AD HALSTRA

Sic Itur Ad Astra

Outil d'importation massive de notices d'ADS vers HAL

Carnet de projet

Table des matières

L'avant projet

Ad HALstra, qu'est-ce que c'est ?

Le nom du projet vient de la citation latine issue de l'Enéide de Virgile «Sic itur ad astra», qui signifie «C'est ainsi qu'on atteint les astres ». Il s'agit également de la devise de l'Observatoire de Paris.

Le projet Ad HALstra a été mis en place à partir d'un constat : ADS, la base de données de la NASA, possède une grande quantité de notices d'articles de chercheurs français avec des informations très complètes qui pourraient être utilisées pour enrichir l'archive nationale en ligne HAL. Ad HALstra a donc pour ambition de transférer les informations des notices provenant d'ADS vers HAL et ainsi permettre à HAL de bénéficier de notices complètes pour tous les laboratoires de l'INSU.

L'équipe

Aurélie Fayard : chef de projet, management et suivi du projet. Exploitation des métadonnées, communication auprès des chercheurs (Bibliothèque de l'Observatoire de Paris).

Alain Courgey : extraction, conversion et diffusion des métadonnées. Co-conception de l'application comprenant les tables de correspondances (Observatoire de Paris).

Charlotte Bultel : suivi du projet, travail préparatoire à l'élaboration de l'application, valorisation du projet, mise en qualité des métadonnées (Bibliothèque de l'Observatoire de Paris).

Soizick Lesteven : suivi du projet, communication auprès d'ADS (Centre des données astronomiques de Strasbourg / Observatoire astronomique de Strasbourg).

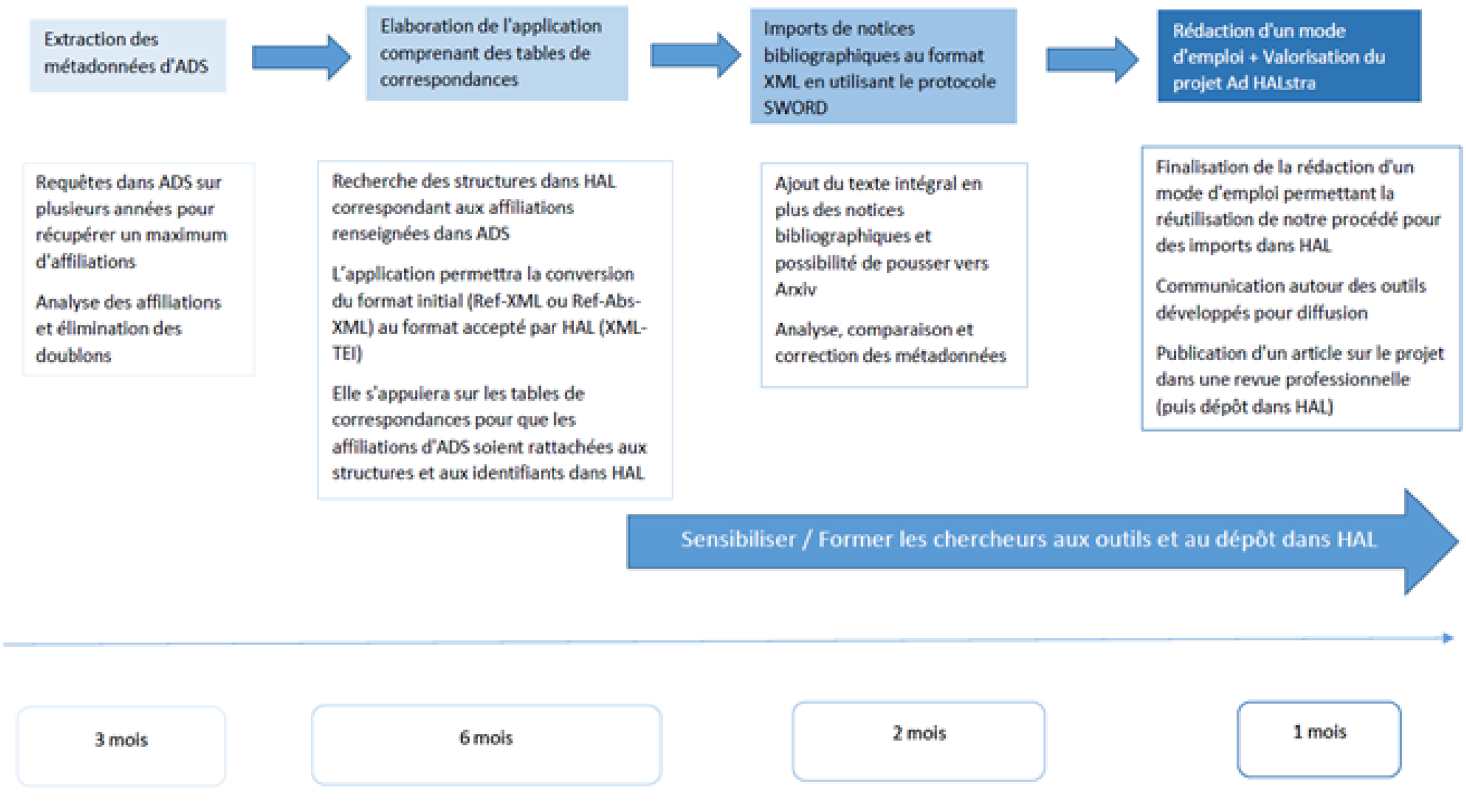
Nathalie Pothier : suivi du projet, appui à l'exploitation des métadonnées, administratrice du portail HAL INSU (Observatoire des sciences de l'univers du Centre Val de Loire).

