Proposition de formulaire pour l'élaboration d'un plan de gestion de données au CC-IN2P3

1. Informations administratives

Nom et identifiant du projet
Description concise du projet
Agences de financement et identifiant éventuel (facultatif)
Nom et identifiant éventuel (exemple : ORCID, ISNI…) du responsable principal de projet
Date de la 1ère version du DMP
Date de la dernière mise à jour du DMP

2. Description des données

Décrire de façon concise pour chaque jeu de données collectées ou générées :

- nature (exemples: données issues d'expériences de physique, d'observatoires astronomiques, d'imageurs médicaux, de simulations numériques,…),
- discipline,
- volumétrie,
- méthode de collecte et/ou création,
- gestion du flux de données,

Spécifier le ou les formats des fichiers qui seront employés pour le versement, la distribution et éventuellement la préservation des données produites ainsi que de leurs produits dérivés.

Pourquoi est-ce important ? Dans l'éventualité de la préservation et de l'utilisation à long terme des données, le format employé doit être indépendant de la plateforme employée. L'utilisation de format de données non propriétaire permet d'améliorer la capacité à réutiliser les données à l'avenir (attention : il ne s'agit pas d'une garantie à 100%).

Indiquer comment les données seront organisées durant le projet : conventions de nommage des répertoires et des fichiers, contrôle des versions etc…

Pourquoi est-ce important ? Il est important d'établir dès le démarrage du projet une convention de nommage des données notamment dans le cadre de projets multipartenaires. En effet, il peut être parfois délicat et lourd de devoir revoir cette organisation en cours de projet, d'où l'intérêt de réfléchir longuement à celle-ci au tout début.

Indiquer si vous réutilisez des jeux de données préexistants (oui/non) ? Si oui, préciser comment elles peuvent être réutilisées (licences, convention d'échanges de données…)

Indiquer s'il existe des données au CCIN2P3 portant sur le même sujet de recherche (oui/non)

3. Métadonnées et documentation

Les métadonnées sont des informations structurées servant à décrire des données. Ce sont littéralement des données à propos de données.

Quels types de métadonnées produirez-vous afin de décrire vos données ?

Est-ce qu'un standard de métadonnées sera utilisé (exemple : Dublin Core, ISO 19115, DataCite Metadata Schema…)? On peut se référer aux listes de standards de métadonnées fournies par le Digital Curation Centre http://www.dcc.ac.uk/resources/metadata-standards ou par la Research Data Alliance http://rd-alliance.github.io/metadata-directory/.

Comment seront gérées ces métadonnées ?

Comment sera géré le catalogue et l'indexation des données ?
Est-ce que la gestion de ces métadonnées sera à la charge du CCIN2P3 (par la mise en place d'outils de gestion de métadonnées comme des bases de données relationnelles par exemple) ?

Une partie "documentation" sous forme de fichiers README peut être parfois associée aux métadonnées pour comprendre les données :

- informations sur le projet : hypothèses, méthodologie, échantillonnage, instruments ;
- informations sur les fichiers ou bases de données et sur les paramètres : unités de mesure, vocabulaire, abréviations.

Pourquoi est-ce important ?
Les métadonnées sont essentielles pour assurer la compréhension des données car elles peuvent être le seul lien entre le producteur des données et les personnes qui les réutilisent.
Elles peuvent se trouver sous forme de bases de données relationnelles, sous forme étiquetées (XML, JSON etc…). Comme pour les données, le choix des métadonnées est important plus particulièrement dans le cadre de préservation à long terme des données.

4. Responsabilité

Nom de la personne responsable de la mise en œuvre et mise à jour du plan de gestion de données

Nom de la personne responsable de chaque activité de gestion des données (facultatif) :

- création/ collection de données,
- documentation et métadonnées,
- stockage et sécurité,

- préservation à long terme

Nom de la/des personne(s) responsable(s) de ces données (« responsable stockage »)

Pourquoi est-ce important ? La personne désignée responsable des données (« responsable stockage ») le sera durant toute la durée du projet. S'il quitte le projet, une nouvelle personne responsable devra être désignée. Cette personne sera le référent pour les données que son projet stockera au CCIN2P3 : que ce soit en terme de droits d'accès, de transfert de propriété, en cas de départ de collaborateurs, de demande d'espace de stockage, d'établissement et de révision de la politique de gestion des données.

