Appel à projets ANR « Corpus, données et outils de la recherche en sciences humaines et sociales »

Très Grand Equipement « Adonis »

Infrastructure de Recherche « Corpus »

Note technique

Pour l'appel à projet « Corpus, données et outils de la recherche en SHS », les infrastructures SHS concernées, le TGE Adonis[1] et l'IR Corpus[2], offrent un cadre de recommandations, de développements et de pérennisation pour les projets retenus.

La grille Adonis : services techniques aux unités et aux projets

Dans le cadre de ses missions, le TGE Adonis fournit, pour la production numérique issue des laboratoires SHS du monde académique français, une infrastructure de services. Depuis l'été 2010, la « grille Adonis » offre aux équipes et aux projets des solutions pour l'hébergement, la diffusion, le stockage, le calcul et l'archivage à long terme des données.
Cette grille s'insère dans l'infrastructure animée par l'Institut des Grilles[3], pionnier en Europe dans la construction d'un réseau de serveurs partagés pour des besoins de calcul ou de stockage. La grille de services est hébergée au Centre de Calcul de l'Institut National de Physique Nucléaire et Physique des Particules (CC-IN2P3). L'archivage à long terme est réalisé avec le Centre Informatique National de l'Enseignement Supérieur (CINES).

Cette grille de services numériques offre aux équipes et à leurs projets les possibilités suivantes :

- un hébergement web pour les projets de recherche. Cet hébergement propose une grande gamme de solutions basées sur des technologies ouvertes ;
- un stockage intermédiaire sécurisé afin de mettre en sécurité hors de leurs murs des données produites dans le cadre de projets de recherche ;
- un service de calcul scientifique sur des grandes masses de données utilisant la ferme de calcul du centre de calcul de l'IN2P3 (modélisations, calculs d'images 3D, statistiques) ;

[1] http://www.tge-adonis.fr/
[2] http://www.corpus-ir.fr/index.php?page=procedure
[3] http://www.idgrilles.fr/

- l'archivage pérenne des données de recherche sur le modèle international de l'Open Archival Information System (OAIS) en coopération avec le CINES ;
- prochainement, la mise à disposition d'un certain nombre d'outils pour faciliter l'hébergement, l'archivage, l'interopérabilité et le moissonnage des données (outils de conversion de formats, édition et vérification des métadonnées, etc.).

Ces services sont accessibles à tout projet de recherche respectant les critères d'éligibilité et les conditions générales d'utilisation de la grille Adonis. Cette grille est basée sur les normes et standards internationaux pour le stockage et la diffusion de données numériques. Travailler sur la grille ou y déposer des données donnent l'assurance de se conformer aux exigences nécessaires d'interopérabilité et de pérennité des données. Par ailleurs, le service est continu : il ne s'arrête pas à la fin des projets de recherche. La grille peut être utilisée par les projets ne disposant pas de structures avec service informatique comme par les projets disposant de cette infrastructure ; la grille peut alors offrir du stockage « hors murs », du calcul et de l'archivage.

Guide des bonnes pratiques

Une mission essentielle de l'IR Corpus et du TGE Adonis est d'accompagner les laboratoires dans leurs réalisations de projets numériques, d'encourager et de promouvoir la standardisation des modes de production et de diffusion des données numériques. Quatre guides sont disponibles sur le site du TGE Adonis et de l'IR Corpus : guide généraliste des bonnes pratiques numériques, guides spécifiques sur le choix de formats numériques pérennes dans un contexte de données orales et visuelles, sur l'OAI-PMH et les techniques de moissonnage de données. D'autres sont en cours de réalisation au sein du réseau des centres de ressources numériques structuré par le TGE. Les équipes sont invitées à se mettre en relation avec l'IR Corpus (http://www.corpus-ir.fr) et ses consortiums afin de bénéficier de conseils et d'un suivi dans le domaine des formats et des standards numériques à utiliser et de leurs développements.

Ces guides sont conçus comme des *vade mecum* présentant les formats, les standards et les pratiques avec les critères d'applications et cas de figures possibles, pour numériser, sauvegarder ou exploiter des données sur multiples supports (texte, image, vidéo, son) ou adapter les corpus numériques à ceux d'autres initiatives, comme celles qui moissonnent via OAI-PMH ou RDF, ou comme la plate forme de recherche ISIDORE. L'adoption de ces bonnes pratiques favorisera l'interopérabilité et la pérennité des données et préparera aux évolutions prévisibles.

Les Infrastructures de Recherche Adonis et Corpus assurent une veille sur les besoins et usages émergents et les accompagnent dans leurs évolutions ; le TGE Adonis adapte sa grille et ses guides en fonction des évolutions technologiques sans modifier la vie des projets accueillis.

Recommandations pour la pérennité et l'interopérabilité des données produites dans le cadre de l'appel à projets « Corpus, données et outils de la recherche en SHS »

Pour une bonne insertion d'un projet dans les dispositifs nationaux et internationaux de mise à disposition, d'interopérabilité et de pérennisation de contenus numériques, nous recommandons les différents points suivants (se conformer à ces dispositifs ne signifie pas rendre les données accessibles à un large public ; des verrous et protections des données sont possibles, même en suivant des pratiques standardisées) :

Corpus

- Avoir une réflexion sur les modalités scientifiques mais aussi technologiques de construction des corpus et de traitement des données (justification des formats ou des logiciels).
- Réfléchir dès l'origine du projet à l'évolution potentielle des formats et standards et adopter autant que possible ceux qui ne sont pas propriétaires mais ouverts ou « libres ».

