Livrable 2.4.1 - mai 2016

Analyse des pratiques et instruments existants

Résumé : L'objet d'étude en astrophysique dans le cadre du WP2 d'Epistémè est le "ground segment" ou segment sol d'une mission spatiale scientifique. Dans ce premier rapport, nous présentons d'abord le segment sol et le rôle du numérique dans sa préparation et dans son mode opératoire, puis comment nous nous intéressons aux pratiques des scientifiques qui y travaillent, à la fois au niveau théorique (quelle approche épistémologique) et pratique (quels sont les outils numériques que nous pourrions observer). Nous concluons en présentant notre plan de travail pour 2016-2017.

Introduction

Ce rapport s'insère dans le lot de travail 2.4 du projet Episteme (*Collecte, modélisation des traces et outils réflexifs*) piloté par le LINA, en lien étroit avec le lot 2.2 (*Etude de Cas 1 : Phénoménotechnique et cycle des images en astrophysique*) piloté par le CEA IRFU.

Cette tâche porte sur l'analyse des traces produites par les scientifiques au cours de leurs pratiques journalières. Il s'agit de documenter au mieux ces pratiques en s'appuyant sur les outils numériques utilisés, avec trois objectifs principaux :

- Fournir un support à l'analyse épistémologique de l'étude de cas en apportant aux chercheurs épistémologues des outils de visualisation permettant de confirmer, préciser, interroger voire réorienter le cas échéant les concepts et approches théoriques considérées.
- Fournir un support réflexif aux chercheurs eux-mêmes, en leur apportant des outils de visualisation de leurs pratiques suffisamment pertinents pour leur permettre d'interroger celles-ci, notamment pour ce qui concerne la mobilisation plus ou moins automatique de traitements algorithmique, l'influence médiée numériquement des pairs, etc. On contribuera ainsi à l'enrichissement des pratiques à partir des traces et à la « mise en place de procédures de réflexivité dans les milieux associés ».
- Fournir un support de médiation permettant l'exposition des pratiques scientifiques réelles sur des cas exemplaires. Il ne s'agit pas ici de publier directement les traces brutes, mais de définir les transformations ou les simplifications nécessaires afin de garantir l'accessibilité minimale des pratiques au grand public.

Le travail devait être assuré par un post-doctorant en IHM/Visualisation du LINA en collaboration avec un post-doctorant épistémologue du CEA. Le recrutement de ces post-doctorants n'a pu être effectué au cours des premiers mois du projet comme cela avait été préssenti dans le planning initial. C'est finalement un ingénieur de recherche qui a été recruté pour la partie IHM / visualisation;. Sa mission a débuté le 3 mai 2016. Il aura pour tâche la collection des traces d'activité du segment sol, puis leur exploitation et leur visualisation.

Le segment sol et le programme d'observation Gould Belt

Dans la chaîne des traitements des données en astrophysique, le segment sol débute à la réception des signaux satellitaires, se poursuit avec leur étalonnage et leur analyse, et se termine avec l'archivage des données[1]. Nous nous focaliserons a priori sur les deux missions Herschel et Euclid, la première étant terminée au sens que le satellite n'est plus opérationnel mais la phase post-opération continue, la seconde étant en cours de conception et s'appuie fortement sur la simulation numérique dans sa préparation.

Herschel[2] est un téléscope spatial de l'Agence Spatiale Européenne (ESA), lancé en mai 2009 et cessant ses observations en avril 2013. La mission est aujourd'hui en phase post-opération. Son objectif est d'étudier la formation des galaxies et des étoiles. Herschel contient trois instruments (PACS, SPIRE, HIFI) fournis par trois consortia de laboratoires qui les ont financé et ont en retour des heures d'observation garanties. Les équipes du consortium qui participent à ces programmes de recherche clé (Guaranteed time Key programmes) gèrent en interne la répartition des heures d'observation et sont les premières à recevoir les données. La planification est effectuée en amont du lancement du satellite via le logiciel HSPOT (Herschel Observation Planning Tool), puis, chaque jour, des AORs (Astronomical Observation Requests) contenant la paramétrage des zones à balayer (coordonnées, vitesse de scan…) sont transmis au télescope pour son travail de la journée à suivre.

Le segment sol d'Herschel[3] (Herschel Science Ground Segment, H-SGS) se compose de trois pipelines, un par instrument. Les données sont tout d'abord reçues à Darmstadt au centre de commandement des opérations (Mission Operations Centre - MOC), puis envoyées à Madrid dans le centre scientifique (Herschel Science Centre - HSC), puis dispatchées vers les trois centres de contrôle liés aux instruments spécifiques (Instrument Control Centres - ICC), et enfin envoyées aux scientifiques des différents laboratoires. Les ICC ne sont pas réellement des centres géographiques, mais les noeuds d'un réseau.