Isabelle Dubigeon : suivi du projet, appui à l'exploitation des métadonnées, référente HAL (Observatoire de Rennes).

L'avant projet

La chronologie

L'Institut des Sciences de l'Univers est intéressé par ce projet et souhaite financer la mise en place de l'outil. Le projet est prévu pour 6 mois, de juin à novembre 2021. Il sera prolongé pour un mois et sera disponible à partir de janvier 2022. Le coeur du travail autour d'Ad HALstra consiste à extraire des métadonnées d'ADS à l'aide de requêtes et la mise en place d'une application avec les tables de correspondance entre les balises d'ADS et celles acceptées par HAL pour permettre l'importation des notices. Un mois de plus sera nécessaire en raison de l'ajout d'une fonctionnalité permettant l'importation massive des notices via le protocole SWORD mis en place par le CCSD.



Présentation de la chronologie du projet

Le projet

Comment ça marche ?

Pour utiliser l'outil, il faut extraire de la base de données de la NASA, ADS, les notices à intégrer dans l'archive nationale en ligne HAL. Pour faciliter l'extraction de ces notices, des requêtes par laboratoire sont proposées par l'équipe d'AD HALstra pour chaque Observatoire des Sciences de l'Univers (OSU) ainsi que pour chaque laboratoire associé à l'INSU. Ces notices sont extraites au format REFABS XML et insérées dans l'outil Ad HALstra.

L'outil permet d'adapter le format des notices extraites d'ADS pour qu'elles puissent passer du format REFABS XML (un des formats d'extraction permis par ADS) à un format XML TEI (le seul format accepté pour l'import dans HAL). Cette importation se fait grâce au protocole d'interopérabilité SWORD mis en place par le CCSD.

L'outil associe les affiliations auteur renseignées dans ADS aux numéros de structure dans HAL, permettant aux informations d'être référencées correctement.