Outils

- Prévoir des outils génériques non propriétaires dont la capacité est de traiter des formats généralement connus et acceptés par la communauté.
- Préférer autant que possible des applications gestionnaires de données qui soient libres et qui soient le produit d'une communauté de chercheurs et de techniciens qui font évoluer ces outils.
- Prévoir en amont du projet les besoins en outils ou méthodes d'exploitation des données recueillies ; définir les besoins d'interrogation et de navigation ; établir un état de l'art du domaine concerné et faire des choix en fonction de l'avis des partenaires impliqués.

Structuration des données

Dans tous les cas, les données numériques doivent être structurées. Plusieurs conseils sont à suivre au sujet de la description, de l'organisation, du classement, de l'exploitation et du moissonnage des données :

- Choisir un ou plusieurs formats de métadonnées et mettre en place des méthodes et des ressources pour les renseigner. Les méta-données (données sur les données) doivent au minimum comprendre le DC de base (Dublin Core), qui est largement utilisé par les communautés scientifiques.
- Dans le cas de projets basés sur des données hétérogènes, il est nécessaire de de choisir un socle commun de méta-données, ou alors de diversifier les réservoirs de données (bases de données, répertoire de fichiers, etc.) et de définir les modalités logiques, scientifiques et techniques qui permettent de mettre en relation ces réservoirs.

À chaque choix, se positionner face aux standards et normes en vigueur (par exemple, pour les informations géographiques, avec Inspire) et expliquer par quels moyens humains et techniques ils seront mis en œuvre.

Recommandations pour les métadonnées et la structuration

- Archivage et entrepôt de données : utilisation des formats OAIS, EAD, OAI-PMH (qui se fonde sur le Dublin Core)
- Encodage des données : utilisation du Dublin Core, des schémas XML dont METS
- Web 3.0 et interopérabilité : utilisation des formats RDF, RDFa. Ces deux derniers sont particulièrement utiles pour l'interopérabilité des données complexes ou hybrides
- De manière générale, le Dublin Core accompagné d'une interopérabilité via OAI-PMH sont le minimum souhaitable. Pour des données plus complexes, d'autres formats et standards peuvent être adaptés dès qu'ils sont décrits, publics et appliqués dans le domaine en question.

Recommandations pour les ressources textuelles

En France, l'utilisation de la TEI pour l'encodage des ressources textuelles est en constante augmentation. Dans ce type de format, deux recommandations :

a) utiliser une norme d'encodage des caractères (en particulier l'Unicode UTF-8) ;
b) utiliser une norme de structuration du contenu (en particulier la TEI : adapter les Guidelines for Electronic Text Encoding and Interchange au corpus et documenter les choix).

Recommandations pour les ressources visuelles

Un projet utilisant des ressources visuelles statiques ou animées, explicitera ses choix quant au niveau de qualité et de densité des informations (résolution), ainsi que le type de schéma de métadonnées choisies (pour les images EXIF, IPTC…).

Dans ce domaine, trois recommandations :

a) différencier un format de conservation et un format de diffusion : pour les images par exemple, les formats TIFF pour la conservation et JPEG ou PNG pour la diffusion (en qualité maximale), la numérisation devra être à 300 dpi minimum. Pour la vidéo, l'utilisation de formats ouverts comme VRML ou X3D est recommandée ;
b) documenter les métadonnées à travers un standard du domaine comme pour les images EXIF ou IPTC ;
c) créer une structuration de ces métadonnées dans des fichiers indépendants mais se conformant au Dublin Core *a minima*, ceci afin de faciliter leur pérennité et leur exploitation à travers les plates-formes de recherche.

Recommandations pour les ressources sonores

Pour les données sonores, les recommandations sont les mêmes que pour les données visuelles :

a) différencier un format de conservation et un format de diffusion : utiliser de préférence des formats « normalisés » comme pour la conservation WAV ou BWF et pour la diffusion comme MP3 ou Vorbis ;
b) documenter les métadonnées à travers un standard du domaine comme pour les fichiers sonores à usage linguistique, les modèles proposés par OLAC ;
c) créer une structuration de ces métadonnées dans des fichiers indépendants mais se conformant au Dublin Core *a minima* afin de faciliter leur pérennité et leur exploitation à travers les plates-formes de recherche.

Recommandations générales pour la diffusion et l'interopérabilité des données

Dans le respect des règles juridiques nationales et internationales en vigueur, les données, aussi bien sous formes de corpus de fichiers que dans une base de données devront :

- être accessibles selon le standard du domaine ;
- dans le cadre d'une diffusion ouverte, être interopérables selon les langages en vigueur :
    - pour les métadonnées, avoir au minimum OAI-PMH comme format d'exposition ;
    - pour les données brutes, être dans des formats ouverts.

Dispositifs de consultations des données

- Avoir une stratégie de diffusion et de valorisation des données produites à travers l'utilisation de différents outils et plates-formes et non d'une seule
- Avoir une présence internet basée sur des technologies ouvertes
- Mettre en place une stratégie pour une continuité de services après la fin du projet

Pérennité

- Choisir une nomenclature des fichiers numériques et un plan de nommage, compris et acceptés par l'ensemble des participants au projet.
- En cas de diffusion web, choisir un identifiant unique et pérenne, en fonction de la volumétrie et de la durée du projet.
- Définir en amont du projet
    - une solution de stockage sécurisé des données, « hors murs »
    - un système d'archivage à long terme