Diplôme national de master

Domaine - sciences humaines et sociales

Mention – sciences de l'information et des bibliothèques

Parcours – politique des bibliothèques et de la documentation

Mémoire de fin d'études / août 2023

Quand l'Univers nous ouvre sa science : la gestion des données de la recherche astronomique

Eléonore KOLAR

Sous la direction de Frédéric Saconnet-Girszonas
Directeur adjoint – Bibliothèque de l'Observatoire de Paris, site Meudon

Remerciements

Je tiens tout d'abord à remercier mon directeur de mémoire Frédéric Saconnet-Girszonas pour sa présence et son accompagnement éclairant tout au long de l'année.

Je remercie l'équipe pédagogique de l'ENSSIB pour leur suivi et leur disponibilité au cours de ma scolarité.

Mes remerciements vont également à celles et ceux qui ont donné de leur temps pour s'entretenir avec moi. Ces échanges ont été riches et déterminants pour l'écriture de ce mémoire.

Enfin je remercie mes proches, mes amies et mon compagnon dont le soutien et la présence sont précieux.

Résumé :

En tant qu'objets complexes, les données de la recherche demandent une gestion particulière. De leur production à leur réutilisation, les professionnels de l'information scientifique et technique sont sollicités à chaque étape de leur traitement, dans un dialogue constant entre divers corps de métiers et institutions. Ce mémoire s'attache à poser un regard interrogateur sur la place qu'ils et elles occupent dans l'écosystème de la production des données et plus particulièrement des données de la recherche en astronomie et astrophysique. Dans ce domaine chaque problématique est étirée et offre matière à penser les défis à venir dans la gestion des données de la recherche.

Descripteurs :

Données – Données de la recherche – Science ouverte – Astronomie – Astrophysique – Information scientifique et technique – Bibliothèques universitaires et de recherche

Abstract :

As complex objetcs, research data require special management. From their production to their reuse, scientific and technical information professionals are called upon each step of their treatment, in a constant dialogue between various trades and institutions. This work examines the place professionals have in the data production ecosystem, especially in the astronomical and astrophysical research data. In this field each question is deployed and gives food for thought on future challenges in research data management.

Keywords :

Data – Research data – Open science – Astronomy – Astrophysics – Scientific and technical information – Academic and Research libraries

Droits d'auteurs

Sommaire

Sigles et abréviations

ADS Astrophysics Data System

AFNOR Association Française de NORmalisation

ANR Agence Nationale de la Recherche

BIS Bibliothèque Interuniversitaire de la Sorbonne

CCSD Centre pour la Communication Scientifique Directe

CDS Centre de Données Stellaires / Astronomiques

CNRS Centre National de la Recherche Scientifique

CRAC Compte Rendu annuel d'Activité des Chercheurs

DMP Data Management Plan

ESA European Space Agency

ESO European Southern Observatory

FAIR Findable, Accessible, Interoperable, Re-usable

FITS Flexible Image Transport System

HST Hubble Space Telescope

IFLA International Federation of Library Associations and Institutions

IGESR Inspection Générale de l'Éducation, du Sport et de la Recherche

INSEE Institut National de la Statistique et des études économiques

INSU Institut National des Sciences de l'Univers

IRAP Institut de Recherche en Astrophysique et Planétologie

IST Information Scientifique et Technique

IVOA International Virtual Observatory Alliance

JWST James Webb Space Telescope

LIRIS Laboratoire d'InfoRmatique en Images et Systèmes d'information

LSST Large Synoptic Survey Telescope / Observatoire Vera-C.-Rubin

NASA National Aeronautics and Space Administration

OAI-PMH Open Archives Initiative Protocol for Metadata Harvesting

OCDE Organisation de Coopération et de Développement économiques

OV Observatoire Virtuel

PGD Plan de Gestion de Données

RDA Research Data Alliance

RIBAC Recueil d'Informations pour un oBservatoire des activités de recherche en Sciences humaines et sociales

SCD Service Commun de la Documentation

SDSS Sloan Digital Sky Survey

VISTA Visible and Infrared Survey Telescope for Astronomy

INTRODUCTION

La première étape d'un travail de recherche autour des données astronomiques a été de partir de l'existant et d'avancer en allant de plus en plus précisément vers l'objet de la recherche qui est très spécifique. Partir du général pour aller au particulier en éclairant les points saillants de cette avancée. En effet faire le choix de travailler sur les données en astronomie relève d'un certain défi.

L'astronomie recouvre des champs disciplinaires pointus. C'est un terme générique qui recouvre des champs disciplinaires distincts, se divisant en trois branches : l'astrométrie qui étudie le positionnement et le mouvement des astres, la mécanique céleste qui traite des lois du mouvement des astres et enfin l'astrophysique qui cherche à comprendre la constitution, le fonctionnement et l'évolution des astres observés en s'appuyant sur les principes de la physique. Ainsi il découle de ces branches des sous champs tels que la physique solaire et stellaire, la planétologie, la physique du milieu interstellaire, la cosmologie, l'astérosismologie, la physique galactique, l'astrobiologie, l'instrumentation. De toutes ces disciplines dérive un nombre important d'autres champs disciplinaires car ce secteur est aussi constitué de chercheurs en biologie, physique, mécanique, mathématique et informatique, de techniciens et d'ingénieurs.

Ma première interrogation a été : où se situent les professionnels de l'information scientifique et technique dans ces branches ? Les bibliothécaires et documentalistes ont-ils un rôle, un espace dans ce kaléidoscope de domaines et de professions ? Les bibliothèques universitaires, dans leur rôle de soutien à la recherche et valorisation du travail scientifique y ont-elles leur place ?

Les bibliothèques universitaires ont un rôle important dans la recherche en ce qu'elles concentrent un certain nombre de services auprès des chercheurs. En effet des postes sont dédiés à l'accompagnement des chercheurs quant à l'utilisation des outils numériques dans la démarche de partage scientifique. L'écriture d'un plan de gestion de données est devenue un élément incontournable dans la recherche, et les bibliothèques interviennent de plus en plus à travers la science ouverte. Elles deviennent un acteur incontournable de la science ouverte et notamment en astronomie : c'est un domaine avec beaucoup de données à traiter et les appels à projets demandent une expertise documentaire dans le traitement des données. Nous pouvons constater le rôle grandissant de la bibliothèque et des personnels de l'information scientifique et technique dans la recherche. Ces personnes deviennent des interlocuteurs face au chercheurs. C'est en ce sens que le métier évolue, et c'est en ce sens que se concentrer sur l'astronomie permet de montrer l'évolution de ces métiers vers des problématiques actuelles.

Travailler à mettre en lumière le rôle des documentalistes et des bibliothèques dans la recherche en astronomie est une façon pertinente d'approcher et questionner les prochaines évolutions du secteur des bibliothèques universitaires. En effet, comment les professionnels de l'information scientifique et technique interviennent-ils sur le plan de la gestion des données de la recherche en astronomie, à l'heure d'une production scientifique exponentielle ? Comment ce questionnement permet d'entrevoir l'évolution de ces métiers ? Dans un premier temps j'ai axé mes recherches sur la collecte de sources traitant des données astronomiques : quelles sont-elles ? où sont-elles produites ? qui les produit ? Puis, dans une perspective documentaire les questionnements sur la gestion de celles-ci sont rapidement venus

à moi : comment traite-t-on ces masses de données ? comment sont-elles partagées et réutilisées ? J'ai réalisé des recherches documentaires en croisant les sources des institutions astronomiques avec des ouvrages et articles traitant plus généralement des données de la recherche et de la science ouverte.

Dans un prisme d'interrogation, se questionner sur la place du documentaliste et des services qu'on peut offrir auprès des chercheurs allait en grandissant, car peu voire aucun ouvrage ne traite de la place des documentalistes dans la gestion des données de la recherche en astronomie. Et cette problématique piquait ma curiosité car un panel important de corps de métiers sont sollicités pour les projets astronomiques. Le point essentiel de cette première approche épistémologique a donc été de comparer les recherches sur les données et sur les données en astronomie, ainsi que sur la gestion documentaire. Sur ce dernier point, les entretiens ont été essentiels.

En effet dans un second temps, il a fallu aller à la rencontre de ces communautés techniques et scientifiques par le biais d'entretiens semi-directifs. J'ai fait le choix d'entretiens car j'avais besoin de collecter des données qualitatives, et ma curiosité sur le sujet m'a poussée à aller à l'encontre des personnes interrogées pour creuser en profondeur mes questionnements. Les entretiens offrent un espace propice à l'échange et à la discussion. Ils permettent de faire le lien entre les données brutes issues des sources et les données quantitatives issues des ouvrages.

Ces entretiens ont été réalisés en direction de trois types de publics : les professionnels de l'information scientifique et technique, le personnel de direction des grandes institutions, les chercheurs. J'ai effectué huit entretiens d'une durée comprise entre 45 minutes et une heure. Ces entretiens ont fait l'objet d'une retranscription partielle de ma part pour que je puisse être en mesure de tirer les informations nécessaires à ma recherche. Lors de ces entretiens, j'ai également appliqué la méthode de l'entonnoir (aller du général vers le particulier) pour recueillir un maximum d'informations précises, en commençant souvent par des questions génériques : comment définiriez-vous une donnée de la recherche en astronomie ou astrophysique ? De quelle nature est votre lien avec les chercheurs/les documentalistes, ou encore les institutions dans votre travail quotidien ? Comment êtes-vous arrivé à votre poste actuel ?

Les réponses à ces questions ont nourri en profondeur mon travail et j'ai pu articuler ces informations avec ma problématique pour aboutir à un découpage qui venait directement y répondre. En effet je me demande comment se positionner en tant que professionnel de l'information scientifique et technique dans la chaîne de gestion des données de la recherche.

Tout d'abord en considérant que les données sont des objets complexes. Pourquoi ? car leur nature ainsi que leur production sont complexes. Des régimes de valeurs leurs sont associés, tant juridiques que personnels (de la part de ceux qui les produisent). Aussi la production des données de la recherche est à la croisée de demandes institutionnelles et des pratiques des chercheurs qui ne sont pas forcément alignés avec ces demandes. Ainsi, cet écosystème complexe s'illustre dans la gestion des données astronomiques. Par ailleurs, s'est posée très tôt la question de la gestion des données dans cette discipline et de la mise en place d'outils performants.

J'ai souhaité aborder la gestion des données par cette science qu'est l'astronomie, car je nourris un intérêt personnel profond pour les thématiques de l'Univers et qu'il m'a semblé intéressant de poser des problématiques actuelles sur une discipline si singulière. C'est enfin l'occasion de (re)penser les relations entre

les différents acteurs de la recherche, et de réfléchir à comment chaque compétence apportée nourrit un dialogue profondément scientifique : ouvert et partagé.

I) LES DONNEES DE LA RECHERCHE : UNE PRODUCTION SCIENTIFIQUE SINGULIERE

1) DONNEES DE LA RECHERCHE : ETAT DES LIEUX

1. Définitions des données de la recherche

Une donnée est avant tout un objet informationnel c'est-à-dire un objet qui permet de construire l'information. C'est somme toute la matière première de l'information, le matériau de base avec lequel on va pouvoir constituer une information. Dans le domaine des technologies de l'information et de la communication, la donnée permet une description élémentaire d'un objet, d'un événement, d'une transaction. Elle est une matière « brute » qui va permettre de construire d'abord un contexte puis une recherche.

Les données se définissent également en trois grands types qui correspondent à des étapes de leur traitement :

Figure 1 Schéma par trois grandes typologies de données. Source © CCSD, CNRS

On trouve dans un rapport de recherche produit par le Service Commun de la Documentation (SCD) de l'Université Bordeaux Montaigne une définition des données de la recherche comme étant :

« […] toutes les productions, numériques ou non, collectées et réalisées par les chercheurs en amont de leur travail d'écriture proprement dit et à partir desquelles ils bâtissent leurs hypothèses. Les données de la recherche regroupent donc un ensemble hétéroclite de sources et matériaux de recherche, aussi appelés données primaires, ainsi que toutes les formes de traitement et d'analyses desdites sources, que l'on appelle alors des données dérivées. »[1]

Cette définition met l'accent sur le caractère généraliste des données de la recherche : « Toute production numérique ou non ». Si ce qui nous intéresse ici est le traitement numérique des données, il reste important de souligner qu'une donnée n'est pas nécessairement sous ce format, elles peuvent être des objets physiques comme des documents papiers, des ouvrages, des matériaux organiques ou non

[1] DUPRAT, Julie. *Les données de la recherche à l'Université Bordeaux Montaigne : Synthèse d'une enquête qualitative auprès des chercheurs* [en ligne]. 2019 Disponible sur : ⟨hal-02020141⟩

organiques, vivants ou non vivants. Nous pouvons également en donner une typologie. Les données de la recherche sont souvent classées en 5 types :

Typologie des données de la recherche :

- **Données de références** : données extraites, triées, agrégées. Ce sont des jeux de données revus par les pairs et mis à disposition

Exemples : base de données de cristallographie, collection de lettres ou archives d'images historiques

- **Données expérimentales** : données qui sont le résultat d'un travail en laboratoire, obtenues à partir d'équipement de laboratoire. Ces données sont reproductibles, bien que coûteuses à reproduire.

Exemples : puces à ADN, calculs, chromatographies

- **Données d'observations** : données capturées en temps réel, ces données sont uniques et non reproductibles.

Exemples : photographies de phénomènes astronomiques, neuroimageries, données d'observations sur des espèces animales ou végétales

- **Données dérivées** : données qui sont le résultat d'un traitement ou de la combinaison de données "brutes". Ces données peuvent être reproductibles bien que coûteuses.

Exemples : bases de données compilées, données issues de fouilles de texte

- **Données de simulation ou données computationnelles** : données générées à partir de modèles informatiques ou simulés. Elles peuvent être reproductibles si le modèle est précisément documenté.

Exemples : modèles de simulations sismiques, météorologiques, simulations cosmologiques, modèle économique

Cette matérialité plurielle de la donnée issue de la recherche est à l'image de la production scientifique hétérogène et qui couvre tous les domaines. Dans la définition produite par le SCD de l'Université Bordeaux Montaigne se trouve déjà deux types de données : les données primaires qui s'apparentent à la matière brute, et les données dérivées qui sont un premier traitement de la donnée afin de la rendre lisible. Les « formes de traitement » mentionnées dans la définition font déjà appel à un choix sur comment traiter la matière première récoltée et comment analyser la source d'où elle provient.

Pour l'OCDE, la matière brute des données est indissociable de l'utilisation qui va en être faite. En 2007, elle avance que les données de la recherche « […] sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont

généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de recherche. »[2].

Ces deux définitions décrivent les données de la recherche comme étant imbriquées dans un tout informationnel comprenant un contexte scientifique et une démarche épistémologique. Les données de la recherche sont collectées par des chercheurs et servent, plus encore qu'à bâtir une hypothèse, à pouvoir valider ou non les résultats de la recherche. Les données sont la source de toute recherche, ce qui permet de construire le raisonnement scientifique. On peut dès lors affirmer que les données de la recherche, loin d'être produites ex-nihilo, sont indissociables des organismes de recherche et de la communauté des chercheurs qui les produit. On retrouve cette idée dans le Courrier des statistiques N°5 de l'INSEE, qui cette fois-ci sur les « données » en tant que telles, insiste sur l'importance des choix dans la production et le traitement des données :

« Toute donnée se caractérise par un vaste faisceau de conventions (sémantique, nomenclatures, formats, etc) et par l'infrastructure de connaissances dans laquelle elle s'inscrit, impliquant des choix qui n'ont rien de neutre. Une donnée se révèle aussi dépendante de l'environnement qui lui a donné naissance, et des processus productif qui l'utilisent. On constate alors que les données ne sont pas pures et parfaites, ne vont pas de soi : paradoxalement les données ne nous sont pas données »[3].

Par conséquent, ces définitions mettent d'ores et déjà l'accent sur la façon dont les données mettent en évidence les liens étroits qui unissent tous les acteurs du traitement de l'information, de sa production à sa diffusion. Comment définir alors un paramètre exact des données de la recherche ? Cette dernière définition des données issue du courrier des statistiques de l'INSEE apporte un regard paradoxal : les données ne nous sont pas données. Elles émaneraient déjà de choix, ici pour les données de la recherche, des politiques publiques de l'Enseignement supérieur et la recherche en matière de production, de gestion et de partage du savoir scientifique. Ce « vaste faisceau de conventions » fait ici référence à des instances de recommandation pour les normes et les standards : par exemple l'AFNOR et les normes relatives à la conception et à l'exploitation des systèmes informatiques[4] Force11[5] et les principes FAIR[6], OpenDataFrance[7], OPIDoR[8], ou encore l'IVOA

[2] OCDE, *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*, p.18 [en ligne]. 2007 Disponible sur : https://www.ouvrirlascience.fr/wp-content/uploads/2018/11/Principes-lignes-directrices-de-l%E2%80%99OCDE-pour-l%E2%80%99acc%C3%A8s-aux-donn%C3%A9es_38500823.pdf

[3] RIVIERE, Pascal. *Courrier des statistiques N5, 2020, « Qu'est-ce qu'une donnée ? ». 2020. Disponible sur : https://www.insee.fr/fr/information/5008707?sommaire=5008710*

[4] AFNOR, Norme NF 3 Z42-013 *Recommandations relatives à la conception et à l'exploitation de systèmes informatiques en vue d'assurer la conservation et l'intégrité des documents stockés dans ces systèmes*

[5] *The Future Of Research Communication and e-Scholarship* : Site du groupe de recherche disponible à cette adresse : https://force11.org/info/the-fair-data-principles/

[6] Cf note 5

[7] Association française créée en 2013 et financée par l'Union Européenne qui a pour objectif de regrouper et soutenir les collectivités territoriales françaises dans leur démarche d'ouverture de leurs données. L'association œuvre dans un but pédagogique en collectant et produisant des ressources auprès de leurs publics. https://www.opendatafrance.net/

[8] Optimiser le Partage et l'Interopérabilité des Données de la Recherche : portail qui met à disposition de la communauté de l'Enseignement Supérieur et de la Recherche des outils et de services pour accompagner la gestion des données de la recherches. Différents services sont proposés tels que DMP OPIDoR pour élaborer des plans de gestion de données, Cat OPIDoR qui est un wiki des services dédiés aux données de la recherche et PID OPIDoR qui est un service d'attribution d'identifiants pérennes.

pour les données recueillies par les observatoires astronomiques. Force 11 nous donne un exemple intéressant sur le processus de rédaction dont sont issues certaines normes. En effet Force 11 est une communauté de chercheurs bibliothécaires, archivistes, éditeurs et professionnels de l'information scientifique qui œuvrent pour améliorer le partage des connaissances savantes grâce à l'utilisation des technologies de l'information. Ils ont activement participé à l'élaboration de principes d'échange de données en 2014 qui irriguent aujourd'hui la communauté de la recherche : les principes FAIR (pour *Findable, Accessible, Interoperable, Re-usable*).

