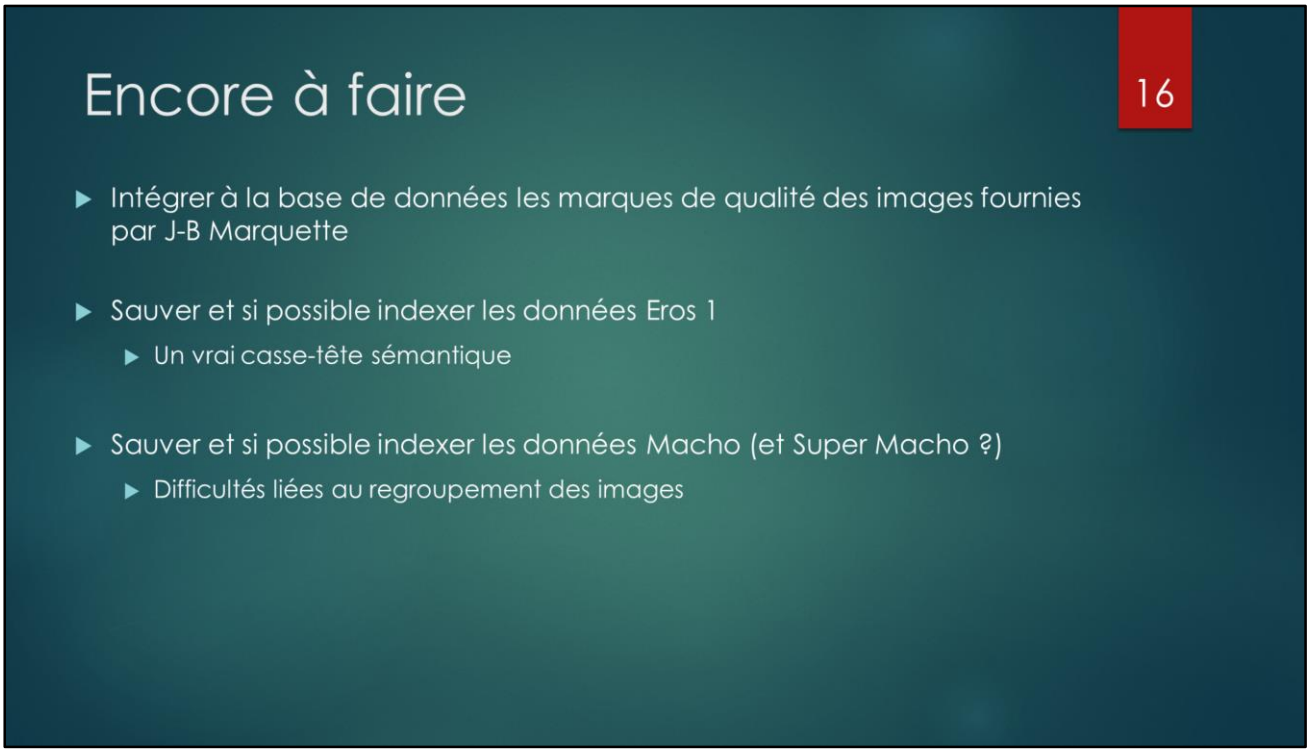
L'exemple présente la recherche des images composées correspondant au CCD 1 du programme LMC, champ 003, couleurs et caméras réunies.

La commande ***GetImages*** constitue le pendant de *ReportImages*, destiné au transfert des images sélectionnées.

Il existe des équivalents pour les suivis nommés ***ReportSuivis*** et ***GetSuivis***.

Les noms des images, des suivis, des références Eros II sont précis et cohérents, ce qui est un sérieux progrès par rapport à Eros 1 (!).

Mais la compréhension de ces noms n'est pas pour autant simple, surtout si on ne pratique pas très régulièrement ! La commande ***ObjectName*** décode les noms des objets Eros II et les présente de manière claire.



J-B Marquette nous à fournir une série d'indication sur la (mauvaise) qualité de certaines images qu'il faut intégrer à la base de données. Ceci va sans doute nécessiter d'étendre la définition de la table des images.

Marc Moniez et Jim Rich souhaiteraient que les données Eros 1 soient sauver elles-aussi. Copier les fichiers dans Irods restent possibles, mais la manière dont ces fichiers sont nommés est un vrai casse-tête !

Marc souhaiterait aussi sauver les données Macho afin constituer une seconde source pour ces données. Ici encore, la recopie des fichiers reste possible, mais la difficulté pour leur indexation dans la base de données réside dans le fait que les fichiers FITS regroupent plusieurs images.

A suivre : migration vers Irods

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/