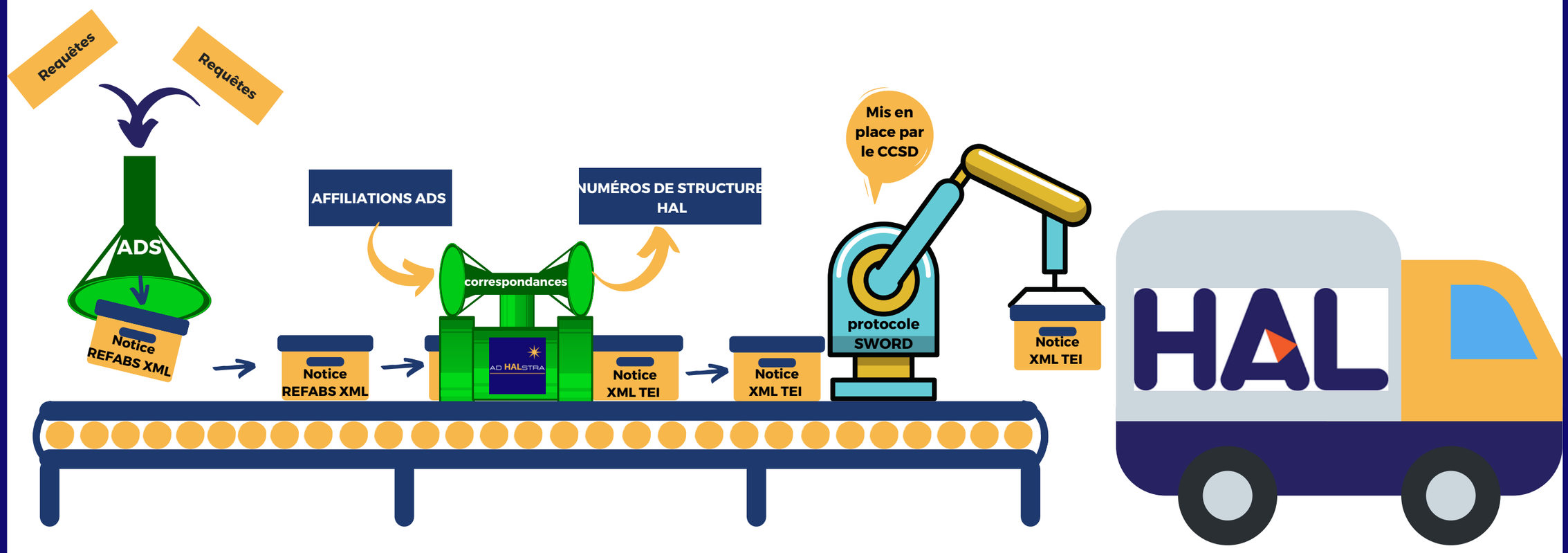


Schéma explicatif du fonctionnement de Ad HALstra

La création du projet

L'extraction des métadonnées d'ADS

Définie sur 3 mois, la première étape, l'extraction des métadonnées, consistait à créer des requêtes pour récupérer un maximum d'affiliations des chercheurs des laboratoires associés à l'INSU. La création de ces requêtes se faisait à partir de la recherche de plusieurs données sur les laboratoires, à savoir le nom du laboratoire, son acronyme, son numéro d'UMR et son nom en anglais.

Exemple de format type d'une requête pour ADS, pour trouver les publications produites par les chercheurs des laboratoires de l'OSU EOST pour l'année 2020 :

(aff:("IPGS" OR "Institut de physique du globe de Strasbourg" OR "UMR 7516" OR "Strasbourg Institute of Globe Physics" OR "LHyGeS" OR "Laboratoire d'HYdrologie et de GEochimie de Strasbourg" OR "Laboratory of Hydrology and GEochemistry of Strasbourg" OR "UMR 7517" OR "ITES" OR "Institut Terre et Environnement de Strasbourg " OR "Earth and Environment Institute of Strasbourg" OR "UMR 7063" OR "LabEx" OR "LabEx G-eau-thermie Profonde") AND 2020)

Nous avons pour cela créé une base de données recensant toutes ces informations ainsi que des informations sur la présence des laboratoires sur HAL. Ces informations permettront de créer des correspondances entre les affiliations d'ADS et les affiliations HAL.