Les données reçues du satellite sont traitées par le HSC et les ICC qui exécutent systématiquement le pipeline de traitement des données d'Herschel (Herschel Data Processing system[4]). Les traitements effectués sont essentiellement de l'étalonnage (*data reduction*, e.g. traduction volt vers puissance/énergie, projection dans un système de coordonnées, etc.) et du contrôle de qualité.

Les données deviennent alors accessibles aux chercheurs sous la forme de fichiers FITS (Flexible Image Transport System[5]). L'environnement open source HIPE (Herschel Interactive Processing Environment) permet d'accéder aux données, de ré-exécuter les traitements effectués automatiquement, de tester ses propres traitements soit pour la réduction de

[1] Cette définition peut être étendue si on prend en compte le travail de simulation avant une acquisition de données, la préparation a priori de traitement d'étalonnage, la publication scientifique, etc.
[2] http://www.cosmos.esa.int/web/herschel
[3] http://sci.esa.int/herschel/43788-science-ground-segment/
[4] http://www.cosmos.esa.int/web/herschel/data-processing-overview
[5] http://fits.gsfc.nasa.gov/fits_documentation.html. On pourra trouver des exemples de fichiers FITS sur le site http://archives.esac.esa.int/hsa/aio/doc/postcardGallery.html (par exemple en tapant M31).

données, soit pour l'analyse, de visualiser les résultats, etc. mais aussi de réintégrer les résultats dans l'archive Herschel. HIPE est composé de modules de traitement (Tasks). Certains modules livrés avec l'outil sont basés sur les étalonnages des instruments issus du travail des ICC, sur les connaissances du satellite par les ingénieurs de l'ESA et de l'industrie (ex. le modèle de pointage du télescope), etc. D'autres modules peuvent également être ajoutés par les chercheurs dans HIPE.

Par ailleurs, les chercheurs des laboratoires ne sont pas obligés d'utiliser HIPE et peuvent également analyser les données en développant localement des algorithmes et des logiciels permettant l'extraction des sources, la visualisation des données réduites et la réalisation d'images "esthétiques" pour la publication des résultats. La phase initiale de data reduction sera refaite avec HIPE[6], tandis que l'analyse de donnée (qui produit des donnés dites "de niveau supérieur") peut être réalisée avec d'autres logiciels, le cas échéant par des chercheurs qui ignorent les caractéristique de l'instrument ou la technique d'étalonnage.

Les résultats ainsi que les données brutes sont versées dans l'archive ouverte Herschel (Herschel Science Archive - HSA[7]) et mises à la disposition de la communauté. En effet, l'organisation scientifique en consortia fait que chacun doit fournir un relevé complet de ses observations sous la forme d'une archive : images (en utilisant HIPE), catalogue de sources et cartes associées[8], code/algorithme faits maison en open source[9], etc. Les programmes clés ont également leurs propres archives, par exemple celui sur la Gould Belt[10].

Nous allons travailler sur le segment sol de Herschel car cette recherche ouvre sur l'avenir, notamment sur la prochaine grande mission spatiale européenne baptisée Euclid que pilote l'ESA et dont le segment sol est coordonné par le CEA. Le satellite Euclid[11] sera lancé en 2020. Son objectif est détudier la matière noire (Dark matter) en lien avec la formation des grandes structures et l'énergie noire ou sombre (Dark Energy) en lien avec l'expansion accélérée de l'univers. Il aura deux instruments (VIS, NISP). [12]

Comme mentionné précédemment, la mission Herschel comporte de nombreux programmes clés. L'une d'elle, Gould Belt, a pour but d'étudier la formation des étoiles dans l'univers froid via les instruments PACS et SPIRE. Elle a ainsi permis de mettre en évidence la structure

---

[6] Ou avec des outils *ad hoc* qui ont servi à construire HIPE.
[7] http://www.cosmos.esa.int/web/herschel/science-archive
[8] Une source étant un point intéressant de l'univers contenant une certaine densité de lumière, qui est repéré par des algorithmes sur des cartes, et correspond sans doute à un objet physique.
[9] Par exemple *getfilaments* et *getsources* développés au CEA Irfu.
[10] http://www.herschel.fr/cea/gouldbelt/en/Phocea/Vie_des_labos/Ast/ast_visu.php?id_ast=66
[11] http://www.euclid-ec.org/
[12] À la différence d'Herschel, le segment sol d'Euclid, du fait de la spécifité de sa mission scientifique, fera intervenir de nombreuses simulations numériques en amont des observations afin de simuler les étalonnages et toute la chaine de traitement des données. En ce sens, Euclid est plus proche du satellite Planck que d'Herschel. Une erreur sur l'étalonnage des données se traduira par une erreur d'interprétation des résultats du fait de la nature même du signal très faible à extraire, réduire et analyser. Un cas d'école d'une telle erreur de traitement est la fausse annonce de la découverte des empreintes de la polarisation de la lumière et donc des ondes gravitationnelles primordiales par l'observation au sol Bicep 2.

filamentaire d'un nuage moléculaire et la distribution projetée des cœurs préstellaires. C'est sur ce programme que nous allons concentrer les efforts de notre recherche.

