LA TRIBUNE D'HUMA-NUM

Huma-Num#Actu

Huma-Num. La nouvelle très grande infrastructure de recherche pour les humanités numériques

Comme nous l'annoncions dans l'éditorial de la tribune de mars 2013, le TGE Adonis et Corpus IR ont fusionné le 1er mars 2013 pour composer la nouvelle TGIR Huma-Num. L'unité support est l'UMS 3598, placée sous la triple tutelle du CNRS (InSHS), d'Aix-Marseille Université et du Campus Condorcet. Prévue par la feuille de route de la Stratégie Nationale des Infrastructures de Recherche 2012-2020, la TGIR Huma-Num vise à faciliter la transition numérique de la recherche en sciences humaines et sociales (SHS). Son organisation originale repose sur la mise en œuvre d'un dispositif humain (concertation collective au travers de consortiums) et technologique (services numériques pérennes) à l'échelle nationale et européenne (infrastructure DARIAH) en s'appuyant sur un réseau de partenaires et d'opérateurs variés. Nous souhaitons rappeler ici le contexte de création de cette nouvelle TGIR, ses services et ses moyens.

TGIR et données de la recherche : un contexte évolutif

La première TGIR en SHS fut créée en mars 2007 dans le domaine du numérique. Prenant le nom d'Adonis (Accès unifié aux données et documents numériques des sciences humaines et sociales), elle avait pour objectif de réaliser un accès unifié aux données de la recherche. Les obstacles étaient grands reflétant, avec un effet de loupe, la multiplicité des techniques, des méthodes et des positions des acteurs de la recherche en SHS face aux outils numériques. Le projet du TGE Adonis reposait sur l'existence de données numériques structurées selon des schémas identifiables et acceptés par les communautés scientifiques productrices de données. L'interopérabilité des données constituait, de ce fait, une notion clef et relativement nouvelle en particulier chez les chercheurs. En effet, si la citation des œuvres de référence est au cœur de la démarche scientifique de publication, la notion d'interopérabilité des données de la recherche ne s'est progressivement imposée qu'avec le web 2.0. Les opérateurs des premiers corpus de données mis en ligne se préoccupaient plus de la possibilité de produire des métadonnées que de rendre celles-ci compatibles avec d'autres ensembles de données. Cette possibilité était d'ailleurs freinée par une standardisation relativement complexe et parfois inachevée. Le risque de produire des données isolées, enfermées dans des systèmes de publications propriétaires, indisponibles à l'échange, reste aujourd'hui bien réel. Le mode de création de bases de données « étanches » où seul un formulaire permet d'accéder aux données est encore fréquent. A l'heure où la mise en œuvre des projets numériques se trouve facilitée par des offres de qualité – nous pensons ici aux réalisations du CLEO/Open édition – nous souhaitons sensibiliser les chercheurs aux enjeux de l'interopérabilité. La qualité de l'éditorialisation des contenus des bases de données pèse sur la réutilisation de données dans le futur. Dans l'océan du réseau, de nouveaux territoires de données sont ainsi apparus. Il s'agit maintenant de créer les conditions de possibilité de construire des échanges entre eux. Pour passer au stade de la compétition coopérative, les données de la recherche doivent être ouvertes, documentées et contextualisées, interopérables et — autant que possible — libres d'accès dans le respect du droit. En l'état actuel, les données « moissonnables », c'est le terme consacré, sont encore trop limitées, principalement centrées sur les publications scientifiques (HALSHS, revues.org, persee.fr, cairn.info, etc.). Or, depuis une dizaine d'années, l'explosion de la numérisation, par les bibliothèques principalement, a provoqué un tournant numérique (*digital turn*[1]) qui change profondément les rapports aux données de la part des utilisateurs (chercheurs, enseignants, étudiants) ainsi que la façon d'utiliser ces dernières dans les processus de recherche. Si l'effort a été puissant sur la publication électronique et sur les archives ouvertes — pour les résultats de la recherche — les corpus de sources primaires, les archives, les fonds de bibliothèques de recherche et de laboratoire, les archives des scientifiques n'ont pas été aussi bien traitées.

En parallèle, avec l'émergence de l'*open data*, nous assistons à une explosion de l'accès à des données numériques qui peuvent — malgré des problèmes de qualité et d'intégrité — représenter des gisements d'information pour la recherche : faut-il encore pouvoir les gérer.

Ces transformations profondes des habitudes, repères et possibilités dans les façons de produire de la connaissance sont au cœur du mouvement des humanités numériques (*digital humanities*) qui se propose d'accompagner mais aussi de préparer les changements futurs pour le monde SHS tant

1. Paul Bertrand, IRHT, Ecole thématique des CRN, 2008

sur le plan technologique qu'épistémologique. Plusieurs initiatives et expériences ont eu lieu tout au long des années 2000, notamment les centres de ressources numériques lancés en 2005 par le CNRS, à l'heure où étaient publiés les premiers appels ANR Corpus (2006). Elles ont permis de mieux cerner les grands enjeux du *digital turn*. Les relations entre les acteurs scientifiques et culturels se sont largement améliorées au cours de la dernière décennie, si bien que l'on voit des projets de recherche en SHS utiliser naturellement les bases de données du monde culturel. De même, les données de la recherche en SHS, pour ce qui concerne les archives et les corpus, sont également présentes dans Gallica (BnF).