Pour s'assurer d'obtenir un maximum d'informations, mais aussi pour vérifier la véracité des renseignements trouvés, nous avons utilisé plusieurs sources (site des OSU et laboratoires, RNSR, ScanR ou encore AuréHAL).

Cette base de données se veut la plus complète possible. Ces informations serviront tout au long du projet et pourront même servir après, si Ad HALstra évolue.

Pour autant, certaines informations manquent. Malgré la pluralité des sources, toutes les informations pour chaque laboratoire n'ont pas pu être trouvées. Mais les principales informations permettant la construction des requêtes ont pu l'être. Ces données vont nous permettre de passer à la prochaine étape du projet, à savoir le test des requêtes pour l'extraction des notices d'ADS.

Les tests effectués pour l'extraction des métadonnées depuis ADS ont mis en évidence l'absence de certaines informations sur ADS. Nous avons ainsi pu constater que, si les laboratoires des sciences de l'univers sont quasiment tous représentés sur ADS, les affiliations des laboratoires n'étant pas spécialisés dans les domaines de la physique ou l'astrophysique sont moins renseignés. Les articles sont présents sur ADS mais sans affiliations auteurs dans de nombreux cas.

La création du projet

Création du site Ad HALstra

Le projet Ad HALstra a pris la forme d’un site web hébergé par l’Observatoire de Paris. Nous avons fait le choix de rendre Ad HALstra facile d'utilisation avec ce format connu de tous.

Le site Ad HALstra est le cœur même du projet. Il est l’interface qui va permettre aux utilisateurs de prendre en main de la façon la plus simple possible le projet Ad HALstra. C’est depuis le site que les utilisateurs vont pouvoir importer sur HAL les notices extraites d'ADS.

Le site propose trois fonctionnalités principales à l'utilisateur :

- Vérifier avant l'importation les correspondances des affiliations entre ADS et HAL (menu Conversion ADS -> HAL)
- Générer des fichiers au format TEI depuis le format REFABS XML d'ADS (menu génération des fichiers au format TEI)
- Importer massivement des notices depuis ADS vers HAL (menu Importation)

Ces fonctionnalités sont permises grâce à la création de tables de correspondance qui permettent de relier les affiliations auteurs d'ADS aux numéros de structure de HAL.

Les affiliations françaises de ADS nous ont été fournies par l'équipe de la NASA en charge de ADS. Les numéros de structure HAL ont été fournis par l'équipe du CCSD qui s'occupe de l'archive ouverte HAL. Des correspondances ont été créées entre les deux pour qu'ils puissent être associés.

L'outil Ad HALstra a été conçu pour reconnaître les affiliations françaises et, à partir de mots-clés associés à chaque affiliation, les faire correspondre à la bonne structure HAL. L'outil prend également en compte l'année en laquelle l'article a été édité. Si une affiliation a deux numéros de structure HAL, un actuel et un ancien fermé en 2019, et que l'article importé date de 2018, Ad HALstra va associer l'article au numéro de structure en vigueur en 2018, à savoir l'ancienne structure dans ce cas.

Dans un souci d'optimisation des résultats pour les laboratoires associés à l'INSU et aux OSU, nous avons fait une seconde vérification des résultats pour associer les affiliations françaises qu'Ad HALstra n'avait pas su identifier aux structures HAL correspondantes lorsque c'était possible.

La création du projet

L'utilisation de SWORD

Ad HALstra avait pour ambition première de se reposer en partie sur le protocole SWORD mis en place par le Centre de Communication Scientifique Directe. L'utilisateur était invité à utiliser Ad HALstra pour faire les conversions de notices de REFABS XML vers du XML TEI puis d'importer via SWORD les notices au format TEI de façon massive dans HAL.