Décrire l'activité des astrophysiciens sur le segment sol

Une double approche de description de l'activité scientifique des astrophysiciens sur le segment sol est possible. La première correspond à l'approche théorique développée à l'IRFU/CEA autour du « cycle de l'image » scientifique, la seconde à une modélisation de l'activité essentiellement fondée sur les traces numériques qu'il est possible de collecter. À ce titre, il ne faut pas oublier que le segment sol est aussi une organisation humaine, composée d'êtres humains et de matériel, et dont le but est de coordonner les opérations. Une telle organisation a dû faire évoluer ses pratiques au cours du temps, entre le début et la fin de la mission. Le segment sol peut donc être considéré comme appartenant à une double temporalité, l'une axée autour de la transduction de l'image, la seconde autour de son élaboration à travers le temps. Nous allons pour le moment nous concentrer sur la première temporalité.

Comme évoqué dans le document scientifique du projet Epistémè, le cycle de l'image[13] proposé par le CEA Irfu vise à étudier l'image scientifique de son origine dans le champ scientifique à sa valorisation hors de ce champ, tout en passant par les étapes de transformation au moyen des outils numériques de nombreuses disciplines scientifiques, artistiques et médiatiques. Il s'agit de :

1. comprendre les conditions de production des images à partir des instruments, de la détection de lumières à leur traduction en signaux électroniques et le traitement des données.
2. étudier la transformation et la circulation des images dans et hors du champ scientifique, opérées par les technologies numériques : traitement des images et média de diffusion.
3. analyser les processus de valorisation symbolique hors et dans le champ scientifique.

L'approche vise dans le projet à comprendre la chaîne de transduction du segment sol. À ce titre, deux visions sont possibles :

- Le cycle minimaliste de l'information : l'image source provenant d'Herschel est étalonnée(via HIPE ou d'autres pipelines d'étalonnage), puis analysée par diverses techniques, archivée dans le Herschel Science Archive, et enfin diffusée à la communauté scientifique ou au grand public
- Le cycle entier de l'information : initialement il était une hypothèse à valider ou informer. Depuis celle-ci, le groupe de scientifique impliqué a déterminé une zone dans l'univers à étudier. Il a alors fallu orchestrer l'étude de cette zone, par sa définition numérique et sa planification via HSPORT et l'envoi d'AORs. La détection a été effectuée par le satellite Herschel, stockée temporairement puis envoyée au HSC. Les étapes de réduction ont été opérées, puis la donnée a été mise à disposition des scientifiques. D'autres travaux ont alors été effectués (étalonnage, nettoyage…) afin d'affiner et analyser les données. Puis des publications ont été diffusées.

[13] Bontems., V., *Le "cycle de l'image" selon Gilbert Simondon. Une définition génétique de l'image scientifique*, Visible, n°8, 2011.

Comme nous le voyons, Images ou graphes peuvent être considérés comme quasi-organismes en développement qui vivent en couplage avec les humains - différents types d'images correspondant à différents organismes plus ou moins développées.

On visera à collecter différents types de traces (Champin et al. 2013)[14] qui seront autant de marqueurs ou d'indices de l'activité des scientifiques et du cycle de l'image. Celles-ci proviendront de sources variées. Logiciels scientifiques pointus comme outils les plus communs seront considérés, ceci afin de modéliser l'activité dans son ensemble et dans sa diversité.

Nous chercherons à instrumenter l'analyse selon l'approche des épistémologues de façon à pouvoir fournir à ceux-ci des éléments permettant de renforcer les analyses, de faire évoluer la théorie le cas échéant, etc. Il s'agit d'être à même de décrire l'activité des scientifiques à partir des traces numérique de celle-ci.