Selon ces principes, toute donnée devrait pouvoir être facilement trouvée, accessible, interopérable et réutilisable. Or, comme nous l'avons vu, le processus de traitement des données ne va pas de soi : il dépend de choix des chercheurs. On peut distinguer les normes préconisées et les habitudes des chercheurs. Ce processus de production des données est empreint d'un système de valeur, d'une axiologie. Une donnée doit être trouvable ou accessible, mais elle ne l'est pas par définition. C'est pourquoi dans la production et le traitement des données, l'intermédiaire des professionnels de l'information scientifique et technique (IST) va jouer un rôle majeur. C'est grâce à leur expertise que les données peuvent entrer dans les normes qui sont préconisées. Aussi l'expertise des professionnels de l'IST est importante lorsqu'ils se font les interlocuteurs face aux chercheurs et qu'ils sont en mesure de comprendre et recueillir leurs besoins.

En outre, les quatre piliers des principes FAIR exposent en un sens les directions à prendre quant aux politiques d'ouverture de la science. Les données doivent pouvoir être *findable*, c'est-à-dire faciles à trouver. Il est vrai qu'on ne peut commencer un travail de partage sans être en mesure de trouver l'objet ou l'information à partir desquels ce travail commencera. Le fait que l'on puisse trouver les données, savoir où elles sont entreposées pose en filigrane la question des entrepôts de données, de la qualité et de la fiabilité de ceux-ci. Pour ce faire, les consortiums scientifiques nationaux ou supranationaux en matière de science ouverte œuvrent à recueillir les besoins des chercheurs pour une meilleure gestion de l'information, et ce de manière indépendante du monde de l'information scientifique et technique.

En somme, bien que les normes définies par le milieu professionnel de l'IST soient un objectif à atteindre, il est important que les communautés scientifiques réfléchissent elles-aussi à comment elles souhaitent que leurs données soient gérées.

2. Cadre institutionnel et juridique des données de la recherche

L'ouverture des données de la recherche est à replacer dans un contexte national et plus général d'ouverture de la science à grande échelle. Tout d'abord en 2011 une politique d'ouverture des données au niveau des administrations centrales a été menée par divers gouvernements successifs pour aboutir à une circulaire concernant les ministères. En effet les ministères ont pour obligation de diffuser leurs données sur le portail *data.gouv.fr* en apposant une licence ouverte aux jeux de données déposés.[9] Ce mouvement d'ouverture a été prolongé en 2014 par le

[9] Décret et circulaire du 26 mai 2011 : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000024072788

Ministère de l'Enseignement supérieur par l'ouverture de la plateforme *Open Data*. Or les initiatives d'ouverture des données, d'abord ministérielles puis de recherche, se heurtent à la politique des établissements de recherche et à la communauté des chercheurs. Une modification des habitudes de travail pointe avec ces lois et recommandations car cela demande une certaine acculturation à la gestion de données. Par exemple il faut être en mesure de comprendre les licences ouvertes, les plans de gestion de données ou encore les processus de collecte automatisée des données. Un dialogue doit être bâti entre chaque acteur afin de faciliter ces démarches d'ouverture. Et en ce qui concerne les données de la recherche l'acculturation à la gestion des données par les chercheurs se heurte également à une réticence de leur part quant à l'ouverture de leurs données.

A titre d'illustration, les chercheurs et scientifiques rattachés à des laboratoires de recherche publics en France font en effet partie des rares agents de l'Etat à conserver l'intégralité de leurs droits d'auteurs sur leurs productions. Il paraît alors difficilement concevable que les données qu'ils produisent dans le cadre de leur recherche puissent être mise à disposition de manière ouverte et libre. En effet cette crainte de la part des chercheurs de mettre à disposition leurs données a été soulignée par Sylvie Retailleau, alors ministre de l'Enseignement supérieur et de la Recherche lors de l'inauguration de la plateforme *Recherche Data Gouv* en 2021 :

> « J'entends encore, parmi certaines communautés scientifiques, la crainte d'être « pillé » en ouvrant ses données. Un article enrichi par les données voit pourtant son taux de citation augmenter en moyenne de 25%. Une science plus ouverte, c'est une science qui est plus visible, qui rayonne davantage, c'est donc une recherche française reconnue à sa juste valeur et qui contribue à construire les solutions aux défis de notre temps. »

Transparaît dans ce discours la volonté du gouvernement d'« avertir » la communauté des chercheurs sur les dispositions que souhaite prendre l'Etat en matière d'ouverture des données. Toutefois, cette crainte relayée par un discours officiel ne prend peut-être pas la mesure de la réticence de la communauté, ou simplement de l'ignorance de celle-ci sur les enjeux d'une ouverture. Dans une étude menée par Violaine Rebouillat en 2019[10], à la question de savoir « Quelle valeur ont les données que vous avez générées ? », un chercheur a répondu *:* « Ce sont nos enfants ! ». Un autre de répondre que ses données sont « précieuses » dû au coût, à la rareté et à l'effort de production qui sont mis en œuvre pour qu'elles voient le jour. Ce niveau d'attachement à la production des données démontre un lien affectif important entre le chercheur et ses données. Ce lien est comme une continuité du travail subjectif déployé dans un projet de recherche. Violaine Rebouillat évoque même une « intimité créative » du chercheur avec son objet de recherche dont les données sont le produit. Cette intimité ne peut donc pas être ignorée de la part des mouvements d'ouverture et cela pose une réticence réelle de la part des chercheurs face à ce que certains perçoivent comme une injonction à l'ouverture.

Il est également à prendre en compte que le périmètre des données de la recherche, comme nous l'avons vu précédemment, recouvre un vaste champ de définitions complexes qui mettent l'accent sur le caractère institutionnel de la

[10] REBOUILLAT Violaine, « Le partage des données vu par les chercheurs : une approche par la valeur », *Les Enjeux de l'information et de la communication*, 2021/1 (N° 22/1), p. 35-53. DOI : 10.3917/enic.030.0035. URL : https://www.cairn.info/revue-les-enjeux-de-l-information-et-de-la-communication-2021-1-page-35.htm

production des données. Les données ne sont pas produites uniquement par un « auteur » ou un chercheur, elles sont produites par un tout institutionnel qui permet sa production : le financement public, le laboratoire, les outils mis à disposition du chercheur, la gestion et le stockage de ces données, et enfin leur réutilisation. En effet, « Pourquoi nous mobilisons-nous tant pour la science ouverte ? » questionne Sylvie Retailleau ?

« […] C'est peut-être une évidence mais nous touchons là au fondement même de la recherche publique : menée sur fonds publics, elle doit revenir au public, à tous les publics. »

Ainsi l'ouverture des données de la recherche questionne comme on le voit la responsabilité du chercheur dans la production de ses données, et qui se heurte à la valeur qu'il leur consacre. On pourrait dire qu'on ne produit pas de la science ex-nihilo et que tout est à replacer dans une chaîne de production. Le suivi de cette production, notamment à travers le plan de gestion de données requalifie les étapes de cette gestion et place différemment le chercheur. Cependant, cet aspect technique pose une réticence car c'était jusqu'à présent le rôle des professionnels de l'IST de remettre cette chaîne de production en forme. Lors d'un entretien, [Nathalie Pothier, responsable du portail HAL de l'INSU] me confie à propos des publications scientifiques :

« Je comprends la place du chercheur réticent aux dépôts, dont la mission première n'est pas de faire des rapports […], on demande peut-être trop de choses aux chercheurs. »

Cette remarque peut tout à fait être reportée sur l'ouverture des données de la recherche. Pour aller plus loin encore, selon cette interlocutrice : « Le monde de demain tournera autour des données, les évaluations se feront autour des données et non plus des publications. ». On perçoit alors l'importance de la médiation et de l'accompagnement aux chercheurs dans la gestion de leurs données. Ce rôle des médiateurs de l'information ne doit ni couper le lien affectif du chercheur à sa production, ni omettre de transmettre les valeurs attachées à l'ouverture et au libre accès.

Nous l'avons vu, un mouvement profond s'est initié en faveur du libre accès des données ministérielles (dans la mesure de la confidentialité de certaines données sensibles ou classées secret défenses) et les données de la recherche font à présent partie de ce périmètre. Les données de la recherche étant spécifiques et parfois sensibles ou soumises à la volonté légitime du chercheur qui les a produites, le travail va être d'acculturer la communauté scientifique à la science ouverte, à ses aspects techniques, et aux valeurs progressistes sur lesquelles elle repose. C'est là que les professionnels de l'IST vont pouvoir agir, en exposant les avantages liés à l'ouverture des données de la recherche, en composant avec la réserve d'une partie de la communauté scientifique.[11]

On peut ici citer qu'avant la mise en place de *Recherche Data Gouv*, est promulguée en 2016 la loi pour une République numérique qui crée l'obligation par les pouvoirs publics de communiquer gratuitement en ligne leurs bases de données. S'en est suivi le premier Plan national pour la Science ouverte lancé le 4 juillet 2018

[11] YOUNG, Andrew, Stefaan G. VERHULST et Andrew J. ZAHURANEC. Comment la science ouverte peut s'inspirer du libre accès aux données publiques. The Conversation [en ligne]. 16 mars 2021. Disponible sur : https://theconversation.com/comment-la-science-ouverte-peut-sinspirer-du-libre-acces-aux-donnees-publiques-157091

par Frédérique Vidal, alors ministre de l'Enseignement Supérieur et de la Recherche, qui marque un tournant officiel dans la politique publique de l'enseignement supérieur et de la recherche en France. Par ces mesures il s'agit de généraliser l'accès aux publications et d'inscrire la science ouverte comme dynamique fondamentale et internationale de la recherche en France.

Ce plan, d'une durée de trois années a développé un axe dédié aux données de la recherche et rend obligatoire l'ouverture des données. Il fait également appel aux principes FAIR quant à la structuration des données de la recherche et promeut la création d'une fonction d'administrateur des données de la recherche dans les établissements. Ainsi l'un des objectifs est d'inscrire la science ouverte dans les établissements par la création d'un label « Science ouverte » dans les écoles doctorales et de proposer aux chercheurs des formations adaptées à leurs besoins. Ces mesures sont prises pour coordonner au mieux les établissements et créer un réseau d'aide à la recherche. En 2021, est promulgué le deuxième Plan national pour la Science Ouverte qui devrait se terminer en 2024. Il vient renforcer les initiatives lancées lors du premier. En effet le Ministère de l'enseignement supérieur et de la recherche cherche à généraliser l'ouverture et structurer la mise en place des plans de gestion des données associées à des projets de recherche financés sur fonds publics. Notons que pour cela, l'intermédiaire des professionnels de l'IST est nécessaire :

« L'information scientifique et technique (IST) est au cœur de l'activité de recherche. […] L'I.S.T. regroupe ainsi l'ensemble des informations produites par la recherche et nécessaires à l'activité scientifique et intervient donc en amont et en aval dans tout le cycle de production de nouveaux contenus scientifiques quelle que soit leur forme : articles, données, ouvrages, archives ouvertes, etc ».[12]
Par conséquent, les professionnels de l'IST interviennent à chaque étape de production scientifique, et par là à chaque étape du cycle de vie des données de la recherche, de l'élaboration du plan de gestion de données à l'archivage et la réutilisation.

[12] Informations disponibles sur : https://www.enseignementsup-recherche.gouv.fr/fr/information-scientifique-et-technique-51161

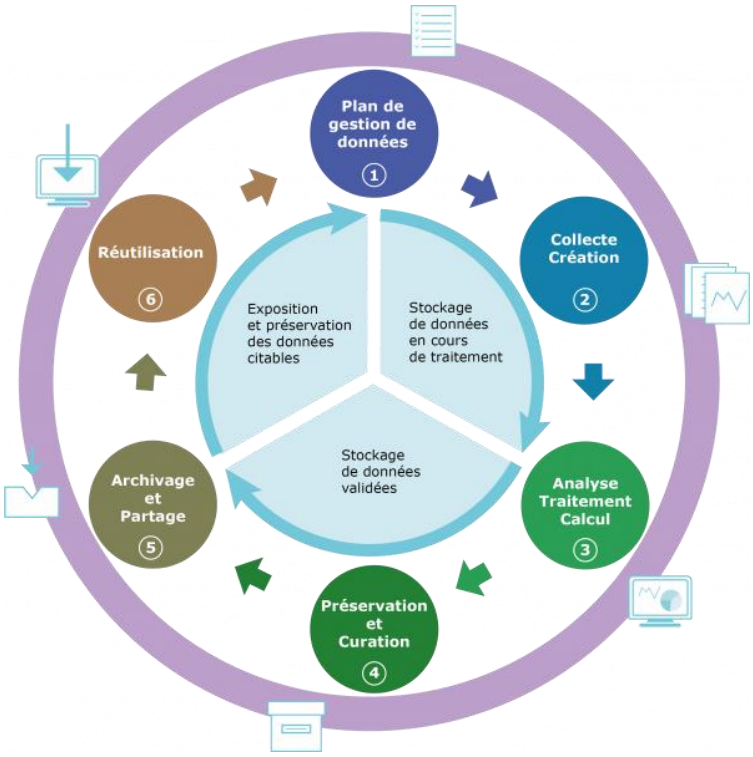

Figure 2 Le cycle de vie des données de la recherche. CC-BY Nicole Lambert / GRICAD / CNRS

3. Quel cadre international pour l'ouverture des données ?

Nous pouvons dès à présent noter qu'il n'existe pour l'instant pas de cadre international strictement juridique ou institutionnel sur l'ouverture des données, il n'existe pas d'instance européenne ou internationale qui fait foi en matière d'ouverture des données. Cependant, la présence de la Research Data Alliance (RDA) depuis 2013 pour la communauté européenne, qui vise à promouvoir le partage et l'échange de données via l'élaboration de recommandations à visée internationale joue un rôle important. Dans un cadre international, les bibliothèques ont également une place de plus en plus importante : par exemple, DataCite qui est un consortium international de bibliothèques qui propose des services autour de l'archivage numérique et des données. En effet en 2004, le comité de la politique scientifique et technologique de l'OCDE publie une déclaration en faveur de l'implication des établissements dans l'ouverture des données issues de projets publics.[13]

C'est en 2019, soit plus d'une décennie plus tard, que l'Agence Nationale de la Recherche (ANR) rend obligatoire le plan de gestion de données (PDG) pour tout projet de recherche financé sur fond public. Enfin dans un décret du 3 décembre

[13] OCDE. *Making Open Science a Reality, OECD Science, Technology and Industry Policy Papers, No. 25* [en ligne]. 2015. Disponible sur : http://dx.doi.org/10.1787/5jrs2f963zs1-en [consulté le 25 octobre 2022]

2021, la science ouverte et la gestion des données sont reconnues comme des questions d'intégrité scientifique. En l'espace de 15 ans, on est passé d'un avis favorable des institutions à prendre en compte l'intérêt de l'ouverture des données à des mesures d'obligation de gestion de celles-ci. Ces décisions viennent bousculer les compétences des professionnels de l'information scientifique. Dans ces avancées, les bibliothèques universitaires et de recherche, en tant qu'institutions au service de la recherche, sont consultées quant à l'établissement des standards documentaires. En effet les données de la recherche sont produites et saisies à travers des valeurs telles que l'ouverture et la réutilisation. Or ces valeurs reposent sur des outils (bases de données, plans de gestion de données, etc) et des techniques documentaires particulières. Quelle expertise peuvent alors apporter les bibliothèques et les universités ?

2) DE LA SENSIBILISATION A L'INVESTISSEMENT DES UNIVERSITES

1. Sensibilisation, accompagnement des chercheurs, tutoriels…

Les universités et les bibliothèques universitaires œuvrent à disséminer des bonnes pratiques de gestion numérique de la production scientifique au sein de leur établissement : ces nouvelles missions « science ouverte » sont au cœur des nouvelles problématiques des bibliothèques et des universités, et de la gestion des données de la recherche. De quelle manière ?

L'ouverture de la science a un impact réel sur les politiques des bibliothèques universitaires, tant sur le plan financier que sur l'organisation interne des bibliothèques. En février 2021, un rapport l'Inspection Générale de l'Education, du Sport, et de la Recherche (IGESR) sur la place des bibliothèques universitaires dans le développement de la science ouverte a été produit[14] et dresse un panorama de l'implication de ces établissements dans la science ouverte. Au sein de ce rapport on trouve les résultats d'une enquête menée du 29 avril 2020 au 13 juin 2020 auprès de 70 établissements dont 6 écoles d'enseignement supérieur, 4 établissements à statut particulier (dont l'Observatoire de Paris), 2 bibliothèques universitaires et 58 services communs de la documentation. L'objet de ce questionnaire est de chercher à savoir combien d'établissements sur ceux ciblés sont dotés d'une politique de science ouverte, si celle-ci est structurée à travers un plan et qui pilote cette politique au sein de l'établissement. Sur 62 bibliothèques interrogées, 88,6% ont répondu qu'elles sont associées à la définition de la politique science ouverte de leur établissement et sur 58 de ces bibliothèques 82,9% sont associées activement c'est-à-dire par la mise en œuvre de la politique science ouverte de leur établissement de rattachement.

De plus les missions confiées aux bibliothèques sont très variées et prennent des formes innovantes : un membre du personnel peut être désigné référent science ouverte. C'est une nouvelle compétence et spécificité du métier. Les missions

[14] LETROUIT, Carole, CACHARD, Pierre-Yves, DUPUIS, Monique, FROMENT, Bernard. La place des bibliothèques universitaires dans le développement de la science ouverte. Rapport à madame la ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation. N° 2021-2022. [en ligne] Février 2021. Disponible sur : file:///C:/Users/El%C3%A9onore%20Kolar/Downloads/igesr-rapport-2021-022-place-bibliotheques-universitaires-developpement-science-ouverte-pdf-88074.pdf

peuvent également prendre la forme de contrat d'objectif ou de projets de service. Par-là, une sensibilisation des établissements à l'égard des chercheurs est menée et vise à élaborer un dialogue dans un cadre ouvert comme contraint, de par la mise en place de l'obligation du PGD et du dépôt en archives ouvertes dans le cadre des évaluations des chercheurs (CRAC[15], RIBAC[16]).