Ces transformations ont été au centre du programme de l'école thématique de Fréjus proposée par les centres de ressources numériques en 2008 et les expériences de ces dernières années ont montré l'importance de l'appropriation par les communautés des questions liées aux formats ouverts, aux méthodes d'interopérabilité, aux standards, etc. Ainsi, une prise de conscience est en cours dans les communautés scientifiques sur la nécessité d'utiliser des standards techniques dans la constitution de bases de données de corpus.

La structuration de la nouvelle TGIR

Les missions de la TGIR sont, pour l'essentiel, le fruit de la réunion des missions du TGE Adonis et de Corpus IR. Le projet est centré sur les corpus de données de la recherche et l'outillage nécessaire pour en garantir la pérennité, la visibilité et l'accès pour des réutilisations maitrisées à l'heure du *big* et de l'*open data*. Il s'agit de mettre en œuvre une politique de réutilisation des données numériques des SHS tenant compte des besoins des scientifiques avant tout dans un contexte d'accès libre aux données (*open access*).

La TGIR Huma-Num intervient sur la production de corpus de sources par l'intermédiaire de consortiums regroupant des acteurs des communautés scientifiques. Initiée dans le cadre de Corpus IR, la création de ces consortiums va être poursuivie par la nouvelle TGIR.

Elle vise également à garantir un dispositif technologique fiable et adapté permettant le traitement, la conservation et l'accès des données de la recherche en développant l'interopérabilité. Cette mission est héritée cette fois-ci du TGE Adonis. La continuité du service est ainsi assurée pour tous les projets hébergés initialement au sein de la grille Adonis, devenue grille Huma-Num. Pour les nouveaux projets, nous avons mis en place une procédure d'examen des demandes qui permet de répondre rapidement aux chercheurs tout en les informant, le cas échéant, des autres opérateurs existants (MSH, Labex, Equipex…)

La TGIR Huma-Num est également attentive à l'évolution des besoins documentaires et technologiques et des activités de recherche et développement. Elle réalise également une veille dans le domaine stratégique de l'utilisation/réutilisation des données sources dans les publications scientifiques en liaison avec les acteurs du domaine (plateformes d'éditions électroniques, archives ouvertes en particulier). Elle peut mener ponctuellement des actions d'expertise et de formation. Elle porte la participation de la France dans le projet européen DARIAH en coordonnant les contributions françaises dans ce projet.

La **concertation collective** est assurée par les relations de la TGIR Huma-Num avec les consortiums disciplinaires, les Maisons des Sciences de l'Homme, les laboratoires et les équipes de recherche.

La TGIR s'appuie en particulier sur l'activité de consortiums disciplinaires ou multidisciplinaires, qu'elle labellise et soutient. Ces consortiums ont vocation :

a) à mutualiser outils et méthodes relatifs à la constitution de corpus numériques pour la recherche en SHS (corpus de données dit « qualitatifs »). Ils fédèrent les initiatives, aident à la mise en commun des sources primaires ou secondaires et facilitent leur diffusion par une sensibilisation à l'indispensable éditorialisation et mise en contexte de leurs métadonnées.

b) à produire des guides de bonnes pratiques, en s'appuyant notamment sur les échanges qu'ils entretiennent avec les pratiques « métiers » des domaines de l'information scientifique et technique, de l'informatique et de l'édition électronique.

c) à accroître la visibilité des différents projets dont ils sont porteurs en les inscrivant dans les initiatives internationales et européennes, telles que CLARIN et DARIAH.
Il s'agit ici principalement de construire une pratique commune d'utilisation de normes adaptées (scientifiques, communautaires, techniques) dans la constitution et la réutilisation des corpus.

La TGIR Huma-Num noue également des relations privilégiées avec les maisons de sciences de l'homme et les laboratoires afin de mutualiser compétences et moyens technologiques pour la conservation, le traitement, l'interopérabilité, le stockage et l'archivage à long terme des données en SHS. La relation avec les équipes de recherche permet d'offrir un appui aux équipes porteuses de projets numériques. Il s'agit ici de conseiller et d'orienter les porteurs de projet, et de leur proposer un partenariat technologique avec la grille de services.

Le **dispositif technologique** met à disposition des équipes de recherche des services spécifiques d'aide à l'identification, au signalement, à la diffusion, à la promotion et à la préservation des données de la recherche. Il s'appuie sur l'articulation de trois composantes gérées par l'équipe de la TGIR :

▶ La *grille des services numériques*, constituée d'un ensemble de services et d'outils mutualisés pour traiter, diffuser, visualiser et stocker des données de nature hétérogène (textuelles, orales, iconographiques, audiovisuelles, cartographiques, tri-dimensionnelles…). À la différence des services d'hébergement généralistes, la grille Huma-Num offre spécifiquement des services dédiés aux SHS. Ce service marque une nouvelle étape dans le support aux données numériques de la recherche car il s'accompagne désormais d'une sensibilisation des chercheurs aux conditions de l'interopérabilité de leurs données, aux formats ouverts pour la diffusion ainsi qu'à l'importance de l'archivage des données à long terme.