Poursuite du travail

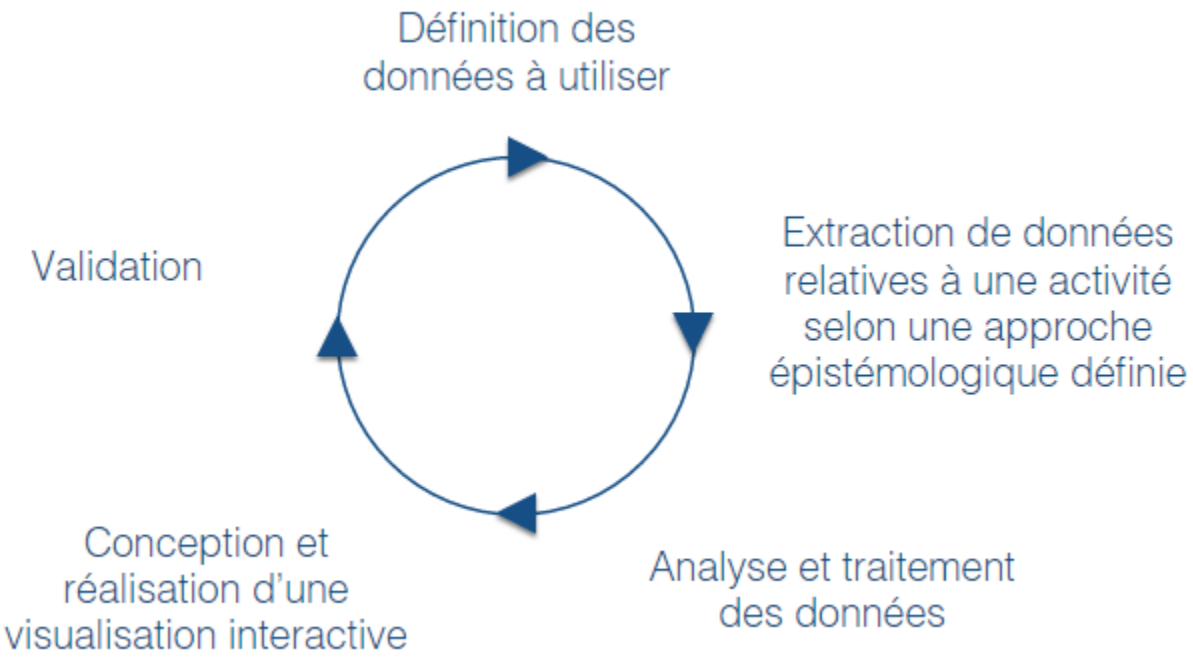
Comme mentionné précédemment, le travail sur la collecte et la visualisation de traces n'a commencé que début mai 2016.

Dans un premier temps nous allons nous concentrer sur la partie "médiation" du projet, en relation avec la journée Episteme de la Digital Week 2016 qui aura lieu le 20 Septembre 2016. La première étape pour l'ingénieur de recherche consistera donc à étudier les étapes de transformation de l'information, depuis sa réception depuis le satellite sous forme brute jusqu'à l'obtention de l'image finalisée. Nous avons déjà mis au point un pipeline théorique qu'il nous appartient de confirmer ou compléter. Pour ce faire, un déplacement au CEA à très court terme est en cours de préparation afin que l'ingénieur de recherche puisse y travailler avec un astrophysicien.
A partir de ces étapes, nous présumons que les traces d'activités correspondantes pourront être mises en évidence. Il faudra en parallèle que chacune des étapes puisse être identifiée et traduite dans un langage "commun", compréhensible par tous, afin de pouvoir les exposer au grand public. Ces traces permettront la conception des premières visualisations en collaboration avec les utilisateurs impliqués.

Par la suite le travail sera itératif : nous commencerons par des visualisations et des traces simples, puis nous les ferons évoluer en fonction des retours des utilisateurs et des possibilités accrues de collecte. Par exemple nous pourrions débuter par ne tracer que les manipulations de fichiers FITS et leur visualisation, éventuellement sans savoir qui les a manipulés, puis passer à des visualisations réflexives notamment pour les doctorants, puis les chercheurs, en complétant avec de nouvelles informations : données liées à l'étalonnage, acteurs des manipulations, échanges de fichiers sur des systèmes de fichier partagés ou par mail, etc. La

[14] Champin, P.-A., Mille, A., and Prié, Y. Vers des traces numériques comme objets informatiques de premier niveau : une approche par les traces modélisées. Intellectica, 59, 2013.

connaissance que cette recherche peut être un moteur puissant pour la conception du programme Euclid.



Pour conclure et afin de rendre davantage concrète l'approche théorique expliquée dans ce document, ci-dessous se trouve un exemple de visualisation interactive de traces numériques résultants de l'activité de wikipedia. Le but est ici d'extraire des informations pertinentes sur les pages afin d'en modéliser le réseau. C'est évidemment un exemple simple, car devant la richesse de l'écosystème "wikipédia", il convient de rester conscient que de très nombreuses visualisations peuvent être proposée, autant dans une approche sémantique que comportementale.


Créations et mises à jour
de pages wikipedia
Production
Logs
• date
• acteur
• contenu
• ...
Extraction
Transformation
Visualisation