Nous l'avons vu précédemment, les chercheurs ont pour certains un lien fort avec la production de leurs données et une sensibilisation aux politiques d'ouverture est alors nécessaire. Pour citer quelques exemples de ces projets de sensibilisation nous pouvons donner celui du SCD de l'Université de Toulouse Capitole qui propose un MOOC dédié à la science ouverte et à la publication *open access* en 2018 à partir de son compte Twitter qui s'intitulait « L'univers de l'Open Access ». Y sont présentés le cadre juridique et institutionnel mais aussi les éléments qui encadrent le dépôt d'une publication en archive ouverte. Également, le Centre pour la Communication Scientifique Directe (CCSD)[17] propose des webinaires « Parlons science ouverte » qui abordent une thématique liée à la science ouverte et qui sont l'occasion de se retrouver entre professionnels de l'IST et chercheurs. En outre, l'Université Paris Saclay propose sur son site des ressources liées cette fois-ci spécifiquement à la gestion des données de la recherche ayant pour thématique : comment produire des données FAIR ? Comment utiliser les entrepôts de données ? Dans ce même registre, la bibliothèque de l'université Lumière Lyon 2 met à disposition sur son site des ressources sur la gestion des données de la recherche. Un accompagnement individuel est proposé par la bibliothèque aux porteurs de projets ANR pour les aider dans leur rédaction d'un PGD. Idem pour la bibliothèque de Nantes Université qui propose une aide à la rédaction d'un PGD, des formations ainsi qu'un soutien au partage des données. Certaines bibliothèques tiennent des « guichets de la science ouverte » où sont abordées les questions liées à la gestion des données, c'est le cas de la bibliothèque de l'Université Grenoble Alpes ou encore de l'Université Paris 1 Panthéon Sorbonne, avec un projet de guichet dont font partie ses bibliothèques (SCD, BIS[18], CUJAS[19]).

Des tutoriels et des ressources sont mis à disposition des chercheurs afin d'aider à la rédaction du PGD. A titre d'exemple l'Observatoire de Paris s'est doté d'une charte des bonnes pratiques pour la science ouverte qui énonce point par point les missions à mener, tel le dépôt en archive ouverte des publications rattachées à l'Observatoire, l'élaboration de plans de gestion de données, l'ouverture des données et enfin la sensibilisation et la formation auprès des chercheurs sur les enjeux de la science ouverte. La bibliothèque de l'Observatoire est garante de ces bonnes pratiques et est chargée de répondre aux demandes de toutes et tous afin de rendre accessibles ces pratiques qui restent parfois complexes. Toute personne désireuse d'en savoir plus ou d'être accompagnée peut envoyer un message à l'équipe de la bibliothèque qui est désormais en mesure d'apporter des réponses sur

[15] Compte rendu annuel d'activité des chercheurs

[16] Recueil d'Informations pour un oBservatoire des activités de recherche en sciences humaines et sociales

[17] Centre pour la Communication Scientifique Directe

[18] Bibliothèque Interuniversitaire de la Sorbonne

[19] Bibliothèque interuniversitaire, spécialisée en droit, sciences économiques et sciences politiques, relevant d'une convention entre les deux universités Paris 1 et Paris 2

les plans de gestion de données, les archives ouvertes, ou encore les identifiants chercheurs.

2. Le plan de gestion de données : une étape essentielle

Le plan de gestion de données est un outil indispensable à une bonne gestion des données, il décrit l'ensemble des étapes du cycle de vie de la donnée au sein d'un projet de recherche. C'est un document évolutif dont trois versions sont attendues par les agences de financement. En effet les chercheurs doivent pouvoir produire un PGD au début du projet (dans les 6 premiers mois), au milieu du projet puis à la fin du projet. On décrit le PGD comme un outil d'anticipation des problèmes liés à la gestion des données pendant un projet de recherche, notamment des problèmes d'ordre juridique ou liés aux coûts des données.

Par exemple, la première version du PGD peut avoir décrit en termes de volume le nombre de données que produira le projet, or ce nombre, au milieu du projet a doublé. En ce cas, le poids des données à stocker sera différent de ce qui était indiqué au départ et il va falloir modifier cette information dans le PGD. Ainsi la rédaction d'un PGD demande aux chercheurs de se poser les bonnes questions dès le début et d'anticiper au mieux leurs besoins. Enfin, le PGD incite les chercheurs à déposer leurs données dans des entrepôts fiables et cela permet d'accroître la visibilité du travail des institutions autour des données de la recherche.

Quels éléments sont attendus dans un PGD ?[20] Dans outils d'aide sont mis à disposition des chercheurs, notamment l'outil DMP[21]OPIDoR[22] qui est un outil national émanant du CNRS et Argos, outil européen d'aide à la rédaction d'un PGD également. On peut lister environ 10 étapes importantes dans le PGD : la préparation du projet, les coûts, la collecte et la réutilisation des données, une documentation, le stockage et l'organisation, l'accès aux données, le partage et la publication des données, les solutions d'archivage, un point sur les aspects éthiques et juridiques et enfin les informations générales sur le projet.

L'importance donnée depuis 2019 aux PGD par les agences de financement de la recherche comme l'ANR pointe l'aspect fonctionnel et complet de ce document. Plus qu'un document juridique ou récapitulatif, le PGD aborde toutes les dimensions de la recherche, du financement à la gestion en passant par la pérennité du stockage des résultats. Par l'intermédiaire des bibliothèques aujourd'hui en mesure d'accompagner les chercheurs sur le PGD, l'objectif de transparence de la science ouverte devient concret. Les professionnels de l'information proposent des outils en ligne et des services en bibliothèque via les *learning center*[23]. Lilliad[24] de l'Université de Lille propose par exemple une aide à la rédaction de propositions de subventions ou encore à la gestion des données de recherche de projets.

[20] Cf Annexe 1 p.64 pour un exemple d'un PGD

[21] Data Management Plan – Plan de gestion de données en anglais

[22] DMP OPIDoR est un outil d'aide à la création en ligne de plans de gestion de données mis à disposition de l'Enseignement Supérieur et de la Recherche. Il est hébergé et géré par l'Inist-CNRS. Disponible sur : https://dmp.opidor.fr/

[23] Désigne un nouveau type de bibliothèques universitaires, proposant une offre de services enrichie. Pour plus d'informations voir : https://www.enssib.fr/le-dictionnaire/learning-center

[24] Learning Center de l'Université de Lille. Disponible sur : https://lilliad.univ-lille.fr/

Ainsi, c'est au sein des services à la recherche que les bibliothèques vont aider à la rédaction de ce type de document, ce qui paraît nécessaire quand lorsqu'en 2017 près de 33% des chercheurs annoncent n'avoir jamais entendu parler d'un PGD[25] et que plus de 80% des données produites sont stockées ailleurs que dans des entrepôts.[26] Ces chiffres évoquent une fracture entre la volonté forte d'ouvrir la science à ce qu'on appellerait une transparence de la recherche, et une communauté de chercheurs encore peu familière avec la gestion des données.

Dès lors, la façon dont est appréhendé l'ouverture des données peut être différente selon la culture scientifique à laquelle appartiennent les chercheurs. L'astronomie en ce sens est une discipline qui a une avance relative, car cette question du partage s'est posée tôt. Or de par la question du plan de gestion de données, qui est un nouvel outil, l'intermédiaire des professionnels de l'IST dans l'accompagnement comme dans la gestion des données est une étape nécessaire et fondamentale, tant pour la démarche scientifique qui se veut de plus en plus transparente, que pour faire face aux nouveaux paradigmes de la gestion de l'information.

3. Science ouverte et bibliothèques : quels enjeux pour les bibliothèques ?

On voit par là qu'une évolution s'est opérée au sein des bibliothèques universitaires car une partie des services se tournent vers la science ouverte et les bibliothécaires sont en mesure d'accompagner les chercheurs sur les problématiques liées aux données. En conséquence, un lien quasi direct est tissé entre la recherche et le corps des métiers documentaires. C'est par la science ouverte et l'ouverture des données qu'un pont se forme entre par exemple la recherche en astronomie et les bibliothèques. En effet une équipe de recherche en astronomie – au même titre que les autres disciplines – a l'obligation de rédiger un plan de gestion de données en amont de tout projet financé sur fonds publics européens. Si la science ouverte et l'ouverture des données demandent de nouvelles compétences côté bibliothèque, nous pouvons également avancer l'idée que les bibliothèques sont dans une continuité quant à leur rôle d'appui et de soutien à la recherche. En effet les bibliothèques ont pour mission essentielle de participer aux activités de formation et de recherche des établissements ainsi que l'énonce le Décret n° 2011-996 du 23 août 2011 relatif aux bibliothèques des établissements d'enseignement supérieur[27]. Le huitième point de l'article 2 va dans ce sens : les bibliothèques ont pour mission de « former les utilisateurs à un emploi aussi large que possible des techniques nouvelles d'accès à l'information scientifique et technique ».

L'implication des bibliothèques dans ce rôle passe notamment par leur participation à l'élaboration des standards pour les normes de catalogage des

[25] EUROPEAN COMMISSION. Providing researchers with the skills and competencies they need to practise Open Science. Report of the Working Group on Education and Skills under Open Science. [en ligne] Juillet 2017. Disponible sur : https://euraxess.ec.europa.eu/sites/default/files/policy_library/ec-rtd_os_skills_report_final_complete_2207_1.pdf

[26] Ibid.

[27] Décret n° 2011-996 du 23 août 2011 relatif aux bibliothèques et autres structures de documentation des établissements d'enseignement supérieur créées sous forme de services communs. [en ligne]. Disponible sur : https://www.legifrance.gouv.fr/loda/id/JORFTEXT000024497856

données, mais aussi par l'appropriation de tout un nouvel écosystème numérique dans lequel se meut la recherche à l'heure actuelle. Les bibliothèques travaillent tout d'abord à l'élaboration des standards comme en témoigne l'implication de ce corps de métier à travers l'International Federation of Library Associtions and Institutions (IFLA), instance internationale reconnue et consultée.[28]

On voit apparaître la notion de *data librarian*, que l'on traduirait grossièrement par bibliothécaire de la donnée. Son travail est d'accompagner le chercheur dans son projet de recherche[29]. Par le biais de services numériques, les bibliothèques vont à la rencontre des chercheurs ou bien se déplacent au sein des laboratoires afin de transmettre leurs connaissances. Dans ce contexte le bibliothécaire devient en quelque sorte médiateur documentaire ou numérique afin d'aider les chercheurs, ils se rendent à la source des publics scientifiques. Les établissements sont également en mesure d'accompagner sur les projets européens, projets qui depuis l'obligation de la publication du plan de gestion de données requièrent une connaissance et des compétences que les chercheurs peuvent ne pas avoir.

C'est effectivement à travers le plan de gestion de données que la passation de connaissance va se faire. En outre ceux-ci ont une portée symbolique forte concernant la transparence de la science : la question est désormais et pour la première fois posée de façon systématique : qu'allez-vous faire de vos données ? où allez-vous les collecter ? Il est à noter que dorénavant, le financement de l'ANR pour un projet de recherche est sous tendu par un PGD. Aller à la rencontre des publics de chercheurs devient alors un enjeu crucial pour les bibliothécaires.

Afin d'illustrer la complexité de ces dialogues et des compétences émergentes, nous nous intéresserons à la gestion des données de la recherche en astronomie et astrophysique qui sont un exemple particulier dans la communauté scientifique. Très tôt s'est posée la question de l'interopérabilité des systèmes, du traitement de donnée en masse, et de la collaboration entre chercheurs et professionnels de l'IST afin de comprendre les besoins de la communauté scientifique. Premièrement les données astronomiques sont des objets complexes dont le traitement demande une expertise scientifique et documentaire. Pour cela nous avons tout d'abord besoin de les caractériser afin d'expliciter leur complexité. Deuxièmement, les outils mis en place pour leur traitement sont des outils dont la performance tient dans l'implication de tous les acteurs : documentalistes, chercheurs, informaticiens. Enfin, cela permet d'entrevoir les nouvelles façons de travailler en bibliothèque comme en centre de gestion de données et pose la question de la monté en compétences, par la formation, de ces professionnels.

[28] IFLA. Déclaration de l'IFLA sur le libre accès à la littérature scientifique et aux documents de la recherche. [en ligne] 5 décembre 2003. Disponible sur : https://www.enssib.fr/bibliotheque-numerique/documents/1972-declaration-de-l-ifla-sur-le-libre-acces-a-la-litterature-scientifique-et-aux-documents-de-la-recherche.pdf

[29] THIAULT, Florence. Data librarian et services aux chercheurs en bibliothèque universitaire : de nouvelles médiations en émergence. [en ligne]. 7e conférence Document numérique et société. Humains et données : création, médiation, décision, narration, Nancy, septembre 2020. Disponible sur : https://hal.science/hal-02972705v1/file/Datalibrarian-THIAULT-%20DOCsocHAL.pdf

II) LES PROFESSIONNELS DE L'INFORMATION SCIENTIFIQUE ET TECHNIQUE ET LE TRAITEMENT DES DONNEES DE LA RECHERCHE : LES DONNEES ASTRONOMIQUES EN PERSPECTIVE

1) CARACTERISTIQUES DES DONNEES ASTRONOMIQUES

1. Des données massives et complexes issues des observations spatiales, des calculs et des publications scientifiques

Dans le cadre de notre exposé qui se concentre sur le rôle des professionnels de l'information scientifique et technique dans la gestion des données de la recherche et plus spécifiquement de la recherche en astronomie, nous allons à présent partir d'une description élémentaire de ce que sont les données en astronomie et de leurs principales caractéristiques

Tout d'abord, l'astronomie étant par définition la science de l'observation, le cœur de la démarche scientifique est l'observation d'objets célestes. A l'heure actuelle les points d'observations se sont multipliés avec les avancées technologiques : télescopes sol et spatiaux, missions spatiales, archives des observatoires, relevés du ciel (informations homogènes sur un grand nombre d'objets), bases de données à valeur ajoutée (CDS), données homogénéisées et extraites des publications et grands relevés, données bibliographiques issues des journaux académiques ou de bases de données comme ADS et enfin les données de modélisation.

Ainsi la combinaison de ces technologies permet d'obtenir des résultats plus fins et plus avancés et un nombre croissant des publications scientifiques autour de l'astronomie sont issus de cette multi-observation et de la modélisation.

On trouve en astronomie des données brutes et des données plus fondamentales qui ont déjà passé un premier traitement scientifique. Tout d'abord il y a différents niveaux de traitement des données brutes qui vont plus ou moins révéler leur caractère exploitable. A titre d'exemple, la National Aeronautics and Space Administration (NASA) a établi des niveaux de traitement des données brutes parmi les plus poussés allant d'un niveau 0 à un niveau 4 allant du moins exploitable au plus exploitable (0, 1A, 1B, 2, 3, 4)[30]. Le niveau 0 correspond à des « données brutes à résolution intégrale de l'instrument » (autrement dit des données nettoyées des artefacts de l'instrument), le niveau 1A correspond déjà à une donnée accompagnée de métadonnées pour lui donner une

[30] BORGMAN, Christine L. Qu'est-ce que le travail scientifique des données ? [en ligne]. Traduit de l'Anglais par Charlotte MATOUSSOWSKY. OpenEdition, 2020. Disponible sur : https://books.openedition.org/oep/14692

description élémentaire (paramètre de l'instrument, données temporelles, etc). Certains instruments traitent les données jusqu'au niveau 4 qui correspond à une donnée avec des métadonnées précises.

La donnée brute est effectivement en soi non exploitable en astronomie car produite en de grandes quantités. Véronique Stoll le formule ainsi lors d'un entretien :

« En astronomie, la donnée brute est inexploitable. Si on prend des gros instruments qui sortent énormément de données de manière quotidienne, on est obligé de réfléchir à ce qu'on va garder ou pas. Cette réflexion se fait quasiment en temps réel parfois, et ceci est asse spécifique à l'astronomie. Consciemment on sait qu'on ne peut pas tout conserver. L'astronomie annonce des problématiques qui arriveront sur les autres disciplines. »

Nous le verrons, comme l'astronomie produit des données en masse, cette discipline donne en quelque sorte un aperçu des problématiques liées aux données qui vont survenir dans d'autres disciplines. Et c'est par ailleurs ce qui est régulièrement souligné lorsqu'on échange avec les professionnels de ce domaine. L'astronomie produit des données massivement.

Des données massives

Lors d'un entretien avec une documentaliste du CDS de Strasbourg, à la question de savoir comment elle définirait les données astronomiques, celle-ci m'a répondu que ce qui l'a marquée, c'est surtout la quantité plus que la spécificité de ces données. Et qu'est-ce qu'une grande masse de données ? Nicolas Lumineaux, enseignant en informatique à l'Université Claude Bernard Lyon 1, chercheur au Laboratoire d'Informatique en image et systèmes d'information (LIRIS) et membre du groupement de recherche MaDICS BigData4Astro[31] m'indiquait en entretien qu'il n'y a pas de valeur précise à donner, mais qu'il s'agit plus de donner une définition de leur gestion. Car à chaque époque on accorde une valeur différente à ce qu'on nomme « grosse masse de donnée ». Nicolas Lumineaux aborde le sujet avec humour, me disant que : « […] ça nous ferait peut-être rire aujourd'hui, car il y a trente ans, on parlait à peine de gigas ! A chaque époque il y a ces enjeux, même si les masses de données d'avant nous font rire aujourd'hui. ». Selon lui, gérer une masse de données, quelle qu'elle soit, se définit ainsi :

«[trouver une] adéquation entre l'infrastructure dont on dispose et les données qu'on veut traiter, avec la satisfaction de leur traitement. »

Ici, plusieurs notions se croisent : la notion d'adéquation entre un outil et les objets qu'il traite. Ensuite transparaît la notion de satisfaction du traitement. Finalement, si on parvient à une gestion satisfaisante d'un volume de données, son « poids » est relatif. Or cette dernière notion recouvre en outre une axiologie plus subjective car de quel point de

[31] Action Big Data for Astronomy, atelier dans le cadre du groupe de recherche (GDR) MaDICS (Masse de données, informations et connaissances en sciences, 2020-2024. Informations disponibles à cette adresse : https://www.madics.fr/actions/bigdata4astro/

vue la gestion doit être, ou peut être satisfaisante ? Et quel niveau de satisfaction est attendu ? Comment intégrer à cela des outils performants ?

Aujourd'hui, la gestion des très grandes masses de données comme en astronomie ouvre sur des problématiques de gestion des stocks mais aussi des flux car des capteurs, satellites ou missions spatiales nous donnent des informations et des données massives, parfois même en continu. Ainsi que le présente Nicolas Lumineaux :

« Il y a 20 ans la problématique était de pouvoir tout traiter. Là c'est plutôt : comment traiter efficacement, en prenant en compte l'enjeu énergétique et en faisant les bons choix technologiques ».

Ainsi on qualifie les données astronomiques de données massives. Les nouvelles générations de télescopes au sol dont sont issues un nombre important de données sont en mesure de produire des téraoctets de données. Pour donner quelques exemples, le programme de cartographie du ciel Sloan Digital Sky Survey (SDSS)[32] produit à lui seul 15 téraoctets de données en une nuit. Le programme Visible and Infrared Survey Telescope for Astronomy (VISTA)[33] quant à lui produit 1,3 téraoctets de données chaque nuit. Il en va de même pour le Large Synoptic Survey Telescope (LSST)[34] qui produit des dizaines de téraoctets chaque nuit. Ces très grosses masses de données ne peuvent pas être transférées ou stockées telles quelles peu de temps après leur production car nous n'avons pas les moyens techniques de transférer vers des centres de traitement de telles masses de données. Et est-ce seulement pertinent, pour reprendre le constat de Véronique Stoll qui dit que consciemment on sait qu'on ne peut pas tout garder ? L'objectif est bien de faire de la science, d'apporter une réelle expertise avant même de traiter dans leur entièreté ces données. Pour opérer cette sélection de données il faut une expertise scientifique à la racine de leur production. Il faut sélectionner et produire des outils capables de faire des traitements intelligents.