Si par accident, il n'y en a plus alors le DAS (Directeur Adjoint Scientifique) de la thématique et/ou le DAS en charge du calcul devient le responsable du devenir de ces données : il devra désigner un nouveau czar stockage.

Pour les projets hors IN2P3 et les services hébergés, le directeur de laboratoire ou d'unité devient responsable du destin des données s'il n'y a plus de czar de stockage : il devra désigner un nouveau référent stockage.

Si tous les recours précédents restent infructueux, on conserve les données sur le support habituel pour une durée de 1 an. Au bout de ce délai, les données seront migrées sur un autre support pendant 1 an.

5. Ethique, cadre légal

5.1. Ethique et confidentialité des données

Est-ce que votre projet de recherche implique des sujets humains ? : o/n.

Si oui, donner les détails de l'accord de conservation et de partage des données (consentement éclairé).
Préciser les étapes pour la protection de l'identité des participants (avis du CCTIRS, autorisation de la CNIL).
Donner les étapes pour assurer la sécurité du stockage et du transfert de ces données, si besoin.

Dans le cas de projets de recherche impliquant des sujets humains, les données doivent avoir été anonymisées avant d'arriver au CC-IN2P3.

5.2. Droits de propriété intellectuelle

Nom de la personne et/ou de l'institution/organisme qui détiendront les droits de propriété intellectuelle pour chaque jeu de données (habituellement spécifié dans l'accord de consortium).

Préciser la (les) licence(s) pour la réutilisation de chaque jeu de données (Conseil : prévoir une liste déroulante : Licences Creative Commons, Open Data Commons, License ouverte (Etalab), Tous droits réservés, autres…)

Pourquoi est-ce important ? Afin de partager les données au sein ou hors du projet, les centres de dépôt ont besoin d'une déclaration claire de la part du producteur de données afin de savoir qui en est le propriétaire. Le responsable du projet ou l'organisme qui l'emploie et finance ses travaux sont habituellement considérés comme le détenteur des droits de propriété intellectuelle sur ces données. Le CCIN2P3 ne demande pas le transfert de ces droits mais demande plutôt la permission de gérer la diffusion des données (au sein et hors du groupe participant au projet) et éventuellement archiver ces données. Dans le cas où des copyrights ou brevets sont utilisés pour produire les données, les producteurs de celles-ci devront entamer une discussion préalable avec le CCIN2P3 afin de déterminer ce que celui-ci est autorisé à faire avec ces données.

Est-ce que ces droits seront transférés à une autre organisation ou institution pour la diffusion et l'archivage ?

Est-ce que des instruments, logiciels, procédés soumis à des copyrights seront utilisés afin de produire ces données ? Si tel est le cas, comment votre projet a obtenu la permission d'utiliser ces instruments, logiciels, procédés et la permission de disséminer les données produites ?

6. Stockage sécurisé et sauvegarde

Cette section concerne le stockage sécurisé à court terme et la sauvegarde des données pendant le déroulement du projet.

6.1. Sécurité physique

Indiquer où vos données seront stockées pendant le projet.

Indiquer le plan de sauvegarde de vos données : nombre de copies, copies hors site, supports, automatisation de la sauvegarde.
Nom de la personne responsable de la sauvegarde.

Pour ce qui concerne le CCIN2P3, combien de copies souhaitez-vous conserver sur ce site ?

Pourquoi est-ce important ? Les données sont fragiles et peuvent être effacées accidentellement ou corrompues suite à un problème matériel ou logiciel. Afin de les protéger au mieux, il est recommandé de stocker plusieurs copies à plusieurs endroits différents.