Le CCSD a mis en place un protocole à partir de SWORD permettant l'importation de notices et de PDF dans HAL à partir de fichiers au format XML TEI. Ce protocole peut être utilisé grâce à son interface graphique ou via ses lignes de commande. Dans le cadre du projet, notre but est de faciliter au maximum l'importation des notices aux utilisateurs d'Ad HALstra. Il était donc important pour nous que l'importation se fasse avec l'interface graphique, accessible à tous, plutôt qu'avec les lignes de commande difficiles d'utilisation pour tout un chacun.

Mais, au cours du projet, nous nous sommes rendus compte que l'importation massive via SWORD ne pouvait se faire qu'avec les lignes de commande et non pas avec l'interface graphique. Pour remédier à cette difficulté, nous avons mis en place une nouvelle fonctionnalité dans Ad HALstra. Cette dernière fonctionne comme une boucle qui permet d'importer une à une les notices extraites d'ADS et ainsi permettre leurs importation massive.

Si une solution pour l'importation massive a été trouvée, malheureusement l'ambition d'Ad HALstra de permettre également l'importation des PDF associés aux notices n'est pas possible pour le moment. Le protocole SWORD ne permet d'importer les PDF que notice par notice et, dans le cadre d'Ad HALstra, nous souhaitions prendre le protocole SWORD comme appui à l'importation massive.

Communiquer autour du projet

A qui communiquer ?

Ad HALstra est un outil qui a été créé pour les laboratoires de l'INSU. La création des requêtes à leur intention a été faite dans le but de leur faciliter au maximum l'importation de notices HAL. La communication leur est donc en grande partie destinée. Les supports de communication seront envoyés à la fois à l'INSU, aux OSU et aux laboratoires pour avoir un impact maximal. Ensuite, à eux de décider s'ils souhaitent utiliser ces supports pour encourager au dépôt dans HAL leurs chercheurs.

L'équipe Ad HALstra a également décidé de communiquer auprès des partenaires du projet, à savoir le CCSD et ADS. Les deux structures ayant été à l'écoute des demandes entourant le projet, un mail de remerciement et de présentation leur a été adressé.

Comment communiquer ?

La communication autour du projet s'est faite à partir de plusieurs supports de communication qui ont été envoyés par mail aux laboratoires de l'INSU. Ces supports ont tous le même objectif : simplifier les dépôts de notices.

Pour cela, nous avons mis en place un flyer pour expliquer les motivations et le fonctionnement d'Ad HALstra. Un schéma présent dans le flyer est également disponible. Ce schéma a une vocation ludique. Il représente le fonctionnement de Ad HALstra sous la forme d'une machine conditionnant les notices pour qu'elles puissent être envoyées dans HAL, représenté par un camion.

Deux modes d'emploi seront également envoyés pour que les utilisateurs de Ad HALstra puissent prendre en main facilement l'outil et sachent rapidement faire les importations massives de notices ainsi que les importations à l'unité de PDF.

Enfin, la méthodologie est ici présente pour informer les utilisateurs d'Ad HALstra du cheminement qui a été à l'origine de la création de l'outil.

Une communication plus générale a également été faite à travers les présentations du projet dans la lettre de l'INSU de janvier ainsi que dans le Bulletin Interne de l'Observatoire de Paris (BIOP), le projet ayant été dirigé par la bibliothèque de l'Observatoire de Paris.

Conclusion du projet

Ad HALstra se veut être un projet conçu pour les chercheurs des laboratoires associés à l'INSU. Son but est de faciliter l'importation de métadonnées depuis la base de données ADS vers l'archive ouverte HAL.

Ad HALstra est un projet réalisé en sept mois mais il n'est pas figé pour autant. L'outil recensant de nombreux laboratoires, des évolutions sont à prévoir comme par exemple des fusions, changements de noms ou d'UMR. Le support technique ne pouvant pas être informé de tous ces changements, nous comptons sur la communauté de l'INSU et sur les laboratoires et OSU concernés pour nous faire savoir les modifications à apporter à l'outil en cas de besoin pour qu'Ad HALstra soit toujours le plus performant possible.

L'équipe Ad HALstra

AD HALSTRA