Des données variées

Parfois, une opération spéciale va consister à prendre directement des données issues de certains relevés comme avec les données issues des relevés de la mission GAIA[35] qui ont été versées dans la base du CDS de Strasbourg. Ces données sont traitées et publiées par les services du CDS avant qu'une publication dans un journal à propos de ces données ne soit publié. Les CDS permettent d'ajouter de la valeur aux données brutes produites initialement comme pour les versements suite à des missions spatiales.

De plus, la complexité des données en astronomie est liée à la nature, l'obtention et la collecte de la donnée. Nous pouvons donner ici l'exemple des données en planétologie ou en physique solaire qui sont complexes à obtenir de par la rareté des missions capables de se rendre au plus proche de l'objet observé ou de l'outil au sol utilisé pour observer.

[32] Programme de relevé des objets célestes via un télescope au sol de 2,5 mètres de diamètres situé à l'Observatoire d'Apache Point (Nouveau-Mexique, Etats-Unis).
[33] Télescope de l'Observatoire européen austral (ESO) situé à l'Observatoire du Cerro Paranal dans le désert d'Atacama (Chili). Il est doté d'un miroir de 4,1 mètres et il est entièrement consacré au recensement des objets célestes.
[34] ou Observatoire Vera-C.-Rubin. Télescope optique américain situé au sommet du Cerro Pachon (Chili).
[35] GAIA : mission lancée par l'Agence spatiale européenne (ESA) en 2013 et qui s'est achevée en 2020. La mission Gaia de l'ESA a permis de dresser le catalogue d'étoiles le plus étendu à ce jour, y compris des mesures de haute précision de près de 1,7 milliard d'étoiles ainsi que des détails inédits de notre propre galaxie.

Véronique Stoll m'indiquait que les données les plus structurées que nous obtenons sont les données issues de l'espace lointain. Pour quelle raison ?

Les données de l'espace lointain sont celles qui sont utilisées et structurées rapidement car elles sont massives et que leur gestion fait partie des plans de financement des missions, le plan de gestion de données étant structurant en amont du projet. Sur les données issues de l'observation spatiale de l'espace lointain, un embargo[36] de 6 à 12 mois est régulièrement préconisé dans les programmes de financement de la NASA. En planétologie les données sont plus chères et plus rares à obtenir et les organismes de production de ces données souhaitent les conserver plus longtemps afin de les traiter. Pour citer Véronique Stoll,

« les grandes missions spatiales sont les arbres qui cachent la forêt »

car on peut plus facilement récupérer des jeux de données issus des grosses missions telles que le James Webb Space Telescope (JWST)[37] ou Cassini-Huygens[38]. Les grandes missions spatiales de ce type ont positivement participé à la visibilité des données, y compris auprès du grand public. Or dans la réalité, selon Véronique Stoll, un certain nombre de données en astronomie ne sont pas faciles à obtenir, et ne sont pas à l'image de ces grandes missions.

Des données issues des publications scientifiques

Outre les données issues des observations, une autre part des données astronomiques sont issues des publications ou des archives scientifiques. Dans les CDS et notamment si l'on prend l'exemple du CDS de Strasbourg, la base de données est dite à valeur ajoutée car les données qu'elle contient ont déjà passé un premier traitement scientifique. En effet les équipes sélectionnent spécifiquement les données qui proviennent des publications *peer-reviewed*[39] par l'intermédiaire d'un accord passé avec une vingtaine des plus importantes revues scientifiques astronomiques. Le CDS a des accords avec ces grands journaux pour utiliser les articles et extraire les données de ces publications scientifiques. Ce sont donc les documentalistes eux-mêmes qui alimentent les bases de données. Lors d'un entretien que j'ai eu avec Esther Collas, qui occupe un poste de chargée du traitement des données scientifiques au CDS de Strasbourg pour la base de données SIMBAD, elle m'indique que les documentalistes traitent en théorie les 20 plus grands journaux d'astronomie pour en extraire les données (en pratique, les plus grands éditeurs en termes de volume de données à extraire vont être EDP Sciences Astronomy & Astrophysics[2], MNRAS et les journaux de l'American Astronomical Society).

[36] Un embargo désigne une période durant laquelle l'accès à une publication, une source, une donnée ou autre est restreint voire interdit

[37] JWST : télescope spatial lancé en 2020 servant d'observatoire fonctionnant principalement dans l'infrarouge, développé par la NASA avec la participation de l'ESA et de l'Agence spatiale canadienne.

[38] Cassini-Huygens : mission lancée en 1997 et qui pris fin en 2007, elle résulte d'un partenariat entre la NASA et l'ESA et l'Agence spatiale italienne. Il s'agit de la plus grosse sonde envoyée dans le système solaire pour récolter des données sur le système saturnien.

[39] « Examiné par les pairs »

Ces bases à valeur ajoutée ont une importance capitale dans la recherche. Il y a par exemple eu plus d'articles scientifiques basés sur les données d'archives du Hubble Space Telescope (HST) que sur les observations originales.[40]

Cette dernière information nous laisse entrevoir l'importance que revêtent les outils pour gérer ces données. Tout ce qui va servir à recueillir, traiter, stocker et diffuser les données astronomiques fait la discipline elle-même car les données brutes sont peu utilisées. Ajouter une valeur aux données par l'expertise scientifique et technique est essentiel au fonctionnement de cette discipline, et ceci ne peut se faire sans une expertise scientifique comme documentaire.

2) ORGANISER LES DONNEES DE LA RECHERCHE PAR DES OUTILS PERFORMANTS : UNE PORTE D'ENTREE DOCUMENTAIRE SUR L'UNIVERS

1. Particularités des identifiants et standards d'échange des données astronomiques

Les caractéristiques des données astronomiques ayant été détaillées, on peut alors se demander alors quels outils ont été mis en place pour gérer ces données. Pour gérer ces données, il existe des bases de données, des formats ou des normes qui permettent de transférer ces données, de les (ré)utiliser, en somme des outils qui permettent à ces données d'être exploitables. Les bases de données mettent en forme et rendre à minima intelligibles ces données. En France – et à visée internationale – des outils très performants sont liés au CDS de Strasbourg, institution qui, nous le verrons, est singulière dans le paysage des unités intégrées aux principaux Observatoires des Sciences de l'Univers (OSU)[41]. A l'Observatoire de Paris, outre les bases de données reliées à celui-ci, des services aux chercheurs ont été mis en place pour les accompagner sur les PGD. En parallèle de la création des outils de gestion des données, de normes et des standards d'échange sont venus irriguer la discipline pour que le partage des données soit plus rapide et plus simple.

Au sein de la communauté astronomique, et ce dès les années 1970 s'est posée la question de réfléchir à un format qui puisse contribuer à la réutilisation des images conservées sous format électronique. Ainsi que l'énonce Françoise Genova, directrice de recherche émérite au CDS de Strasbourg :

[40] MIKULSKI ARCHIVE FOR SPACE TELESCOPES. HST Reports. [en ligne] Disponible sur : https://archive.stsci.edu/hst/bibliography/pubstat.html

[41] Les OSU ont pour mission principale dé décliner le projet stratégique de l'Institut en « organisant les moyens nécessaires à l'acquisition d'observations des systèmes astronomiques ou des composantes du système Terre ». Voir : https://www.insu.cnrs.fr/fr/les-observatoires-des-sciences-de-lunivers

« Certaines disciplines n'ont pas attendu que le sujet soit à la mode pour mettre en place des politiques de partage de leurs données et se mettre en ordre de marche pour prendre en charge et distribuer celles-ci. »[42]

FITS et Bibcode

Afin d'illustrer le propos de Françoise Genova, nous pouvons donner deux exemples de formats standards d'échange en astronomie : le format Flexible Image Transport System (FITS) pour les échanges de données et le format Bibcode pour les bases de données bibliographiques. Car il se trouve que le domaine de l'astronomie a été précurseur sur la gestion et le partage des données et des références scientifiques.

On qualifie aujourd'hui de FAIR les données scientifiques fiables et de qualité. En astronomie, le format FITS un standard d'échange, a commencé à voir le jour en 1977 autour d'un atelier de travail[43]. La date de 1979 est retenue par la NASA pour son élaboration[44], et c'est en 1981 qu'est publié un premier papier qui le définit. C'est l'un des premiers formats ouverts d'échange de données à se qualifier tel quel, en cherchant spécifiquement l'interopérabilité entre systèmes et institutions :

« This interchange of data has traditionally been hampered by the fact that each installation has generated its own software system for image processing tailored to its own computer facilities, which differ enormously. Almost every installation has developped at least one unique data format and produced a large quantity of software based on the use of that internal format. Given this situation, the adoption of a single format for use in all installations would be prohibitively expensive and would lead, in general, to less efficient computing within the individual systems. However, a feasible course of action is the adoption of a unique interchange tape format to be used for transferring digital imagery between cooperating institutions. »[45]

Tout d'abord, l'une des idées développées ici est que produire un seul format dans lequel on enregistrerait par exemple les données d'observations en astronomie était très compliqué. Chacun avait sa façon d'enregistrer ses données avec ses outils, ses logiciels. Adopter un format unique pour toutes les installations aurait été beaucoup trop coûteux en termes humains et en termes financier. La solution trouvée pour essayer de s'aligner entre institutions et entre groupe de chercheurs a été de trouver un format d'échange de données qui pourrait faciliter les transferts entre les outils, à défaut de trouver seul format d'enregistrement de données. C'est donc par là que l'adoption d'un outil

---

[42] GENOVA, Françoise. Du nécessaire partage des données scientifiques. : l'exemple de l'astronomie. *Ar(abes)que* [en ligne]. 2014, (73), 12–13. Disponible sur : https://publications-prairial.fr/arabesques/index.php?id=999

[43] EVANS, J. , KIRSCH, R. and NAGEL, R. Workshop on Standards for Image Pattern Recognition. [en ligne]. National Institute of Standards and Technology, Gaithersburg, MD, 1977. Disponible sur : https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nbsspecialpublication500-8.pdf

[44] FITS Working Group. Definition of the Flexible Transport System (FITS). [en ligne]. Version 4.0 mise à jour le 22 juillet 2016. Disponible sur : https://fits.gsfc.nasa.gov/standard40/fits_standard40draft1.pdf

[45] « Cet échange de données a traditionnellement été entravé par le fait que chaque installation a généré son propre logiciel de traitement d'image adapté à ses propres installations informatiques, qui diffèrent énormément. Presque chaque installation a développé au moins un format de données unique et produit une grande quantité de logiciels basés sur l'utilisation de ce format interne. Compte tenu de cette situation, l'adoption d'un format unique à utiliser dans toutes les installations serait d'un coût prohibitif et conduirait, en général, à un calcul moins efficace au sein des systèmes individuels. Cependant, une ligne de conduite réalisable est l'adoption d'un format de bande d'échange unique à utiliser pour le transfert d'images numériques entre institutions. » WELLS, D., E. GREISEN et R. HARTEN. FITS : A Flexible Image Transport System. Astronomy & Astrophysics [en ligne]. 1981, (Supplément Series 44), 363–370. Disponible sur : https://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?bibcode=1981A&AS...44..363W&db_key=AST&page_ind=7&plate_select=NO&data_type=GIF&type=SCREEN_GIF&classic=YES

d'interopérabilité a été une question posée très tôt en astronomie. Sans un format d'échange de données stable, la discipline n'avance pas. Ce format répond à un besoin de partage de la communauté scientifique.

Dans ce paragraphe, l'utilité et les enjeux d'un format commun sont exposés. Ainsi, élaborer un standard d'échange plutôt qu'un standard de production des données a été retenu par la communauté astronomique. Le format FITS intègre les données et les métadonnées afin qu'elles soient réutilisables par les systèmes informatiques des institutions de recherche. Il permet de partager les observations des télescopes et contribue à un travail scientifique à valeur ajoutée. Sans que ce format concerne les données en elles-mêmes, il demande à celles et ceux qui les produisent qu'elles soient qualitatives afin de pouvoir s'intégrer à ce protocole.

A travers cet exemple de l'élaboration du format FITS on entrevoit l'importance de la collaboration entre les institutions scientifiques. Les valeurs associées aux données (interopérabilité, accessibilité, etc) s'articulent avec des pratiques et des situations précises (ici, le besoin de rendre les outils plus performants). De ce fait, les chercheurs qui faisaient face à un problème d'ordre technique ont « matérialisé » le système de valeurs associé aux données en le concrétisant dans un format. Cette collaboration, qui aboutit sur des outils concrets est un enjeu de taille pour faire avancer la recherche. Cependant elle n'est pas sans poser des difficultés et des coûts humains comme financier pour les institutions. Outre le format FITS nous pouvons donner un autre exemple d'outil cette fois-ci bibliographique, le format Bibcode, devenu un standard d'échange en astronomie.

Les bases de données bibliographiques en astronomie utilisent le format Bibcode[46], initialement utilisé par la base de données SIMBAD du Centre de données astronomiques (CDS) de Strasbourg et par la base de données Astrophysics Data System (ADS) de la NASA. Son but est d'identifier les références scientifiques par un code identique de 19 caractères pour chaque référence[47].

Il est également utilisé par les éditeurs de journaux en astronomie. Ce standard en la matière est original et en fait l'un des premiers standards spécifiques à un domaine de recherche, interprétable facilement et pouvant être créé par des personnes physiques. Ceci facilite donc le travail des professionnels de l'IST pour la gestion des données, car il est facile à comprendre et à utiliser. Par conséquent, le Bibcode, qui au préalable a été conçu pour une base de données spécifique (SIMBAD puis ADS) est devenu un standard de gestion et de partage des ressources en astronomie. C'est un exemple qui relate la standardisation d'un format favorisant l'interprétation et la recherche des ressources scientifiques dans un domaine donné.

[46] ARCHIMBAUD, Jean-Luc. Identifiants des documents numériques : ISBN, ISSN, URL, DOI, OpenURL... [en ligne]. 26 janvier 2015. Disponible sur https://archivesic.ccsd.cnrs.fr/sic_01068135v2/document

[47] Guide de lecture du format Bibcode : CDS, Strasbourg. NED and SIMBAD Conventions for Bibliographic Reference Coding. [en ligne]. Disponible sur : https://simbad.cds.unistra.fr/guide/refcode/refcode-paper.html

2. Les bases de données scientifiques à portée de tous

Les bases de données sont des outils qui permettent d'ordonner et rendre plus facilement réutilisables les données. Ces outils numériques regroupent des données organisées, issues des observations sol ou spatiales, ou issues des résultats de calculs scientifiques. Ces bases ont un but commun : permettre une large réutilisation des données pour la communauté scientifique visée. Elles sont en grande majorité gratuites et accessibles librement depuis le Web. Ainsi les scientifiques du monde entier peuvent réutiliser ces données. Comme l'indique le site de l'Observatoire de Paris dans sa définition :

« Cette mise à disposition de données, effective dans le monde entier et dans tous les domaines de l'astronomie et de l'astrophysique, enrichit les échanges de la communauté scientifique et permet aux scientifiques du monde entier d'en tirer bénéfice ».[48]

On trouve différentes bases de données astronomiques issues de collaborations entre l'Observatoire de Paris et d'autres grandes institutions en France comme à l'étranger telles l'Institut Max Planck, l'IVOA, l'IRAP, diverses infrastructures de recherche et réseaux de chercheurs en astrophysique dans le monde. Lors de mes entretiens, j'ai eu l'occasion de rencontrer plusieurs personnes qui travaillent pour le CDS de Strasbourg. Si cette institution revient souvent, c'est qu'elle occupe une place particulière dans la communauté astronomique française et internationale. En effet, le CDS de Strasbourg propose trois principaux services : ALADIN (atlas interactif du ciel), VizieR (catalogues astronomiques) et SIMBAD (base de données).

Focus sur les outils proposés par le Centre de données astronomiques de Strasbourg

Le CDS de Strasbourg est un centre de données qui a une renommée internationale, c'est l'un des dix gros centres de données astronomiques dans le monde. Les équipes qui le gèrent sont composés d'environ un tiers de documentalistes, un tiers d'informaticiens et un tiers d'astronomes.

La base de données SIMBAD fournit des données de bases, une bibliographie et des mesures pour les objets astronomiques en dehors du système solaire. L'outil réunit en une page web un set de données fondamentales pour un objet astronomique. Grâce à cet outil les utilisateurs ont accès à la liste de toutes les publications associées à un objet recherché. Enfin une publication peut être rentrée dans la base, des listes d'objets et des scripts peuvent être soumis.

[48] Liste des bases de données inclues dans le Paris Astronomical Data Center : https://www.observatoiredeparis.psl.eu/-bases-de-donnees-scientifiques-.html

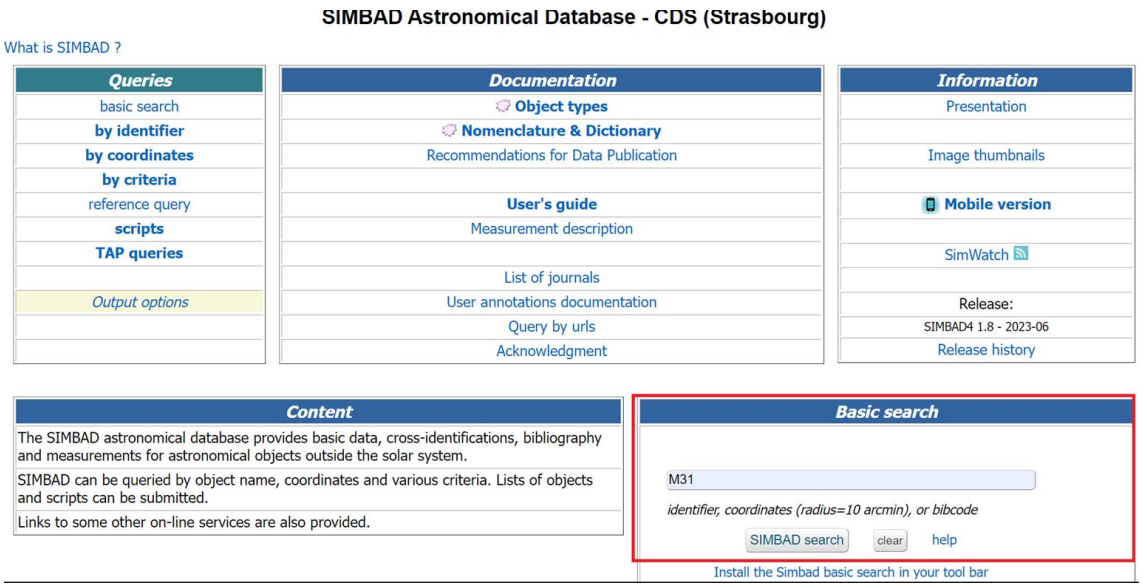

Figure 3 Page d'accueil de SIMBAD

Une entrée accessible directement depuis la page d'accueil de SIMBAD permet de rechercher un objet par son nom. Par exemple ici, nous recherchons des informations relatives à l'objet « M31 », connu également sous le nom de Galaxie d'Andromède, qui est la galaxie la plus proche de la nôtre.