6.2. Sécurité informatique et accès aux données

Protection contre les virus et les intrusions.
Restrictions sur le droit d'accès (authentification nécessaire pour y accéder) (oui/non).
Encryptage des données.
Clause de confidentialité ?
Dispositif éventuel pour le transfert sécurisé et intègre des données à une tierce partie.

Pourquoi est-ce important ?
La sécurité des données numériques est importante au cours de leur cycle de vie. Ces données peuvent inclure des identifiants permettant directement ou indirectement de remonter à l'identité d'individus dans le cas de travaux incluant des êtres humains. Ces données peuvent aussi tomber sous le coup d'un brevet et/ou avoir une valeur commerciale. Un environnement de travail et de stockage sécurisé peut inclure des restrictions sur le droit d'accès (authentification nécessaire pour y accéder), l'encryptage des données, une sauvegarde, une protection contre les virus et les intrusions. Dans le cas d'études impliquant des sujets humains, la protection des données et leur anonymisation font partie des règles fondamentales à respecter. Les données arrivant au CC-IN2P3 doivent avoir été anonymisées au préalable.

7. Partage des données

Cette section concerne le partage des données.

Pourquoi est-ce important ?
Il est fortement recommandé de déterminer dès le début du projet, la manière dont le partage et la dissémination des données seront effectués. Cela peut avoir des conséquences sur les choix technologiques et l'organisation même du traitement des données particulièrement s'il est distribué dans de multiples endroits.

Indiquer si vous souhaitez partager vos données o/n.
Si oui, indiquez la façon dont vous souhaitez le faire :

- Prise en charge intégrale ou partielle par le CCIN2P3 en utilisant les services existants proposés par celui-ci.
- Dissémination prise en charge par votre projet à travers des interfaces tel qu'un site web par exemple. Si cette solution est choisie, il est recommandé que le ou les producteurs de données fassent le nécessaire afin d'effectuer l'archivage éventuel de celles-ci après que la période de dissémination des données se termine.
- Dépôt de données dans un entrepôt disciplinaire (ex : Observatoire Virtuel).
- Préservation des données avec une dissémination différée (période d'embargo). Dans ce cas de figure, le producteur des données passe un accord avec un dépôt publique de données pour l'archivage des données avec une dissémination qui démarre à une date ultérieure.
- Utilisation d'entrepôts institutionnels gérés par un organisme public.

Si oui, l'accès sera ouvert à tous ou restreint : avec qui et conditions de restriction du partage des données (données sensibles/personnelles, données soumises à des droits de propriété industrielle (brevet, valeur commerciale), données soumises au secret défense, secret statistique, autres...)

Si oui, dans quel délai ?

- Accès immédiat
- Embargo (préciser la durée)

Décrivez le type d'utilisateurs qui utiliseront ces données.

Pourquoi est-ce important ? Caractériser la communauté d'utilisateurs peut influencer la façon dont ces données seront gérées et partagées.

Procédure éventuelle d'obtention d'un identifiant pérenne pour les données (DOI, Handle…)

8. Préservation à long terme (archivage)

Envisagez-vous d'effectuer la préservation à long terme de vos données c'est à dire au-delà de l'arrêt de votre projet ? oui/non

Si oui, indiquer quelles données seront sélectionnées pour l'archivage (attention certaines données sont soumises à des obligations contractuelles, légales ou réglementaires).
Que faire des données stockées pendant le projet mais qui ne seront pas archivées ? Faut-il envisager leur effacement/destruction ? Faut-il envisager un délai de grâce avant destruction définitive de celles-ci ?

Quelle sera la durée de préservation des données au-delà du projet ?

Pour les données archivées, quels sont vos projets pour une éventuelle transformation de leur format dans le futur ?

Pourquoi est-ce important ? Toutes les données ne sont pas destinées à être conservées à perpétuité. C'est pour cela qu'il est important de déterminer la durée de rétention des données, particulièrement si ces données ne sont pas conservées de façon permanente. Pour les données à durée de vie limitée, une politique claire de gestion des données concernant leur effacement permet d'utiliser plus efficacement l'espace de stockage disponible et permet de réduire le volume de métadonnées associées. Cette réduction permet aussi de réduire le temps nécessaire à localiser les données d'intérêt.

