Du nécessaire partage des données scientifiques

L'exemple de l'astronomie

Les ministres de la Recherche du G8, réunis avec les présidents des académies des sciences à Londres le 12 juin 2013, ont eu des mots très forts pour soutenir le partage et la réutilisation des données de la recherche[1]. Cette prise de position est un nouveau signe de l'attention croissante portée à ces données – et au fait que celles-ci doivent être « ouvertes » – par les gouvernements et les agences de financement de la recherche.

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. Les astronomes ont commencé dès 1976 à réfléchir à un format qui leur permette de réutiliser les images conservées sous forme électronique. Dans ce format d'échange, les métadonnées qui décrivent l'image et les conditions dans lesquelles celle-ci a été obtenue sont incluses en en-tête du fichier de données.

LES ENJEUX DU PARTAGE DES DONNÉES

Il y a des raisons scientifiques au cœur de la recherche en astronomie pour que les chercheurs souhaitent réutiliser les données. Pour comprendre les phénomènes physiques à l'œuvre, il faut le plus souvent faire appel à la combinaison d'observations de différents instruments. Les phénomènes observés sont aussi très souvent variables, d'où des comparaisons entre observations faites à différentes époques. De plus, la discipline utilise de grands instruments au sol et dans l'espace et la réutilisation des données permet d'optimiser le retour scientifique sur ces investissements.

Le format FITS (*Flexible Image Transport System*), dont discutaient les radioastronomes dans les années 70, est toujours l'un des outils de base de l'astronomie. C'est un bien commun de la discipline, régulièrement mis à jour et maintenu sous l'égide de l'Union astronomique internationale : le partage des données a été dès l'origine conçu pour lever les barrières entre les instruments d'observation, mais il se joue aussi des frontières. Grâce à FITS, tout astronome peut visualiser et utiliser les données de n'importe quel télescope. L'existence d'un format commun a aussi permis de développer des outils partagés, qui peuvent être utilisés par l'ensemble de la communauté pour manipuler les données.

Les astronomes ne se sont pas arrêtés au partage des données d'observation. Ils se sont aussi intéressés très tôt au partage des résultats de la recherche. Le Centre de données astronomiques de Strasbourg (CDS) a été créé dès 1972 par l'Institut national d'astronomie et de géophysique, prédécesseur de l'Institut national des sciences de l'univers du CNRS, en collaboration avec l'université Louis-Pasteur, devenue l'université de Strasbourg. Sa mission est de collecter l'information « utile » sur les données astronomiques, de les améliorer par des évaluations critiques et des comparaisons, de les distribuer à la communauté astronomique internationale et de conduire des recherches utilisant ces données. Une partie de la collecte d'information repose sur le dépouillement des articles publiés dans les journaux académiques de la discipline, d'où l'on extrait entre autres la liste des objets cités dans l'article, et des informations sur ces objets : par exemple, le nom qui leur est donné dans l'article (les objets astronomiques peuvent avoir de nombreux noms différents) ou leur position dans le ciel. Le CDS collecte aussi les grands catalogues des objets astronomiques et des images de référence du ciel et, en collaboration avec les journaux et les auteurs des articles, les tables publiées dans les journaux, ainsi que, de plus en plus, d'autres données (images, séries temporelles…) attachées aux articles.

UNE CONSTELLATION DE COMPÉTENCES ET DE PROFILS

Il est important de noter que le CDS a été créé au sein d'une structure de recherche, l'Observatoire astronomique de Strasbourg (l'actuelle UMR 7550 Université de Strasbourg/CNRS). Il inclut des scientifiques qui mènent leurs propres programmes de recherche en plus de leurs tâches au service de la communauté. Ce contact constant avec la recherche vivante est certainement l'une des clés du succès du CDS sur le long terme, puisqu'il donne un regard direct sur les besoins des chercheurs qui utilisent les services, qui sont donc correctement pris en compte, voire anticipés. Un autre aspect des com-

[1] www.gov.uk/government/news/g8-science-ministers-statement

[2] Christian Huitema, *Et Dieu créa l'Internet*, Eyrolles, 1996.

[3] *Riding the wave : How Europe can gain from the rising tide of scientific data*, European Union, 2010, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

pétences indispensables porte sur la construction du contenu : identifier et valider des données extraites des publications, construire les métadonnées permettant de décrire le contenu d'une table. C'est le travail des documentalistes du Centre de données, qui ont des compétences de pointe sur, par exemple, les objets astronomiques ou les quantités présentes dans les tables des articles. Enfin, les informaticiens de l'équipe développent les bases de données et les interfaces utilisateurs et ont une activité soutenue de veille technologique et méthodologique, ce qui permet de détecter rapidement les nouveautés susceptibles d'améliorer le service rendu aux utilisateurs.