Une page web s'ouvre correspondant à cet objet avec un jeu de données associées :

Figure 4 Set de données associées à M31 ou Galaxie d'Andromède dans la base SIMBAD

- Son nom
- Son type (ici c'est une Galaxie)

- D'autres types d'objets spécifiés dans la littérature : QSO (=quasar[49]) Rad (= source radio) X (=source X), GiC (=cluster, cette galaxie est dans un cluster[50]), GiG (=groupe, cette galaxie est dans un groupe[51])
- Les données fondamentales qui lui sont associées (nous les spécifierons juste après)
- Une image qui provient d'ALADIN
- Enfin toute une bibliographie est associée à cette page :

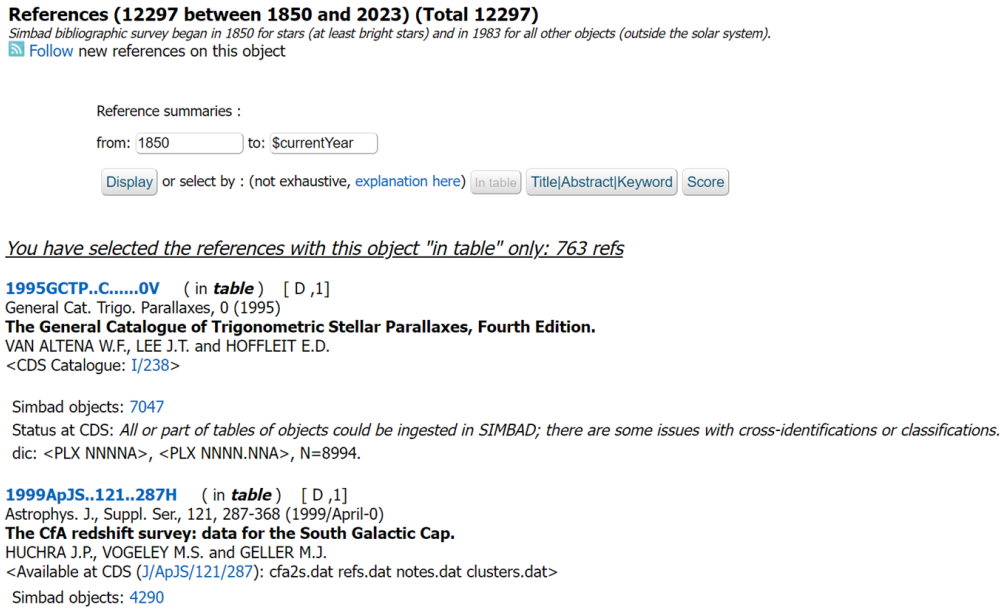

Figure 5 Références bibliographiques associées à la recherche "M31" dans la base de données SIMBAD

Parmi les données déjà traitées et intégrées dans des bases telles que SIMBAD les données vont servir le plus souvent à l'identification des objets. Ces données sont communes au plus grand nombre de type d'objets différents, c'est-à-dire que nous allons retrouver des données ou informations communes pour une galaxie ou une étoile qui sont pourtant des objets très différents et qui servent des sous disciplines de l'astronomie différentes.

On trouve dans ces données fondamentales :

- Les identificateurs[52]
- Les types d'objets (galaxie, étoile, quasar, planète, etc)
- Les coordonnées[53]

[49] Ou « source de rayonnement quasi stellaire » est une source d'énergie électromagnétique incluant la lumière.
[50] Amas de galaxies.
[51] Amas de plus d'une centaine de galaxie parvenu à un équilibre dynamique.
[52] Les identificateurs sont tous les « noms » sous lesquels les objets sont identifiés. A un seul objet peut correspondre un grand nombre d'identificateurs.
[53] Les coordonnées permettent de trouver l'objet dans la carte du ciel et de pointer un télescope dessus.

- La vitesse[54]
- Les flux et les magnitudes[55]
- Les parallaxes[56]

Les données fondamentales décrites ci-dessus ont été établies par le CDS de Strasbourg et sont basées sur le fait que ces informations sont les plus souvent demandées et les plus utilisées par les chercheurs comme par les astronomes amateurs pour reconnaître et identifier les objets. Elles sont devenues un standard en astronomie. On voit par-là que les données en astronomie se constituent comme une sédimentation des habitudes des chercheurs. Si telle donnée est très utilisée dans telle recherche, alors les documentalistes ont estimé qu'elle était importante et peut constituer une norme ou un standard. L'expertise des documentalistes dans la gestion des données est également sollicitée en ce qu'ils sont capables de reconnaître les besoins des chercheurs.

Enfin, ALADIN est un atlas du ciel interactif qui permet de visualiser des images astronomiques numérisées ou des relevés complets. Cet outil permet de superposer des entrées de catalogues ou d'autres bases de données. Il permet également d'accéder de manière interactive aux données des bases du CDS de Strasbourg. Il y a toute une technologie pour pouvoir superposer ces images, naviguer dedans. Les documentalistes l'utilisent quotidiennement car cette carte interagit avec tous les objets rentrés dans la base de données SIMBAD. Lorsqu'on zoom dans la carte du ciel des petits points verts apparaissent. Ces points représentent les objets connus dans la base de données. C'est utile lorsqu'un documentaliste analyse un article pour en extraire des données : cela permet de vérifier par exemple de quelle étoile il s'agit.

Figure 6 M31 dans l'atlas du ciel ALADIN

[54] Les trois types de données fondamentales de vitesse sont les données de calculs de la vitesse radiale d'un objet (vitesse d'un objet mesuré à partir du point d'observation ou depuis l'objet vers le point), le *redshift* (ou « décalage vers le rouge », observation dans une longueur d'ondes données des objets lointains de l'univers) et le cz (calcul de la vitesse de la fuite d'un objet lointain).
[55] Les flux et les magnitudes représentent le taux de brillance d'un objet dans une longueur d'ondes donnée.
[56] Les parallaxes permettent d'identifier à quel point un objet est proche de nous et comment il se déplace par rapport à nous.

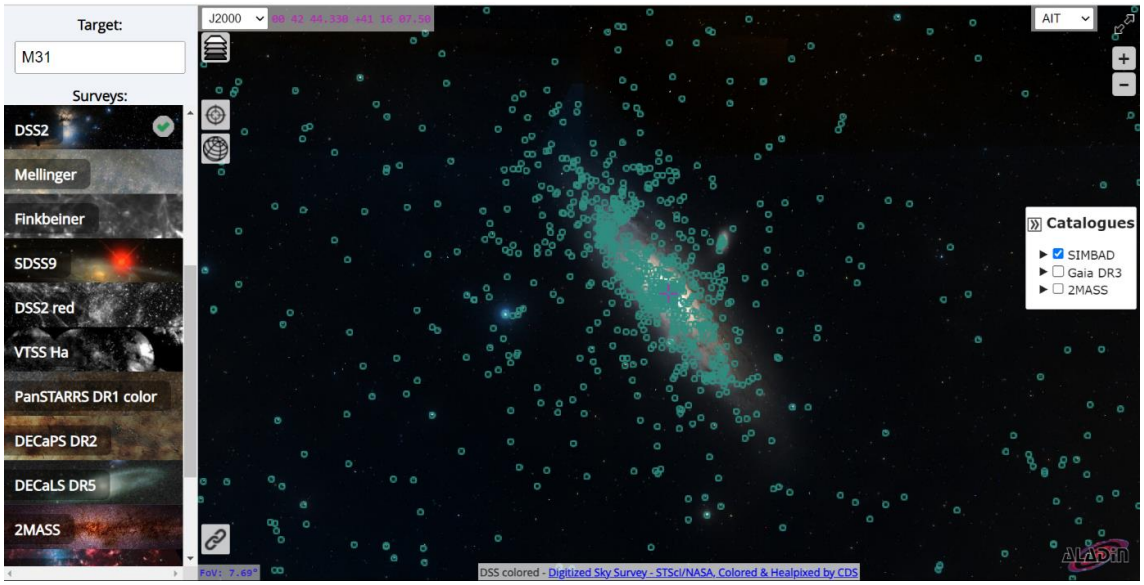

Figure 7 Un clic sur "SIMBAD" permet d'afficher les ressources associées aux objets dans la carte du ciel. Plus on zoom et plus les références apparaissent sous forme de points.

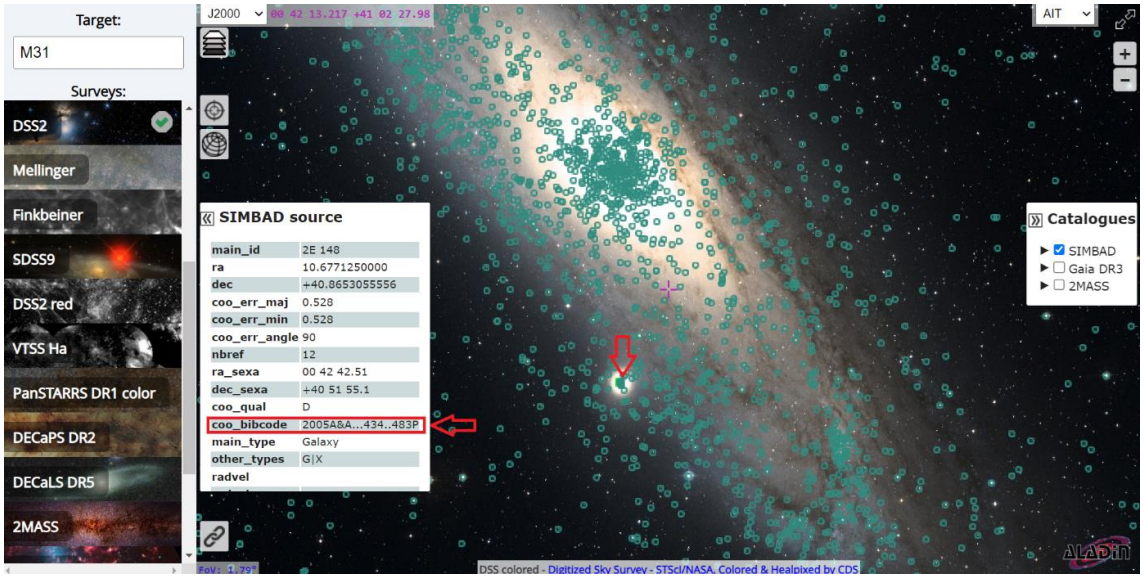

Figure 8 Un clic sur un point permet d'afficher la référence. Un certain nombre d'informations apparaissent dont la référence bibliographique sous format Bibcode. L'utilisateur peut récupérer cette référence et effectuer une recherche dans les catalogues.

Enfin, VizieR est un service de catalogue qui expose et présente les tables de données telles qu'elles proviennent des publications mais formatées et standardisées. Chaque colonne a été décrite avec les standards de l'International Virtual Observatory

Alliance (IVOA)[57]. Cet outil permet une interaction dynamique avec les données, on peut les trier, les croiser, les réutiliser pour plusieurs publications différentes, et des liens sont faits avec d'autres catalogues. L'équipe de VizieR est chargée de mettre en forme les données sous formes brutes avec les standards internationaux pour les colonnes.

3. Observatoires Virtuels et défis sémantiques

Différentes bases de données dans diverses disciplines peuvent converger dans un outil qu'on nomme Observatoire virtuel (OV). Les OV permettent de créer un espace d'interopérabilité entre différentes bases de données, ceci pour former une archive numérique de données organisées. Les utilisateurs peuvent alors réutiliser des jeux de données issus de diverses bases de données dans des domaines différents (planétologie, physique solaire, codes, physique atomique ou moléculaire, etc).

Pour récapituler, les CDS proposent un panel d'outils et de bases de données spécifiquement liées à l'astronomie, et les OV sont des outils qui regroupent des bases différentes et qui peuvent être interrogées simultanément. Les OV concernent des données d'observations de diverses disciplines (OV en sciences de la Terre par exemple, en géophysique, etc). Les CDS quant à eux fournissent un support aux données d'observation et permettent l'interopérabilité de ces mêmes données dans un espace, l'OV. L'OV ne rassemble pas toutes les données au même endroit mais en une requête on peut interroger tous les *data centers*[58]. L'OV sert de passerelle technique, il définit des normes entre le fournisseur de données et l'utilisateur. Les fournisseurs partagent des données et des services pour éviter les flux de données. Les utilisateurs vont quant à eux chercher et puiser dans tous les réservoirs. Ces interactions entre bases s'appuient sur le protocole Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Or l'OV n'est pas sans poser des difficultés de gestion de des défis sémantiques. En effet une gestion efficiente des données passe, nous l'avons plus haut, par une standardisation du vocabulaire utilisé et des métadonnées qui accompagnent les jeux de données. Par exemple un groupe de recherche produit ce type de jeu de données : ils ont des objets observés, ils décrivent leur position dans le ciel, le nom de l'instrument qui leur a servi à détecter l'objet, ainsi que la date d'observation. Comment entrer ces données dans l'OV ? Les documentalistes ont besoin d'un vocabulaire standard, d'un format fixe à utiliser. Ainsi que l'énonce Véronique Stoll :

« Toutes les communautés en astrophysique ne rentrent pas encore dans l'OV car toutes n'ont pas fait ce travail de description de l'ensemble du vocabulaire. Il y a des démarches, et c'est assez lent. Il faut donc d'abord être d'accord sur ce qu'on utilise. »

Ce travail de description ne peut être fourni par la communauté de l'IST et des documentalistes. Toutes et tous n'ont pas l'expertise scientifique pour dire que telle information est nécessaire à l'utilisation d'un jeu de données en astrophysique. Par conséquent, une réflexion au niveau disciplinaire sera à fournir par la communauté

[57] Organisation scientifique internationale créée en 2002 qui a pour but de coordonner et permettre un accès mondial aux données recueillies par les observatoires astronomiques.
[58] « Centre de données »

astronomique et ce ne peut être les bibliothèques ou les centres de données qui s'en occupent. Par exemple, les observations des satellites ou d'autres missions spatiales extraient des données très différentes (spectroscopie, positionnement, vitesses, etc) et il faut faire un choix : par exemple avec tel spectre combiné à tel autre on obtient une donnée importante. Donc quelles données on choisit ? C'est à la communauté des chercheurs d'énoncer ses besoins.

3) DE L'IMPORTANCE DE STANDARDISER ET NORMALISER LES DONNEES DE LA RECHERCHE

1. Une science ouverte en avance émanant des pratiques des chercheurs

Ce tour d'horizon des outils du CDS de Strasbourg permet d'illustrer la diversité de ce que peuvent utiliser les chercheurs comme les professionnels de l'information dans l'utilisation ou la gestion des données. Chaque corps de métier s'en sert et les alimente de concert, et ce depuis leur création. Selon un membre de l'IVOA, ingénieur de recherche à l'Observatoire de Strasbourg et président de la chaire sur la conservation des données au sein de l'IVOA avec qui j'ai pu m'entretenir, les bases de données astronomiques sont des outils qui se sont mis en place presque naturellement quand les masses de données à traiter sont devenus considérables :

« Avant, quand l'ESA envoyait un satellite, les données recueillies étaient accessibles pendant un certain temps et ensuite, elles étaient archivées et/ou transférées sur des disques durs, etc. Très peu de données étaient en ligne. Le CDS de Strasbourg est un des tous premiers à avoir mis ses données en ligne. Il y a 10 ans, les données n'étaient pas accessibles via le web. Le besoin de rendre accessible un maximum de choses en ligne était important. »

En effet, le besoin de la communauté de mettre à disposition ses donnée sur le web est venu également d'un changement de pratique des chercheurs. L'astronomie a eu besoin d'aller vers de l'astronomie multi-bandes : c'est-à-dire que les techniques d'observations ont évolué et les chercheurs voulaient comparer les données entre elles. L'astronomie multi-bandes est un dispositif photographique qui consiste à enregistrer simultanément des images dans différentes bandes spectrales (comparer des données de l'X avec de l'infrarouge ou de l'optique par exemple). Les instruments à bord des satellites sont devenus alors plus pointus et plus gourmands en production de données. Lors de notre entretien, ce membre de l'IVOA parle d'une « avalanche » de données : il fallait alors mettre un maximum de choses en ligne pour pouvoir comparer les données entre elles, autant pour les traiter que pour les utiliser.

Ainsi la gestion de ces volumes de données et le besoin de les mettre en ligne pour que la communauté y ait accès était un défi technologique émanant directement des nouvelles pratiques des chercheurs. Ceci est une particularité saisissante en astronomie, ce qui en fait un domaine dont l'exemple est « extrême » et permet de spéculer sur les défis à venir. Il est de commune mesure dans cette discipline que les outils informatiques actuels ne répondent pas à la demande lors de la création d'un projet, déjà car la construction des projets se fait sur le long terme. En d'autres termes, on sait d'avance qu'un très gros

télescope va produire des téraoctets de données avant sa construction, et les outils pour gérer ces données ne sont pas spécifiquement mis en place. Par exemple, le Large Synoptic Survey Telecsope (LSST) au Chili va bientôt démarrer la plus grosse caméra CCD[59] du monde. Ce projet a commencé il y à 15 ans alors qu'il n'y avait rien pour gérer ses données. Il va s'agir alors de tout mettre en place pendant 15 ou 20 ans pour que les outils soient opérationnels :

« On est toujours en avance sur ce qui est possible techniquement de faire » confie ce membre de l'IVOA.

En ce sens, les formats FITS et Bibcode, ainsi que l'établissement des outils comme ceux du CDS de Strasbourg sont des exemples d'une façon de faire de la science ouverte sans en avoir encore le concept. Ces deux formats montrent que l'astronomie est une communauté spécifique avec une politique des données particulière, où la gestion de celles-ci découlent en grande partie des pratiques et besoins des chercheurs. C'est un exemple de science ouverte émergente, initiale. C'est-à-dire que les formats utilisés ont émergé depuis la base d'une communauté, par les pratiques, et non pas par un standard donné en amont par des institutions. La dynamique est différente d'avec les principes FAIR qui émanent d'une association de professionnels de l'information.