Les services proposés par la grille Huma-Num évolueront en concertation avec plusieurs acteurs : le comité des utilisateurs de la grille, qui a été lancé en 2012 par Adonis, les consortiums et le conseil scientifique de la TGIR. Au delà, nous échangeons aussi avec notre partenaire historique, le centre de calcul de l'IN2P3 qui héberge, dans ses locaux du campus de la Doua à Villeurbanne, notre équipe *Infrastructure*. Fruit de l'expérience de la grille Adonis, la TGIR Huma-Num développe depuis quelques semaines d'autres services d'aide à la diffusion des données SHS : stockage de collections numériques, attribution d'identifiants pérennes, exposition OAI-PMH et RDF par exemple. En raison de la spécificité SHS de ces services, les nouveaux projets scientifiques accueillis sur la grille seront encouragés à intégrer une stratégie de mise à disposition des métadonnées/données pour Isidore ou tout autre portail utilisant les standards internationaux ouverts (OAI-PMH, RDF, etc.). Ces données pourront également s'insérer dans le dispositif d'archivage à long terme.

▶ *L'archivage à long terme des données de la recherche en SHS*, propose un archivage des données numériques des SHS mettant en œuvre la norme *open archival information system* (ISO14721:2003). Ce service, proposé en 2008 et opérationnel depuis 2010, est réalisé en partenariat avec le Centre Informatique de l'Enseignement Supérieur (CINES) et sous le contrôle du service inter-ministériel des archives de France (SIAF).

Il repose sur la mise en place d'une chaîne de responsabilités en matière d'archivage des données numériques allant jusqu'à l'archivage définitif aux Archives de France. Plusieurs communautés et opérateurs bénéficient d'ores et déjà de ce service : les consortiums de linguistique, par exemple, au travers des plateformes CoCooN et Speech and Language Data Repository (SLDR) mais aussi – et cela est moins connu – HALSHS et de MédiHAL dont les données sont archivées sur le long terme.

▶ La plateforme *Isidore*. Plus qu'un simple moteur de recherche accessible en ligne, Isidore est l'instrument numérique de référence pour le signalement, l'enrichissement, l'accès et donc la valorisation des données de la recherche en SHS en France (sources, événements, publications, communautés scientifiques et culturelles). Conçu en 2010 et inauguré le 4 avril 2011, Isidore est réalisé avec l'un de nos partenaires : le CCSD qui en assure la maitrise d'œuvre. Il est le premier dispositif d'open data sémantique du domaine. Cette plateforme évolutive offre tout à la fois un accès unifié aux savoirs des SHS et un bouquet de fonctionnalités relatives au traitement, à l'indexation, au signalement et à l'accès aux données numérisées. Isidore contient plus de 2,4 millions de références indexées et reliées entre elles à l'aide de plusieurs référentiels scientifiques. Plus de 60.000 visiteurs uniques par mois l'utilisent pour trouver de l'information afin de nourrir une recherche ou pour constituer une bibliographie thématique.

Une TGIR pour et avec les chercheurs

La fusion entre le TGE Adonis et Corpus IR a été l'occasion de reposer des questions sur l'évolution des outils et les besoins des enseignants-chercheurs. Les TGIR sont avant tout des instruments collectifs qu'il s'agit de faire évoluer avec l'implication directe des scientifiques. Certains projets ont été lancés il y a plusieurs années : 2008 pour l'archivage pérenne, 2010-2011 pour Isidore. C'est déjà un temps « lointain » dans le calendrier des technologies numériques. Cependant, si les technologies évoluent très vite, les services proposés par Huma-Num s'inscrivent dans la durée afin que les communautés scientifiques, diverses en SHS, aient le temps de les intégrer à leurs processus de recherche.
Le rôle de la TGIR est de contribuer à la prise de conscience des enjeux et à l'appropriation des méthodes et outils du numérique par les scientifiques, tout en anticipant leurs besoins futurs et en garantissant la préservation les données numériques de la recherche.

La TGIR Huma-Num ne doit cependant pas être perçue comme un « guichet » où les chercheurs viendraient sous-traiter le volet numérique de leurs programmes scientifiques. Elle est avant tout le lieu où les communautés font avancer à la fois la mise à disposition, par elles-mêmes, des corpus de données tout en accédant, à leur rythme, à des services technologiques communs et stables.

contact&info
▶ Marc Renneville,
Directeur
marc.renneville@huma-num.fr
▶ Stéphane Pouyllau,
Directeur adjoint technique
stephane.pouyllau@huma-num.fr
▶ Pour en savoir plus
www.huma-num.fr/