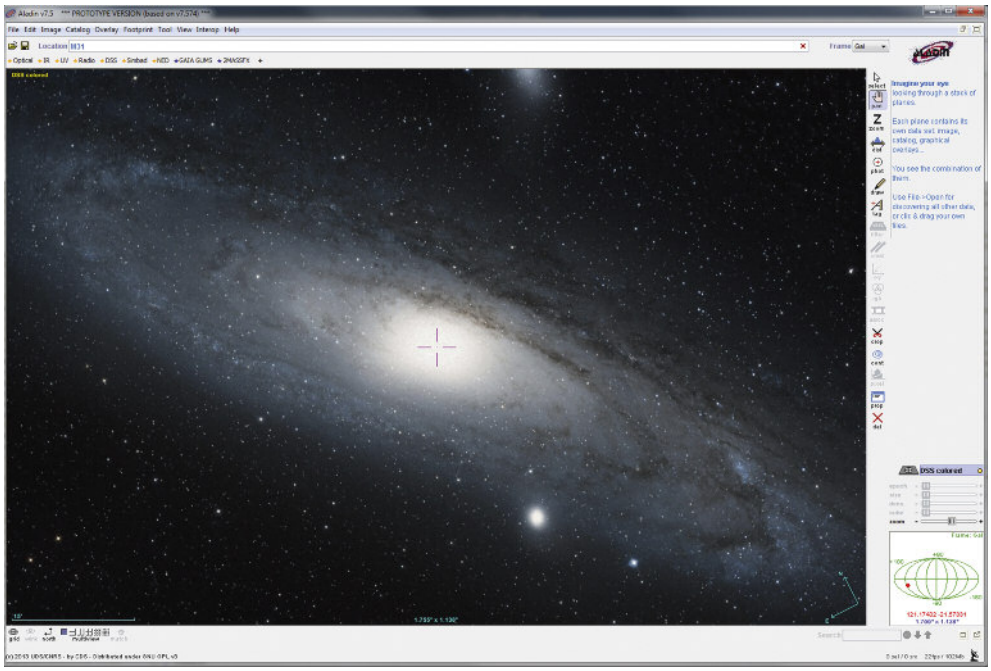

Image du Digitized Sky Survey, STScI/Nasa

***La galaxie d'Andromède*, visualisée avec le logiciel Aladin du Centre de données astronomiques de Strasbourg.**

Sur cette longue durée, le CDS a vécu plusieurs révolutions technologiques majeures : l'arrivée du *World Wide Web* vers 1993 en est évidemment une, mais on peut aussi citer le fait que la première liaison Internet entre la France et les États-Unis, en 1988, a été mise en place pour une démonstration d'une base de données du CDS aux États-Unis[2]. Pour la distribution des catalogues, on sera passé de l'envoi par la poste de cartes perforées, de disquettes ou de bandes magnétiques, sur commande reçue également par courrier postal, à la mise à disposition des fichiers par ftp, ainsi que dans une base de données pour permettre la recherche des éléments individuels de chaque table. Ce qui a été constant depuis l'origine, c'est la quête inlassable de la meilleure qualité possible pour le contenu et de la meilleure réponse possible aux besoins des utilisateurs. Les documentalistes jouent évidemment un rôle majeur sur le long terme pour assurer la qualité du contenu. L'équipe fonctionne en équipe intégrée et ses membres, avec leurs différents profils, collaborent étroitement dans les activités journalières et dans la définition de la stratégie de développement. Le succès, sur le long terme, est au rendez-vous puisque le CDS est labellisé « infrastructure de recherche » par le ministère de l'Enseignement supérieur et de la Recherche, et que ses services reçoivent plus d'un million de requêtes par jour en moyenne.

VERS UNE INFRASTRUCTURE INTERNATIONALE ET INTERDISCIPLINAIRE

Les astronomes utilisent quotidiennement les données disponibles dans les archives des observatoires, les services du CDS, la base de données bibliographique maintenue au Smithsonian Astronomical Observatory avec l'aide de la Nasa, etc. Depuis plus de 10 ans, la communauté construit aussi l'Observatoire virtuel astronomique, qui vise à donner accès de façon transparente à l'ensemble des données en ligne : l'utilisation de protocoles et de vocabulaires standardisés permet de trouver les données disponibles et d'y accéder avec des outils communs, sans devoir passer par les interfaces propres à chaque service. Et les outils sont eux-mêmes interopérables entre eux.

De nombreuses autres disciplines scientifiques suivent des chemins similaires et mettent en place des politiques de conservation et de partage des données. Les agences de financement demandent aussi de plus en plus que les résultats de la recherche soient disponibles. Le dépôt des données dans un « répositoire » institutionnel permet de répondre à cette obligation, mais il ne permet pas forcément que les données soient trouvables et réutilisables. Pour définir des standards d'échange et des métadonnées, il faut un travail au niveau des disciplines, où se trouve la connaissance des données et celle du fonctionnement de la communauté. Ces « piliers disciplinaires » serviront de base à l'« infrastructure collaborative des données » internationale et interdisciplinaire préconisée par le *High Level Expert Group on Scientific Data* mis en place par la Commission européenne en 2010[3].

Françoise Genova
Directrice du Centre de données astronomiques de Strasbourg
francoise.genova@astro.unistra.fr

POUR EN SAVOIR PLUS

Site du centre de données astronomiques de Strasbourg :
http://cdsweb.unistra.fr