Pour autant, cela ne veut pas dire que les professionnels de l'IST n'ont pas leur place dans ces réflexions. Dans le cas de l'astronomie les institutions ont par la suite collaboré pour aider à la mise en pratique de ces formats. C'est une collaboration entre différents acteurs. On peut parler d'une ouverture des données de la recherche à l'échelle d'une discipline internationale qui impulse un projet à large échelle, à l'image du mouvement de la science ouverte actuellement à l'œuvre. Le modèle de l'astronomie, qui favorise le partage entre les scientifiques et les institutions, est un des modèles précurseurs et qui montre, pourrait-on dire, l'exemple à ce qui deviendra la science ouverte. Nous ne pouvons pas nous prononcer sur le fait que sans ces modèles d'interopérabilité, des découvertes majeures n'auraient pas eu lieu. En revanche on peut affirmer que cela a permis un accroissement de la visibilité du travail des chercheurs liés à un institut particulier, et favorise aussi le travail des documentalistes. Le partage et l'accessibilité favorisant la recherche, par là ils favorisent également la découverte scientifique.

Ces exemples illustrent la complexité des données astronomiques et l'importance d'une bonne gestion de celles-ci pour pouvoir correctement les structurer, puis les réutiliser. Véronique Stoll, dans ses missions de directrice de la bibliothèque de l'Observatoire de Paris a une vision nette de l'importance d'une ouverture des données, ainsi que de ses écueils. Car si les chercheurs peuvent identifier leurs besoins, on peut se demander dans quelle mesure la formulation de ces besoins doit ou peut être cadrée. Dans ce cadre, la bibliothèque de l'Observatoire est un lieu propice à l'accompagnement de la recherche, tant dans l'expression des attentes des institutions vis-à-vis des chercheurs que dans la compréhension de leur travail. Au sein des bibliothèques, les missions dédiées à la science ouverte vont appuyer ces demandes.

[59] Capteur photographique basé sur un dispositif à transfert de charges

2. Bibliothèques des Observatoires : entre conservation et valorisation, le cas de la base de données solaires BASS2000

Au sein de cet écosystème varié d'outils, les bibliothèques rattachées aux observatoires tiennent une place particulière en ce qu'elles concentrent des compétences et des services mis à disposition des chercheurs. Ce rôle de médiation qu'ont ces institutions est à mettre en parallèle avec un rôle de participation au travail de recherche en astronomie et de valorisation de celui-ci. Nous pouvons souligner le rôle de conservation qui leur est accordé, notamment pour la bibliothèque de l'Observatoire de Paris qui conserve en son sein des documents anciens qui ont été numérisés pour faciliter leur utilisation. Un exemple particulier est celui de la base de données solaires BASS2000 qui est née d'une collaboration entre sites d'observation (les observatoires de Bruxelles et de Coimbra) et la bibliothèque de l'Observatoire de Paris qui conservait des cartes synoptiques de l'activité solaire[60].

Ces documents sont rares car peu d'appareils sont en mesure d'observer le soleil dans son intégralité et ce quotidiennement depuis des décennies. La bibliothèque conserve un fonds de ces cartes depuis le début du 20e siècle et elles ont été intégrées au projet BASS2000 pour intégrer une base de données accessible en ligne. D'autres données insérées dans le projet (notamment celles issues des observations ultraviolettes) proviennent d'articles en libre accès. Le responsable scientifique du projet BASS2000 m'indique lors d'un entretien que ce type de projet de migration de fonds vers une base de données requiert des compétences diverses qui dépassent l'expertise d'un scientifique. Selon lui que je questionne sur le rôle des documentalistes dans ce projet, ils auraient besoin de l'expertise des documentalistes pour savoir comment utiliser les données publiées dans des atlas numérisés de façon privée et/ou publiées par des instituts qui n'existent plus, et qu'ils auraient besoin d'intégrer à la base :

« C'est typiquement dans ces cas de figure que des documentalistes sauraient mieux que nous [chercheurs] où chercher et quoi faire avec ces données. »

En effet selon lui, et cette position rejoint notre point de vue sur le recueil des besoins des chercheurs par les professionnels de l'IST :

« Dans les bases de données scientifiques, il y a seulement des scientifiques du domaine qui sont capables de dire ce dont on a besoin. Or, il y aussi tout un aspect que les scientifiques ne sont pas capables de faire. Nous sommes beaucoup moins compétents dans les data management plan (DMP) par exemple, ou dans l'écriture de cahiers pour la certification CoreTrustSeal[61], ou encore la gestion des codes utilisés. »

Il m'indique également que sa mission de responsable scientifique pour la base de données occupe 20% de son poste. Ainsi un apport technique d'un documentaliste et d'un

[60] Une carte synoptique de l'activité solaire étale les 28 jours d'observations du Soleil sur une carte rectangulaire (28 jours correspondants à une rotation du Soleil). Elle donne pour chaque rotation du Soleil une vision globale de son activité. La bibliothèque de Paris conserve les cartes synoptiques de 1919 à 2002. Le Soleil ayant un cycle d'activité de 11 ans et un second de 80 ans, cela permet donc une étude sur le long terme du comportement du Soleil. Ce sont les seuls à avoir des données sur un si long terme, ce qui en fait des données rares.

[61] Le CoreTrustSeal (ou CTS) est une organisation internationale, communautaire, non gouvernementale et à but non lucratif qui promeut des infrastructures de données durables et fiables. Il offre à tout référentiel de données intéressé une certification basée sur les exigences des référentiels de données dignes de confiance.

informaticien serait bienvenu pour compléter les compétences techniques qui peuvent être limitées comme indiqué précédemment.

Nous voyons avec l'exemple du projet BASS2000 une continuité entre la recherche en physique solaire et l'apport de la bibliothèque. En effet dans ce projet, le rôle de conservation de la bibliothèque s'allie avec son rôle de valorisation. La bibliothèque conserve les cartes synoptiques dans des versions imprimées puis propose une numérisation de celles-ci. La bibliothèque ne gère pas directement la base de données créée, cependant elle assume une partie de la valorisation de ces données. Ceci est un exemple parmi les outils cités plus haut de la porosité entre différents milieux et du besoin exprimé par la recherche quant à une expertise externe pour l'aider à gérer ses données. On voit par-là que la création des outils techniques spécifiques à la recherche astronomique demande une expertise scientifique.

Or ceci amène à repenser les relations entre les professionnels de la recherche et les professionnels de l'information et de la documentation car dans la gestion de données, il semble que chaque corps professionnel ne peut se passer de l'autre. Dans un troisième volet de notre exposé, nous verrons tout d'abord que cette porosité propose de nouvelles façons de travailler en bibliothèque et en centre de données : en effet sans expertise scientifique les documentalistes sont limités dans leurs apports techniques pour les chercheurs et les centres de données astronomiques sont un exemple probant du lien qu'ils doivent entretenir. Ensuite, pour travailler en étroite collaboration il faut prendre en compte les compétences que chacun peut apporter (informatique, scientifique, documentaire) aussi dans l'idée qu'il faut traiter un nombre croissant de données issues de l'essor de la science ouverte. Enfin, la formation des personnels aux problématiques à la fois institutionnelle et de recherche est une première réponse à ces enjeux.

III) REPENSER LES RELATIONS ENTRE CHERCHEURS ET PROFESSIONNELS DE L'INFORMATION : L'APPORT DES DONNEES DE LA RECHERCHE A TRAVERS LES DONNEES ASTRONOMIQUES

1) ENTRE RECUEIL DES BESOINS ET ADAPTABILITE : OU SE SITUENT LES DOCUMENTALISTES ?

1. Traiter les données astronomiques : une étroite collaboration

Dans notre exposé, nous nous demandons comment se positionner en tant que professionnel de l'IST dans la chaîne de gestion des données de la recherche. Tout d'abord nous avons vu qu'il faut considérer que les données de la recherche sont des objets complexes, et plus particulièrement à travers l'exemple des données de la recherche en astronomie et en astrophysique. Ces « objets informationnels » spécifiques que sont les données de la recherche donnent lieu à des discussions entre différents corps de métiers afin de pouvoir les gérer au mieux. Il émane également de ces objets une volonté d'ouvrir la science pour la rendre plus accessible à tous les niveaux de la recherche comme de la société. Effectivement, c'est en ouvrant les données et en partageant les résultats de la recherche que des disciplines comme l'astrophysique se sont fortement développées. Cette volonté de développement, on la retrouve plus tard en France dans les politiques publiques émanant de l'Enseignement supérieur et de la recherche et ceci nous amène à repenser le rôle des bibliothèques et des documentalistes dans la recherche.

Cette troisième partie s'attache à montrer comment l'ouverture des données et le partage scientifique, à l'image du partage scientifique en astronomie et en astrophysique met en évidence les liens qui unissent les chercheurs avec le monde de l'IST. Cela transparaît en ce que les chercheurs doivent énoncer leurs besoins et les documentalistes s'acculturer à la discipline qu'ils traitent. Cette tension entre recueil des besoins et adaptabilité du documentaliste sous-tend à l'heure actuelle le développement de la science ouverte et de la gestion des données de la recherche. Il paraît alors nécessaire, pour répondre à cette problématique, d'accentuer le dialogue entre chaque partie (chercheurs, professionnels de l'IST, institutions) et de proposer des formations adéquates pour monter en compétences.

A la question de savoir si l'on peut traiter les données astronomiques sans connaissance dans la discipline, la réponse est contrastée voire négative. Nous l'avons vu, les données astronomiques sont des objets complexes, et aujourd'hui avec une production massive des données dans cette discipline, les documentalistes ont besoin de dialoguer avec les chercheurs pour identifier ces données et les traiter au mieux. Comment alors appréhender leur gestion en tant que professionnel de l'IST ? C'est une question que je suis allée poser

auprès des personnes qui traitent quotidiennement ces données. La façon dont fonctionne aujourd'hui un CDS comme celui de Strasbourg expose une manière de travailler innovante, en constante discussion avec les informaticiens et les chercheurs.

A ce titre l'exemple – quasi « extrême » – du CDS de Strasbourg est comme une porte ouverte sur la gestion future des données de la recherche. En effet, le CDS sélectionne spécifiquement les données qui proviennent des publications.

Afin de rendre ce travail de fouille de texte possible pour des non-scientifiques spécialisés du domaine, des réunions inter-équipes sont organisées et permettent de passer au crible les articles scientifiques. Les trois corps de métiers – documentalistes, informaticiens et astronomes – se réunissent ainsi deux fois par semaine. Ils vérifient la qualité des données, et grâce au savoir des astronomes ils identifient les données intéressantes pour la recherche. La coordinatrice de l'équipe, Soizick Lesteven, avec qui j'ai pu m'entretenir également, expose l'importance de la collaboration entre les chercheurs et les documentalistes :

« […] là, il y a vraiment une expertise des astronomes pour savoir comment intégrer les données ». Soizick Lesteven est elle-même titulaire d'une thèse en astrophysique, ce qui lui confère une double expertise, et scientifique, et documentaire, dans sa manière de gérer son équipe.

Enfin l'expertise des chercheurs comme des documentalistes est sollicitée pour le travail de l'établissement des standards et de ce qu'on peut appeler un travail de *cross identification*. On pourrait définir le travail de *cross identification* par le fait de croiser des données entre elles afin d'en tirer des constantes qui pourront devenir par la suite des données fondamentales. C'est en somme repérer, identifier des objets et croiser les différentes sources et noms sous lesquelles elles apparaissent, pour en tirer des constantes. Le fait de repérer puis nommer des données fondamentales est un point sensible au CDS comme dans toute la communauté documentaire lorsqu'il s'agit de mettre en commun le travail de fouille. Pour quelles raisons est-ce un point complexe dans la gestion des données de la recherche en astronomie, comme dans tout autre domaine ?

Pour en donner un exemple concret, lorsqu'une équipe d'astronomes va étudier un le ciel ou quand elle va prendre des mesures, une description des objets repérés va être faite. Ces objets vont être nommés or, dans cette communauté scientifique, ce n'est pas une habitude de vérifier si certains objets ont déjà été nommés. Certaines équipes vont y faire plus attention et utiliser les termes préconisés par les grandes bases de données qui effectuent ce travail de *cross identification* (notamment ADS et le CDS de Strasbourg). Or, si c'est un satellite qui observe tout le ciel il ne va pas effectuer de lui-même ce travail de *cross identification*. Il est donc possible de se retrouver avec une nouvelle publication dont les objets sont nommés avec d'autres noms.

Nous pouvons donner un autre exemple avec les métadonnées associées aux publications scientifiques déposées dans des archives ouvertes, notamment le cas de

HAL[62]. HAL est une archive ouverte générique où il est possible de déposer des documents provenant de disciplines différentes. Un travail important a été mené par la communauté documentaire et scientifique utilisant la plateforme pour savoir quelles sont les métadonnées « fondamentales » attendues pour chaque document (article dans une revue, communication dans un congrès, chapitre d'ouvrage, etc) et qui soient assez génériques pour que toutes les disciplines puissent déposer dans l'archive.

Ainsi le travail de nomenclature dans les catalogues ou pour les référentiels est une étape essentielle, et elle ne peut se faire sans l'expertise des chercheurs. C'est là un point clé du travail des documentalistes au quotidien qui notamment gèrent les dictionnaires de nomenclatures : ils décident le nom qu'ils considèrent comme le plus important et le plus utilisé en suivant les recommandations internationales et l'usage le plus fréquent dans la littérature. Aussi ces noms ne sont pas choisis au hasard : il y a des règles de nomenclatures au niveau international pour qu'ils soient facilement identifié et durables dans le temps. Il y a quelques décennies la communauté avait pour habitude d'utiliser le type d'objet dans le nom. Par exemple, la « nébuleuse d'Andromède » est un nom historiquement utilisé pour désigner la Galaxie d'Andromède. Or le terme « nébuleuse » signifiait « objet flou ». Aujourd'hui on a de meilleures données et on arrive à identifier cet objet, « nébuleuse » ne signifiant plus quelque chose de précis, il a été délaissé par la communauté scientifique. Or il se peut que ce nom réapparaisse dans certaines publications et l'idée est d'inciter la communauté scientifique (comme documentaire) à utiliser des noms d'objets précis pour homogénéiser les données et les publications. Identifier et standardiser les nomenclatures c'est en somme ce que les principes FAIR sont à la standardisation du partage des données et c'est ce sur quoi aboutit une gestion efficace des données.

De cette façon l'apport scientifique des chercheurs est direct dans le travail quotidien des documentalistes. En retour, les documentalistes sont régulièrement formés lors et en dehors des réunions inter équipes, ils deviennent par conséquent experts dans les données astronomiques. Leur niveau de connaissance devient important bien que l'intervention des chercheurs soit essentielle. Esther Collas souligne ceci en rapportant lors de notre entrevue :

« […] chaque fois que nous avons une question pendant le traitement d'une référence ou d'un article, on va directement dans le bureau des astronomes pour avoir une meilleure compréhension du contexte. L'équipe des astronomes du CDS est pour ça spécialisée dans chaque domaine, on trouve des spécialistes des longueurs d'ondes, de différents objets ou d'études. Cela nous permet d'avoir des compétences dans tous les domaines qu'on retrouve dans les publications ».

Dans cette remarque, on voit que ce qui fait la force de l'équipe est l'interdisciplinarité et la collaboration entre les divers corps de métiers. Cette situation de travail concrète reflète l'importance de la mise en commun des compétences de chacun y compris ici au

[62] Archive ouverte pluridisciplinaire française gérée par le Centre pour la communication scientifique directe (CCSD)

sein de l'équipe des chercheurs. Plus elle est diversifiée et plus ils et elles pourront répondre aux questions précises des documentalistes.

2. L'informatique au cœur du métier

Dans les missions confiées aux documentalistes dans la gestion des données de la recherche, une part importante est également donnée à l'informatique et à la connaissance de langage de requête ou de langage de programmation. En somme tout un panel de connaissances techniques sont demandées aux documentalistes. Lors de notre entretien Esther Collas confie qu'elle a appris de la programmation lors de son entrée en poste, bien qu'elle eût des connaissances dans quelques langages de programmation. En effet les documentalistes manipulent quotidiennement la programmation. Il y a des scripts à disposition qui peuvent être utilisés pour interroger directement les bases de données du CDS. Il y a différents types de manière d'utiliser les bases de données pour extraire des données et une bibliothèque de requêtes est mise à disposition des documentalistes. C'est une sorte de boîte à outils de programmes préparés par les informaticiens ou en lien avec eux, un « package » qui leur est mis à disposition pour interroger facilement les bases de données depuis des scripts python. Il est également possible d'interroger les bases directement depuis les pages web mais cette façon d'opérer est la moins utilisée par les documentalistes. C'est ce qu'utilisent le grand public et notamment les astronomes amateurs (pour obtenir les positions précises de chaque objet afin de savoir où pointer le télescope).

En outre, elle a appris à utiliser des logiciels techniques internes lors de son entrée en poste. L'utilisation technique des logiciels internes a été apprise lors de l'entrée en poste. Pour cette dernière phase d'apprentissage, des formations internes sont proposées par les astronomes ou les documentalistes qui ont déjà acquis ces connaissances. En termes de compétences, la capacité d'adaptation et la curiosité sont de rigueur afin de pouvoir exercer ses missions.

Le centre de données devient ici un lieu où les informaticiens et les documentalistes travaillent de concert. C'est aujourd'hui le cas des lieux de travail où les plateformes à gérer demandent des compétences en informatique. Il est important que les documentalistes et les informaticiens puissent se comprendre lorsqu'ils travaillent ensemble, en témoignent les métiers autour de la gestion de ressources documentaires ou les *data steward*[63].

Cependant, cela est à mettre en perspective avec la difficulté que peut représenter cette multi-compétence pour les documentalistes. Encore une fois, un travail d'acculturation à l'informatique est nécessaire et les établissements doivent pouvoir accompagner au mieux cette mise à niveau. Pour cela, des formations sont par exemple proposées par le CNRS

[63] Coordinateur de données

pour tout agent (contractuel ou titulaire) qui souhaite acquérir des compétences informatiques « pour les non-informaticiens ». Des formations proposées par des partenaires externes sont également financées pour les agents. Par exemple le réseau de formation Orsys Formation propose une formation « Base de données et langage SQL pour les non-informaticiens » qui vise à comprendre les bases de données relationnelles et être capable de construire des requêtes avancées pour extraire des données. Aussi les plateformes animent des formations, par exemple *Recherche Data Gouv* propose un webinaire « Introductions aux APIs de *Recherche Data Gouv* » pour apprendre à construire des requêtes qui interrogent leur service.