Si les données doivent être remises en libre accès durant la période d'archivage, sous quelle forme doivent-elles être accessibles (ex : archive tar, système de fichiers avec une organisation de l'arborescence identique à celle utilisée originellement) ?

Plusieurs classes de stockage sont envisagées en ce qui concerne l'archivage au niveau du CC-IN2P3:

- Les données restent en ligne mais sont figées : elles restent accessibles en lecture seule dans le système de stockage d'origine, tout en conservant les droits d'accès en lecture pour tous les ayants droit et les informations sur le ou les propriétaires des fichiers.
- Les données restent en ligne mais sont figées : elles restent accessibles en lecture seule dans une zone de stockage confinée, tout en ne conservant les droits d'accès en lecture que pour le « responsable stockage » et les informations sur le ou les propriétaires des fichiers.
- Les données sont stockées sous forme d'archive tar dans notre système de stockage de masse et sont uniquement accessibles par le « responsable stockage » en lecture seule. Un seul exemplaire est conservé : ce cas couvre la situation où ces données sont aussi conservées dans un autre centre de données.
- Les données sont stockées sous forme d'archive tar dans notre système de stockage de masse et sont uniquement accessibles par le « responsable stockage » en lecture seule. Une double copie de ces données est effectuée au CCIN2P3 sur deux media différents: ce cas couvre la situation où ces données sont précieuses et pour lesquelles il n'existe pas d'autre copie sur un autre site.

Dans tous les cas de figure, un délai de rétention est défini pour les données archivées qui sera de 5 ans par défaut. Il peut être redéfini au cours de la durée de vie du projet (révision du plan de gestion de données) suite à un accord entre le «responsable stockage » et le CCIN2P3.

Pourquoi est-ce important ? Les données numériques ont besoin d'être gérées activement afin d'être sûr qu'elles seront toujours disponibles et utilisables.

GLOSSAIRE

Archivage : Opération qui consiste à dupliquer des données qui ne sont plus modifiées, qui ont une grande valeur pour le projet scientifique et qui, par conséquent, doivent être conservées. L'archive est un système de stockage séparé de l'ensemble « stockage primaire » + « sauvegarde ».

Dissémination : Diffusion des données au sein ou hors du projet au cours de sa durée de vie ou après la fin de celui-ci. La diffusion des données doit aussi prendre en compte le rôle de chaque utilisateur (hiérarchie entre administrateurs, curateurs, simples utilisateurs etc…) et définir alors les privilèges d'accès aux données qui peuvent différer suivant le rôle de chacun (droits d'accès en lecture, écriture, modification, effacement…).

Donnée personnelle : Toute information relative à une personne physique identifiée ou qui peut être identifiée, directement ou indirectement, par référence à un numéro d'identification ou à un ou plusieurs éléments qui lui sont propres (art. 2 "Loi Informatiques et Libertés"). Source : http://www.cil.cnrs.fr/CIL/spip.php?rubrique299

Données sensibles : Les données sensibles sont celles qui font apparaître, directement ou indirectement, les origines raciales ou ethniques, les opinions politiques, philosophiques ou religieuses ou l'appartenance syndicale des personnes, ou sont relatives à la santé ou à la vie sexuelle de celles-ci. Source : http://www.cil.cnrs.fr/CIL/spip.php?rubrique300

Durée de rétention : Durée pendant laquelle les données stockées seront conservées avant destruction.

Métadonnées : Informations servant à définir ou décrire des données. Ce sont littéralement des données à propos de données.

Sauvegarde : Opération périodique qui consiste à dupliquer et à mettre en sécurité les données contenues dans le ou les système(s) primaires de stockage (ceux directement utilisés pour accéder et traiter les données). Elle permet de restaurer ces données en cas de perte, effacement ou corruption sur ce ou ces système(s) de stockage.

Versement : Dépôt des données dans le ou les services de stockage qui seront utilisés pour l'accès et la dissémination, à destination des utilisateurs, des données produites par le projet.