2) DES ESPACES DEDIES POUR IMPLIQUER LES ACTEURS DE LA RECHERCHE

1. Créer un lieu propice à l'échange

Outre l'investissement des établissements, des espaces permettent la discussion entre les acteurs de la recherche. En effet les groupes de recherche, outre l'intérêt de se regrouper autour d'un sujet de recherche qu'on en commun les membres, laissent place à la discussion. Les groupes de recherches rendent possible la collaboration en faisant se rencontrer les chercheurs. Le travail est alors rendu collectif de manière très concrète en avançant sur des questions d'intérêt commun. J'ai eu l'occasion de m'entretenir avec Nicolas Lumineaux, co-responsable d'un groupe de recherche (GDR) centré sur la gestion de gros volumes de données en astronomie, le GDR Masse de Données, Informations et Connaissances en Sciences (MaDICS) qui a produit l'atelier BigData4Astro[1]. Cet atelier propose aux participants d'échanger à propos des données qu'ils ont à traiter dans le cadre de leur propre projet de gestion de données. Nicolas Lumineaux m'indique que le projet se rencontre autour de deux réunions annuelles, les Symposiums MaDICS. Il qualifie ces symposiums de « vivers d'échange » En effet, ils permettent de faire se rencontrer des communautés qui ont parfois du mal à se comprendre, par exemple les chercheurs qui travaillent dans le calcul de haute performance et les chercheurs en astronomie pure. Dans ce cadre d'échange, chacun vient avec ses « propres données » et ils discutent ensemble sur leur caractérisation et leur potentielle gestion. Le GDR BigData4Astro développe alors essentiellement des services pour les chercheurs, services dont ils travaillent à leur amélioration constamment.

Au-delà de la spécificité du Big Data[64] et de l'astronomie, le but premier d'un GDR est de créer un lieu propice à l'échange. Dans le cadre des ateliers BigData4Astro, Nicolas Lumineaux l'explicite ainsi :

« Le lien se crée car on va aller chercher des carnets d'adresses pour voir qui va présenter des travaux récents. On se base sur le principe de communication et sur de l'animation scientifique conviviale, sans le stress de la publication. C'est un moment agréable. Derrière il peut se monter des collaborations. Nous [les organisateurs] on est juste là pour donner un espace où on peut se retrouver et échanger, et cet échange se base sur les

[64] Données massives, mégadonnées

travaux personnels des chercheurs. Viennent se greffer à ces rencontres des conférences entre physiciens et astrophysiciens, cela permet de se rencontrer. »

Nicolas Lumineaux témoigne ici de l'importance de se retrouver entre chercheurs pour discuter de comment ils peuvent gérer leurs données. Une part importante est donnée à la convivialité et ces espaces sortent les chercheurs du cadre de l'évaluation. La publication n'est pas un enjeu, ils viennent dans ces ateliers pour discuter de leur production scientifique dans un cadre qui diffère de l'institution ou de celui du laboratoire. Il en résulte que des collaborations peuvent se monter. Ces lieux apparaissent alors comme des lieux propices au partage de compétences et d'une certaine matière comme des lieux propices à une science plus ouverte. On peut avancer l'idée que partager sur ses pratiques est une des manières de faire de la science ouverte et que ces initiatives sont à favoriser. En se structurant et en recueillant leurs besoins entre eux, les chercheurs sont plus à même de transmettre ensuite ces besoins aux documentalistes quant à la gestion des données de la recherche.

Les groupes de recherche aboutissent à l'établissement de standards, en témoigne également la construction de l'International Virtual Observatory Alliance (IVOA) : « The International Virtual Observatory Alliance (IVOA) is an organisation that debates and agrees the technical standards that are needed to make the VO possible »[65]. L'IVOA se réunit tous les 6 mois et des travaux sont discutés pendant ces sessions. A partir de là des discussions ont lieu et des *working draft*[66] sont produits. Ces notes peuvent ensuite être discutées par la communauté de chercheurs comme des documentalistes pendant des mois voire des années avant d'être validées. Ensuite le comité exécutif de l'IVOA (composé de personnes qui représentent les divers observatoires virtuels nationaux) appose sa validation finale.

J'ai pu m'entretenir avec un ingénieur de recherche qui est membre de l'IVOA depuis ses débuts. Il m'explique ainsi que leur rôle est de faire un lien entre la communauté astronomique internationale et les documentalistes dans une perspective « science ouverte » :

« Certains documentalistes impliqués dans l'IVOA participent beaucoup à des projets nationaux et internationaux sur la science ouverte en général. La science doit devenir la plus transparente et réutilisable possible. On mène une réflexion importante sur l'utilisation des données par les scientifiques. »

Cette remarque à propos de l'IVOA corrobore notre idée selon laquelle les lieux propices à la discussion entre scientifiques (et documentalistes) participent à une science plus ouverte. Dans le cadre de la gestion des données de la recherche ils permettent aux scientifiques de construire leurs besoins. Ce qui pointe est la question des compétences que chaque corps de métier apporte. En effet la chaîne de gestion des données de la

[65] « L'IVOA est une organisation qui débat et valide les normes techniques nécessaires pour rendre l'Observatoire virtuel possible. »

[66] Désigne un papier qui réunit toutes les versions d'un document technique de travail émanant d'un groupe de recherche

recherche demande de multiples compétences et pour trouver sa place, le documentaliste dialogue avec les chercheurs et se forme pour mieux comprendre les données qu'il gère. Pour construire ses connaissances il ne suffit pas au documentaliste d'être impliqué dans ses missions. Les établissements doivent aussi permettre le dialogue et la formation (entre chercheurs comme entre documentalistes) afin d'assurer la coopération et le développement des services d'appui à la recherche. Car comment parvenir à une gestion efficiente sans une implication totale de chaque partie ?

3) UNE IMPLICATION OPERATIONNELLE DES ETABLISSEMENTS DE RECHERCHE

1. Vers une montée en compétences des professionnels de la documentation

En effet les institutions universitaires et notamment les bibliothèques universitaires et de recherche sont aujourd'hui confrontées aux problématiques de la gestion des données de la recherche. Notre hypothèse est aque ces institutions doivent de plus en plus s'acculturer au domaine scientifique qu'ils accompagnent, en témoigne ce que font les documentalistes rattachés à des lieux comme les centres de données astronomiques. Finalement la particularité de la gestion des données astronomiques peut s'étendre à tous types de données de la recherche.

Cette acculturation des bibliothécaires à la communauté scientifique qu'ils et elles servent est aussi bien présente, et Véronique Stoll, directrice de la bibliothèque de l'Observatoire de Paris m'en fait part lors d'un entretien :

« Dans le domaine de l'astronomie, on est obligé de s'acculturer en arrivant. Il reste à apprendre les spécificités disciplinaires et effectuer une montée en compétences ».

Plus encore qu'une acculturation scientifique, les questions liées à la gestion des données de la recherche et de la science ouverte lui ont également demandé de s'adapter. En tant que personnel de direction d'une bibliothèque, Véronique Stoll me confie que lorsqu'elle a suivi sa formation de conservatrice des bibliothèques, l'étude des données de la recherche restait marginale, et ce n'était pas au cœur des politiques des bibliothèques. Il s'agissait donc en arrivant à la direction de la bibliothèque de l'Observatoire d'avoir un socle de connaissance sur la science ouverte et la gestion des données, et dans un second temps d'adapter ce socle à la spécificité astronomique et astrophysique. Le tout non seulement pour une bonne gestion des données mais aujourd'hui pour une bonne ouverture de celles-ci.

Les bibliothèques universitaires sont des établissements structurés et qui travaillent réseau. On y trouve une communauté d'échanges de pratiques, en relate par exemple le Système universitaire de documentation (Sudoc) qui est un catalogue alimenté collectivement par les bibliothèques. Il n'est donc pas spécifiquement propre aux bibliothèques d'Observatoires comme celle de l'Observatoire de Paris de suivre les pratiques métiers (catalogage, ouverture des données et des publications). Or, Véronique Stoll m'indique que les chercheurs ou les professionnels de l'IST viennent chercher

conseil auprès de la bibliothèque de l'Observatoire sur des problématiques propres à l'astronomie, et c'est à ce titre que la bibliothèque est présente au sein de l'université Paris Sciences & Lettres (PSL). Sur les sujets propres à la science ouverte également, l'expertise de la bibliothèque de l'Observatoire est consultée.

On voit par là qu'accentuer la formation des futurs bibliothécaires, conservateurs d'État et professionnels de l'IST à la gestion des données de la recherche est un enjeu de taille. Les établissements abordent dorénavant des formations à la gestion des données de la recherche et plus généralement à la science ouverte. En relate par exemple la formation des conservateurs d'État dispensé par l'École nationale supérieure des sciences de l'information et des bibliothèques (ENSSIB)[67] où un module entier d'enseignements professionnels est dédié aux données et à la science ouverte. Aussi l'Université Claude Bernard Lyon 1 propose un module de cours entièrement dédié aux données de la recherche depuis 2021 dans le cadre du Master Information et Médiation scientifique et technique (IMST), outre les diverses formations proposées par les établissements[68]. Ainsi l'enseignement des problématiques liées à la gestion des données de la recherche commence à pointer dans les modules d'enseignement des universités et grandes écoles. Nous ferons ici l'hypothèse que ces modules seront de plus en plus présents, afin de répondre à la demande croissante du milieu scientifique.

2. Un défi pour le pilotage des établissements

Une des problématiques majeures que soulève la gestion des données de la recherche pour les personnels de direction des institutions es également d'être au fait de ces nouveaux enjeux au sein du pilotage d'un établissement ou d'une équipe. Un travail d'identification des besoins et de l'évolution des métiers est une tâche de fond qui est à prioriser, ainsi que celle de repérer les nouvelles compétences dont les établissements ont besoin. Lors de mon entretien avec Soizick Lesteven, en tant que responsable de l'équipe des documentalistes, elle insiste sur cet aspect :

« On va chercher des personnes qui ont un minimum de connaissances scientifiques (mathématique, chimie, botanique), des personnes à l'aise avec les données, avec les chiffres. Le niveau de compréhension [des données] est aujourd'hui beaucoup plus élevé. Avant on recrutait niveau assistant, maintenant on recrute niveau ingénieur d'études. Le travail s'est densifié et complexifié. Il faut une rigueur très importante pour assurer la qualité de ce qu'on distribue. »

Ici Soizick Lesteven souligne les défis que posent le recrutement de nouveau personnel en appui à la recherche. Le niveau de compréhension des données s'étant intensifié, il en va de même pour le niveau de recrutement. Les formations internes

[67] Livret de formation des conservateurs disponible à cette adresse : https://www.enssib.fr/sites/enssib.fr/files/inline-files/LivretDCB32-web.pdf

[68] Le service « Question/Réponse » de l'ENSSIB décrit dans cette réponse à un usager les différentes offres de formations relatives aux données de la recherche : https://www.enssib.fr/services-et-ressources/questions-reponses/formation-continue-gestion-des-donnees-de-la-recherche

viennent en cela compléter les connaissances des personnels débutants et confirmer celles des personnels qui ont acquis de l'expérience. Ce constat n'est pas propre à la gestion des données astronomiques bien qu'elle en soit un exemple tout à fait exacerbé. Car Véronique Stoll, en tant que conservatrice des bibliothèques relate la même complexité, ainsi que de futurs enjeux :

« On doit s'interroger sur comment se passe la formation des personnes qu'on recrute, comment les collègues font cette montée en compétences. [En somme] comment faire en sorte que les bibliothèques soient des acteurs pivots sur l'ensemble des établissements. Faire de la bibliothèque pas simplement un fournisseur de documentation mais aussi un fournisseur de services. Enfin La 3e partie qui attend les établissements c'est tout ce qui concerne la préservation des codes et logiciels et qui va nécessiter un autre mode de réflexion. ».

Véronique Stoll fait part de la complexité des missions de pilotage qui sous-tend les établissements quant à la gestion des données de la recherche. Elle souligne l'importance de la place des bibliothèques dans cette gestion et dans les questions plus générales de science ouverte. Les établissements sont de véritables « fournisseur de services » et pour cela la montée en compétences des personnels est une étape importante. Il reste alors à penser les futurs services que vont pouvoir fournir les bibliothèques et les établissements de recherche, notamment à travers les nouveaux types de documents et de données à conserver. Pour cela, la mise en commun des compétences des différents acteurs de la recherche sera à privilégier et à développer.

CONCLUSION

Les données de la recherche sont des objets complexes et leur gestion soulève des problématiques qui croisent divers intérêts. Il apparait que les établissements qui viennent en appui à la recherche tels les bibliothèques universitaires ont un rôle structurant pour une gestion efficace des données. En outre, cet aspect fonctionnel des bibliothèques est une continuité de ses missions fondamentales d'un service public tourné vers la science. En focalisant notre étude sur les données de la recherche astronomique, on voit que les enjeux d'ouverture et de création d'outils performants quant à leur (ré)utilisation se pose depuis toujours dans certaines disciplines scientifiques. A la question de savoir quelle place occupent les professionnels de l'information scientifique et technique dans la chaîne de gestion des données, on répond qu'ils occupent un rôle fondamental. Ce rôle tient en ce qu'ils accueillent les besoins des chercheurs et sont en mesure de mettre en ordre les données à chaque étape de leur cycle de vie.

Cependant, nous l'avons vu, ce recueil des besoins de la part des documentalistes ne peut se faire sans une communauté scientifique elle-même au fait de ce dont elle a besoin. Et la gestion des données par les documentalistes est limitée s'ils ne sont pas formés à la discipline qu'ils accompagnent. Sur la formation des personnels en appui à la recherche, c'est là que s'insèrent également des problématiques de pilotage des équipes.

En effet la place que prend les services dédiés aux données de la recherche, dans les formations des futurs professionnels de l'information comme dans les établissements va en grandissant. Le milieu astronomique est en ce sens une prévision sur ce qui attend les équipes en termes de partage d'échanges et de partage des besoins. Ce présent mémoire constitue une première étape dans la définition des données de la recherche astronomique et permet d'entrevoir les enjeux qui attendent les établissements en appui à la recherche quant à la gestion des données. Il serait pertinent d'observer comment s'est constituée (et comment évolue) la chaîne de gestion des données de la recherche dans d'autres disciplines afin de comprendre plus profondément encore le rôle des professionnels de l'information scientifique et technique, à l'heure d'une science qui s'ouvre et qui produit davantage.

SOURCES

Tableau récapitulatif des entretiens				
NOM Prénom	**Poste, institut de rattachement**	**Date de l'entretien**	**Durée**	**Accord pour utilisation des données personnelles**
ABOUDARHAM Jean	Chercheur au Laboratoire d'Etudes Spatiales et d'Instrumentation en Astrophysique (LESIA) et responsable scientifique pour BASS2000	21/04/2023	1h	Oui
COLLAS Esther	Chargée du traitement des données scientifiques, Centre de données astronomiques de Strasbourg	11/01/2023	1h15	Oui
LESTEVEN Soizick	Ingénieure de recherche au Centre de données astronomiques de Strasbourg - Responsable de l'équipe des documentalistes	14/02/2023	40min	Oui
LUMINEAUX Nicolas	Enseignant chercheur en informatique à l'Université Claude Bernard Lyon 1, membre de l'équipe Bases de Données du LIRIS et du GDR MaDICS BigData4Astro	21/03/2023	45min	Oui
STOLL Véronique	Directrice de la Bibliothèque de	06/01/2023	45min	Oui

	l'Observatoire de Paris - Site Meudon			
-	Administratrice du portail HAL de l'Institut National des Sciences de l'Univers (INSU)	11/01/2023	45min	Non communiqué
-	Ingénieur de recherche à l'Observatoire Astronomique de Strasbourg, membre de l'IVOA et président de l'IG sur la conservation des données de l'IVOA	05/05/2023	45min	Non communiqué

BIBLIOGRAPHIE

Articles dans une revue

FEUILLEBOIS, Geneviève. Les bibliothèques d'observatoires. [en ligne]. *Bulletin des bibliothèques de France (BBF)*, 1958, n° 3, p. 175-184. Disponible sur https://bbf.enssib.fr/consulter/bbf-1958-03-0175-001. [consulté le 18 janvier 2023]

GENOVA, Françoise. Du nécessaire partage des données scientifiques. : l'exemple de l'astronomie. *Ar(abes)que* [en ligne]. 2014, (73), 12–13. Disponible sur : https://publications-prairial.fr/arabesques/index.php?id=999 [consulté le 8 décembre 2022]

RIVIÈRE, Pascal. Qu'est-ce qu'une donnée ? Impact des données externes sur la statistique publique. *Courrier des Statistiques* [en ligne]. 2020, (5) Disponible sur : https://www.insee.fr/fr/information/5008707?sommaire=5008710 [consulté le 9 décembre 2022].

SCHÖPFEL, Joachim. Hors norme ? Une approche normative des données de la recherche. *Revue COSSI : communication, organisation, société du savoir et information* [en ligne]. 2018, Processus de normalisation et durabilité de l'information, 5. Disponible sur : https://hal.science/hal-01979798/document [consulté le 10 décembre 2022]

Articles de blogs scientifiques

APPÉRÉ, Edern. La science ouverte sur orbite. *savoirs-archives.unistra.fr* [en ligne]. 10 octobre 2022. Disponible sur : https://savoirs-archives.unistra.fr/recherche/science-ouverte-un-nouveau-contexte-pour-la-recherche/la-science-ouverte-sur-orbite/ [consulté le 10 novembre 2022]

CNRS, Info. ESCAPE : les physiciens préparent leur science ouverte. *CNRS.* [en ligne]. 20 octobre 2023. Disponible sur : https://www.cnrs.fr/fr/cnrsinfo/escape-les-physiciens-preparent-leur-science-ouverte [consulté le 10 mars 2023]

CNRS, Info. Lancement du projet ESCAPE. *IN2P3.* [en ligne]. 19 novembre 2018. Disponible sur : https://www.in2p3.cnrs.fr/fr/cnrsinfo/lancement-du-projet-escape [consulté le 10 janvier 2023]

YOUNG, Andrew, Stefaan G. VERHULST et Andrew J. ZAHURANEC. Comment la science ouverte peut s'inspirer du libre accès aux données publiques. *The Conversation* [en ligne]. 16 mars 2021. Disponible sur :

https://theconversation.com/comment-la-science-ouverte-peut-sinspirer-du-libre-acces-aux-donnees-publiques-157091 [consulté le 18 février 2023]

Autres publications scientifiques

ARCHIMBAUD, Jean-Luc. *Identifiants des documents numériques : ISBN, ISSN, URL, DOI, OpenURL...* [en ligne]. 26 janvier 2015. Disponible sur https://archivesic.ccsd.cnrs.fr/sic_01068135v2/document [consulté le 5 février 2023]

Chapitres d'ouvrages

MAUREL, Lionel. La réutilisation des données de la recherche après la loi pour une République numérique. *La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques*. [en ligne]. Presses Universitaires de Provence, 2018. Disponible sur : https://hal.science/hal-01908766 [consulté le 22 février 2023]

Communication dans un congrès

DEBRAY, Bernard. *L'observatoire virtuel astronomique : permettre l'accès à tous les données et les services pour la recherche*. [en ligne]. Journée « Base de données », Besançon, le 21 mars 2013. Disponible sur : https://theta.obs-besancon.fr/IMG/pdf/BDebrayJourneeBD20130321.pdf [consulté le 8 décembre 2022]

GENOVA, Françoise, DUVERT, Gilles. *L'archivage des données en astronomie*. [en ligne]. 20 novembre 2019. Datarchivage : archivage numérique des données de recherche, Grenoble, le 20 novembre 2019. Disponible sur : https://dataarchivage.sciencesconf.org/data/pages/ArchivageDonneesAstronomie_20nov2019_Genova_Duvert.pdf [consulté le 9 décembre 2022]

GENOVA, Françoise. *Les données astronomiques de l'Observatoire Virtuel*. [en ligne]. 15 novembre 2017. Colloque Data BFC : ouvrir et gérer les données de la recherche en Bourgogne Franche-Comté, Besançon, les 13 et 15 novembre 2017. Disponible sur : https://databfc.sciencesconf.org/data/FGenova_DataBFC_IVOA.pdf [consulté le 5 février 2023]

OBSERVATOIRE DE PARIS. *Numériser les données astronomiques contemporaines.* [en ligne]. Journées Collex-Persée, Paris, le 4 avril 2019. Disponible sur : https://www.collexpersee.eu/wp-content/uploads/2018/04/Pr%C3%A9sentationCollExPro19Projet20.pdf [consulté le 8 décembre 2022]

SCHAAFF, André. *Vers le « Big Data », l'exemple de l'astronomie*. [en ligne]. Séminaire Audipog, Paris, le 9 avril 2015. Disponible sur : http://www.audipog.net/pdf/seminaires/seminaire_2015/pres03.pdf [consulté le 8 décembre 2022]

THIAULT, Florence. *Data librarian et services aux chercheurs en bibliothèque universitaire : de nouvelles médiations en émergence*. [en ligne]. 7e conférence Document numérique et société. Humains et données : création, médiation, décision, narration, Nancy, septembre 2020. Disponible sur : https://hal.science/hal-02972705v1/file/Datalibrarian-THIAULT-%20DOCsocHAL.pdf [consulté le 15 décembre 2022]

Documents de travail

FITS Working Group. *Definition of the Flexible Transport System (FITS)*. [en ligne]. Version 4.0 mise à jour le 22 juillet 2016. Disponible sur : https://fits.gsfc.nasa.gov/standard40/fits_standard40draft1.pdf [consulté le 4 février 2023]

EVANS, J. , KIRSCH, R. and NAGEL, R. *Workshop on Standards for Image Pattern Recognition*. [en ligne]. National Institute of Standards and Technology, Gaithersburg, MD, 1977. Disponible sur : https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nbsspecialpublication500-8.pdf [consulté le 9 décembre 2022]

Documents juridiques

Circulaire du 26 mai 2011 relative à la création du portail unique des informations publiques de l'Etat « data.gouv.fr » par la mission « Etalab » et l'application des dispositions régissant le droit de réutilisation des informations publiques. [en ligne]. Disponible sur : https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000024072788

Décret n° 2011-996 du 23 août 2011 relatif aux bibliothèques et autres structures de documentation des établissements d'enseignement supérieur créées sous forme de services communs. [en ligne]. Disponible sur : https://www.legifrance.gouv.fr/loda/id/JORFTEXT000024497856 [consulté le 25 octobre 2022]

Numéro spécial de revue

WELLS, D., E. GREISEN et R. HARTEN. FITS : A Flexible Image Transport System. *Astronomy & Astrophysics* [en ligne]. 1981, (Supplément Series 44), 363–370. Disponible sur : https://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?bibcode=1981A&AS...44..363W&db_key=AST&page_ind=7&plate_select=NO&data_type=GIF&type=SCREEN_GIF&classic=YES [consulté le 2 mars 2023].

Ouvrages

BORGMAN, Christine L. *Qu'est-ce que le travail scientifique des données ?* [en ligne]. Traduit de l'Anglais par Charlotte MATOUSSOWSKY. OpenEdition, 2020. Disponible sur : https://books.openedition.org/oep/14692 [consulté le 22 février 2023]

ROBIN, Agnès. *Droit des données de la recherche. Science ouverte, innovation, données publiques*. Larcier, 24 mars 2022.

Rapports

COURBE, Thomas. *Recherche Spatiale*. [en ligne] Programme n°193. Budget général, mission interministérielle, projets annuels de performances. Annexe au projet de loi de finances pour 2022. Disponible sur : file:///C:/Users/El%C3%A9onore%20Kolar/Downloads/FR_2022_PLF_BG_PGM_193.pdf [consulté le 25 octobre 2022]

EUROPEAN COMMISSION. *Providing researchers with the skills and competencies they need to practise Open Science*. *Report of the Working Group on Education and Skills under Open Science.* [en ligne] Juillet 2017. Disponible sur : https://euraxess.ec.europa.eu/sites/default/files/policy_library/ec-rtd_os_skills_report_final_complete_2207_1.pdf

IFLA. *Déclaration de l'IFLA sur le libre accès à la littérature scientifique et aux documents de la recherche*. [en ligne] 5 décembre 2003. Disponible sur : https://www.enssib.fr/bibliotheque-numerique/documents/1972-declaration-de-l-ifla-sur-le-libre-acces-a-la-litterature-scientifique-et-aux-documents-de-la-recherche.pdf [consulté le 15 décembre 2022]

LETROUIT, Carole, CACHARD, Pierre-Yves, DUPUIS, Monique, FROMENT, Bernard. *La place des bibliothèques universitaires dans le développement de la science ouverte. Rapport à madame la ministre de l'Enseignement supérieur, de la Recherche et de l'Innovation. N° 2021-2022.* [en ligne] Février 2021. Disponible sur : file:///C:/Users/El%C3%A9onore%20Kolar/Downloads/igesr-rapport-2021-022-place-bibliotheques-universitaires-developpement-science-ouverte-pdf-88074.pdf [consulté le 25 octobre 2022]

MIKULSKI ARCHIVE FOR SPACE TELESCOPES. *HST Reports*. [en ligne] Disponible sur : https://archive.stsci.edu/hst/bibliography/pubstat.html] [consulté le 4 février 2023]

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR DE LA RECHERCHE ET DE L'INNOVATION. *Feuille de route nationale des infrastructures de recherche 2021*. [en ligne]. Version mise à jour le 17 mars 2022. Section « Astronomie et astrophysique ». p.17-40. Disponible sur : https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2022-03/feuille-

de-route-nationale-des-infrastructures-de-recherche---2021-v2--17318.pdf [consulté le 2 novembre 2022]

OCDE. *Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*. [en ligne]. 2007. Disponible sur : https://www.ouvrirlascience.fr/wp-content/uploads/2018/11/Principes-lignes-directrices-de-l%E2%80%99OCDE-pour-l%E2%80%99acc%C3%A8s-aux-donn%C3%A9es_38500823.pdf [consulté le 25 octobre 2022]

OCDE. *Making Open Science a Reality, OECD Science, Technology and Industry Policy Papers, No. 25* [en ligne]. 2015. Disponible sur : http://dx.doi.org/10.1787/5jrs2f963zs1-en [consulté le 25 octobre 2022]

Ressources pédagogiques en ligne

CDS, Strasbourg. *Vers le Big Data ? Exemple de la gestion des données astronomiques au Centre de Données astronomiques de Strasbourg* [en ligne]. 2015. Disponible sur : https://docplayer.fr/2847175-Vers-le-big-data-exemple-de-la-gestion-des-donnees-astronomiques-au-centre-de-donnees-astronomiques-de-strasbourg.html?fbclid=IwAR0wDnu-d8rbfF4xY9HiuzePS-B827wj601vLu8B2kIh-Ekkzfv22Yl08xE [consulté le 18 décembre 2022]

CDS, Strasbourg. *La curation des données dans VizieR* [en ligne] Disponible sur : https://cdsweb.u-strasbg.fr/~landais/presentations/curationVizieR.pdf [consulté le 17 janvier 2023]

CDS, Strasbourg. *NED and SIMBAD Conventions for Bibliographic Reference Coding*. [en ligne]. Disponible sur : https://simbad.cds.unistra.fr/guide/refcode/refcode-paper.html [consulté le 17 janvier 2023]

GRAVET, Romaric. *Articles scientifiques, données : comment ça marche ? Exemple de l'astronomie.* [en ligne] 19 novembre 2021. [consulté le 17 janvier 2023] Disponible sur : https://media.afastronomie.fr/RCE/PresentationsRCE2020/SAB-19-GRAVET.pdf

LANDAIS, Gilles. *Retour d'expérience sur les méthodes développées par le centre de données astronomiques de Strasbourg.* [en ligne] Disponible sur : https://indico.cern.ch/event/338461/contributions/1730296/attachments/663158/911592/predon_CDSSI.pdf [consulté le 17 janvier 2023]

ANNEXES

Table des annexes

ANNEXE 1 : PLAN DE GESTION DE DONNEES DU PROJET ADVANCED RADIO ASTRONOMY IN EUROPE – RADIONET

H2020 Grant Agreement No. 730562 – RadioNet

PROJECT TITLE:	Advanced Radio Astronomy in Europe
STARTING DATE	01/01/2017
DURATION:	48 months
CALL IDENTIFIER:	H2020-INFRAIA-2016-1
TOPIC:	INFRAIA-01-2016-2017 Integrating Activities for Advanced Communities

Deliverable 1.2 Data Management Plan

Due date of deliverable: 2017-06-30

Actual submission date: 2017-06-30

Leading Partner: MPG

Document information

Document name:	Data Management Plan
Type	Report
WP	WP1 – RadioNet Management
Version date:	2017-06-30
Authors (Institutes)	G. Kruithof (ASTRON), I. Rottmann (MPG)

Dissemination Level

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Introduction

This deliverable provides the RadioNet data management plan (DMP) version 1.0.

It is important to mention that the RadioNet project does not create data in the true sense of the word. However, one of the activities is designed to develop software for astronomical data. Thus some of the astronomical data will be used in the development phase of the software testing.

This document outlines how the collected or created research data will be managed during and after the RadioNet project. The DMP describes, which standard and methodology will be followed for data collection and generation, and whether and how the data will be available.

This document follows the template (ver. 3.0, 26.7.2016) provided by the European Commission in the Participant Portal69.

Data Summary

The RadioNet project will generate software products for radio astronomy data sets in the JRA RINGS (WP7). RINGS will generate reference data, both actual data and simulations, from various facility sets (LOFAR, e-Merlin, etc.). The purpose of these data sets is the validation and verification of the developed algorithms. The generated astronomical data will be in the common formats, i.e. the MeasurementSet [see http://dx.doi.org/10.1016/j.ascom.2015.06.002] and FITS images. If data sets are already available then those will be reused. Otherwise, new observations and simulations will be requested to generate a reference data set. The software will reuse the CASA and CASACORE libraries. The origin of the data are the radio astronomy facilities. The expected size of the data is several TBytes to 1 PByte. Software will be of the order of several kloc. After the lifetime of the project, the data sets will be kept available on Github and maintained by the RINGS partners for any future improvements of the algorithms. The main users of the results are radio astronomers.

FAIR data

Making data findable, including provisions for metadata

The data sets will be stored in the existing archives of the facilities.

TABLE 2.1 ARCHIVES OF THE RAW DATA OF THE RADIONET INFRASTRUCTURES

TA Name	Archive address	Access conditions
EVN	http://www.jive.nl/select-experiment	Free after 1 year
e-MERLIN	http://www.e-merlin.ac.uk/archive/	Free after 1 year
Effelsberg	http://www.mpifr-bonn.mpg.de/en/effelsberg	Free, upon request
LOFAR	http://lofar.target.rug.nl/	Free after 1 year

69 http://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tpl-oa-data-mgt-plan_en.docx

IRAM	http://www.iram-institute.org/EN/content-page-240-7158-240-0-0.html; new to be ready in 2016	Free after 1 year
TA Name	Archive address	Access conditions
APEX	http://archive.eso.org/wdb/wdb/eso/apex/form	Free after 1 year
ALMA	https://almascience.eso.org/alma-data/archive	Free after 1 year
LOFAR	http://lofar.target.rug.nl	Free after 1 year
WSRT/ALTA	https://www.astron.nl/wsrt-archive/php/QueryForm.php (ready in 2018)	Free

The metadata standards and discovery from those facilities will be used. Simulated data will be reproducible by the sets of parameters to generate them. These parameters will be documented wherever the simulations are used. Software products will be integrated in CASA/CASACORE and use the associated discovery mechanisms. For the data products, we will follow the naming conventions of the facilities. For the software products we will use the naming conventions of CASA/CASACORE. The metadata in the archives of the various facilities adequately describes all relevant parameters and keywords for searching. For software products the search keywords are not applicable. There will be no clear version numbers in case of the archives, as the metadata contains timestamps that uniquely identify the observation data. However, in case of the software – version numbers with the versioning schemes of CASA/CASACORE will be followed. The metadata created for observations is described in the Measurement Set standard [see reference above] and in the FITS standard. The metadata associated with software products are the headers in the code and the software documentation.

Making data openly accessible

The archives of all RadioNet facilities comply with the open standards policies. All data will be available during and after the project´s lifetime. Software products will become available as open source. The archives are accessible via web interfaces, most of them complying with the Virtual Observatory (VO) standards. The software products will be made accessible via the CASA/CASACORE repository in Github.

Depending on the data size, the data is directly downloadable or is accessible by interaction with the observatory staff. Where possible VO tools can be used to access images. For software products, direct download from the repositories will be available and do not require additional tooling.

The various archives have their own documentation about the software needed to access the data. It is not necessary to include the relevant software, as all tools are openly available. Data and the associated metadata will be stored in the archives of the RadioNet facilities. Code implementing calibration algorithms and the

associated documentation will be integrated with the CASA(-CORE) repositories. Both are open source and there are no restrictions on use. An appropriate arrangement with the identified repository has been explored. There is no need for a data access committee.

Making data interoperable

The data produced is stored in common formats and standards of the astronomical communities and the software products will adhere to the interoperability conventions of CASA/CASACORE, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. The Measurement Set and FITS standards will be used for data and metadata vocabularies in order to make the data interoperable. The standard vocabularies will be used for all data types present in the data set, to allow inter-disciplinary interoperability. A use of uncommon or generated project specific ontologies or vocabularies is not foreseen.

Increase data re-use (through clarifying licences)

Software products are published in the CASACORE repository under GNU General Public License v3.0. Data products are subject to the data policies and licenses of the RadioNet facilities (see Table 2.1). If new data is required, the data will generally become available 1 year after the observation has taken place (see also Table 2.1). No re-use of the data outside the radio astronomy community is currently foreseen. However, the data is openly accessible for third parties. The project will seek interaction with industrial partners to investigate the reuse of the software products in other domains. The data storage terms determined by the archive policy of the facilities, which is commonly to store data indefinitely. There are no limits foreseen to the reusability of software products delivered by RINGS. The data quality is ascertained by the quality procedures of the facilities.

Allocation of resources

There are no costs required to make the data FAIR. The JRA RINGS leader will be responsible for the data management. There is no need for plans for long-term preservation, as they will be designed by the facilities and the CASA/CASACORE collaboration partners

Data security

The data is secured according to the policies and arrangements of the RadioNet facilities, which are publicly available (See table 2.1 for the address). They assure a long-term preservations and curation of the data.

Ethical aspects

There are no ethical or legal issues that can have an impact on data sharing.

Other issues

The RadioNet JRA RINGS is using the data generated by RadioNet facilities, which follow their own procedures for the data management. However, since the RadioNet facilities follow the open policy procedure, no particular influence on the FAIR is expected.

ANNEXE 2 : FICHE DE POSTE : CHARGE DU TRAITEMENT DES DONNEES SCIENTIFIQUES – CDS DE STRASBOURG

F : Culture, communication, production et diffusion des savoirs

Ingénieur d'études

Concours N° 146

Délégation organisatrice : Ile-de-France Meudon (DR 05) (MEUDON)

Nbre de postes :	1
Emploi-type :	Chargé-e du traitement des données scientifiques
Affectation :	Observatoire astronomique de Strasbourg, STRASBOURG
Groupe de fonction :	Groupe 3

Mission :

Au sein de l'équipe du Centre de Données astronomiques de Strasbourg (CDS),
l'ingénieur-e d'études sera chargé(e) de valoriser des données scientifiques publiées dans les journaux académiques d'astronomie et de les ingérer dans les bases de données astronomiques du CDS : SIMBAD et VizieR.

Activités :

- Analyser, sélectionner et récupérer les informations pertinentes venant des articles scientifiques pour les bases de données,
- Mettre en forme les données, notamment à l'aide d'outils spécifiques du CDS,
- Synthétiser et décrire les données astronomiques,
- Utiliser et rechercher l'information dans d'autres bases de données astronomiques,
- Enrichir les données grâce aux outils spécifiques du CDS et ajouter les métadonnées conformes aux standards de la discipline,
- Vérifier et valider les résultats après ingestion,
- Collaborer à l'évolution des procédures avec les informaticiens et astronomes du CDS,
- Veiller à la qualité des bases de données,
- Valoriser l'activité de l'unité en intervenant lors de conférences scientifiques ou grand public.

Compétences :

Savoirs généraux:
- Maîtriser le travail avec des données scientifiques (dimensions physiques, unités, ...),
- Être en mesure d'apprendre des notions d'astrophysique (formation interne),
- Maîtriser l'anglais de niveau B1 (selon le cadre européen commun de référence pour les langues).

Savoir-faire:
- Maîtriser l'outil informatique (environnement Linux de préférence),
- Des notions en Python (ou autre langage de programmation) seraient un plus,
- Des notions de documentation (lecture rapide, analyse et synthèse, requêtes sur les bases de données).

Savoir-être:
- Capacité à travailler en équipe,
- Rigueur et fiabilité,
- Autonomie,
- Capacité d'adaptation et persévérance,
- Curiosité intellectuelle.

Contexte :

Le Centre de Données astronomiques de Strasbourg (CDS), hébergé par l'Observatoire
astronomique de Strasbourg, met à la disposition de la communauté scientifique internationale plusieurs services en ligne : SIMBAD, VizieR et ALADIN.
Ces différents services génèrent plus de 2 millions de requêtes par jour venant du monde entier.
Ces bases de données, développées par les informaticiens-informaticiennes et supervisées par les astronomes du CDS, sont alimentées par l'équipe de chargé(e)s de traitement des données scientifiques. Les activités du traitement des données scientifiques se déroulent dans un environnement de travail Linux et utilisent des procédures internes au CDS.

L'ingénieur-e intègrera une équipe qui comprend des astrophysiciens, des informaticiens et des chargés du traitement des données scientifiques, et interagira avec ces différents personnels. Il ou elle aura l'entière responsabilité de son travail, dont les résultats seront directement visibles dans les bases de données et dont la qualité est essentielle pour que le service continue à répondre aux attentes des utilisateurs.
La personne recrutée aura la possibilité d'accroître ses compétences pour contribuer à l'évolution des travaux dans le contexte de la science ouverte.
La personne recrutée pourra solliciter des jours de télétravail sous couvert de l'accord de son responsable.

L'Observatoire astronomique de Strasbourg est situé sur le campus de l'Esplanade ;
ce site est facilement accessible en transports en commun et dispose d'une offre de restauration collective.

TABLE DES MATIERES