MI
Mission Interdisciplinarité

DEFI MASTODONS

Troisième appel à projets

La qualité des données dans les Big Data

Différents programmes de recherche (ANR, PIA, H2020) ont été consacrés aux masses de données. Depuis 2012, le CNRS a également contribué à ces avancées en lançant le **défi Mastodons** qui vise à favoriser l'émergence d'une communauté scientifique interdisciplinaire autour des **Big Data** et de la **Science des Données,** capable de rivaliser avec les grands laboratoires internationaux et de produire des solutions originales sur le périmètre des **données scientifiques**.

De nombreux domaines *scientifiques* (ex : le séquençage haut débit, observatoire virtuel en astronomie, simulation en physique et énergie, imagerie médicale, données environnementales biotiques comme abiotiques), *économiques* (ex : e-commerce, systèmes décisionnels), ou *sociaux* (ex : réseaux sociaux, bibliothèques numériques, patrimoines culturels) produisent et consomment des volumes de données considérables. L'ouverture des données (*Open Data*) et la corrélation entre sources de données (*Linked Data*) sont devenues des instruments de valorisation des données et posent à ce titre de nombreux problèmes d'hétérogénéité, de sémantique et de droits d'usage. Le CNRS, à travers ses unités de recherche et ses grands instruments, concentre plusieurs centaines de bases de données et de corpus d'informations dont les volumes croissent de façon exponentielle et dont la valorisation se révèle un enjeu stratégique. Cette valorisation ne peut être effective que si les données et les connaissances qui en dérivent sont caractérisées qualitativement, quoi qu'il en soit du caractère subjectif ou contextuel de la notion de qualité.

Qu'elles proviennent d'observations, de calculs ou de numérisations, de simulation ou modélisation les données peuvent souffrir de multiples problèmes dus à leur hétérogénéité, leur sémantique ou leur transformation. Les erreurs ou les imperfections (biais expérimentaux) des données peuvent être d'origine *technologique* (survenant au niveau du capteur, de l'instrument de production, de la mémoire de stockage, du transfert sur le réseau…), *d'origine humaine* (générées lors des saisies, des annotations ou de l'interprétation des données…) ou, *d'origine logicielle* (erreurs de calcul, mauvaises transformations de formats, données tronquées, ...). Les facteurs de qualité peuvent être de nature diverse : données inconnues ou incomplètes, données incohérentes, données incertaines, données obsolètes, données peu ou non crédibles, données ambiguës, types ou formats incompatibles, échelles ou unités de mesure incompatibles, conflits de nommage, conflits structurels... Ces problèmes de qualité sont exacerbés par l'explosion des volumes de données et de métadonnées et par la mise en œuvre de systèmes critiques où la décision est susceptible d'impacter des vies humaines ou des structures économiques et sociales. Ils concernent toute la chaîne de traitements de données, allant de la production ou l'acquisition multi-sources, à leur intégration, leur agrégation et leur transformation en connaissances servant de base à la décision.

La recherche sur la qualité des données est aux confluent de plusieurs disciplines (la biologie, la physique, les statistiques, l'informatique, le traitement du signal… et ne peut pas être envisagée de façon indépendante des processus de transformation des données et de leur usage. Les algorithmes de calcul, de recherche d'information, d'extraction de connaissances ou de transport des données doivent intégrer en leur sein des mécanismes de redressement, de transformation et de filtrage des données, et tenir compte du caractère incertain, incomplet ou incohérent de celles-ci. D'autres approches, plus globales, peuvent être envisagées. Elles peuvent viser le **diagnostic** (mesure, analyse statistique, profilage, test de cohérence, détection d'anomalies diverses...), la **prévention** (échantillonnage, règles de filtrage, modèle de cohérence logique...) ou la **correction** (nettoyage, transformation, complétion, rétro-action...). A ces approches, il faut la nécessaire phase de sécurité des données et de protection de la vie privée, en mettant en place des protocoles reconnus d'accès aux données et des techniques d'anonymisation de ces données. Toutes ces approches doivent tenir compte du caractère multicritères du problème et proposer des techniques dont les coûts sont en adéquation avec les enjeux applicatifs visés. Ces approchent s'inscrivent généralement dans une méthodologie globale de gestion de la qualité, propre à chaque domaine d'application.

Le **défi Mastodons** (http://www.cnrs.fr/mi/spip.php?article53) existe depuis 2012 et a progressivement constitué une communauté scientifique interdisciplinaire autour des **Big Data** et de la **Science des Données**. Deux appels à projets, en 2012 et 2013, ont soutenu 26 actions ainsi que l'émergence d'une communauté interdisciplinaire, structurée au sein d'un nouveau Groupement de recherche crée en 2015, le GDR MaDICS

(http://www.madics.fr). Ce **troisième appel à projets du défi Mastodons** a pour objectif de compléter les précédents en suscitant des actions de recherche sur **la qualité des données et des connaissances** tant au niveau de leurs sources de production qu'au niveau de leurs processus de transformation et d'exploitation.

Dans cette perspective, le CNRS souhaite soutenir une troisième vague **de projets visionnaires, capitalisant des connaissances de plusieurs disciplines, et permettant de franchir un pas significatif dans la valorisation et la sécurité des grandes masses de données**. Ces recherches peuvent concerner, mais de façon non exclusive, les axes de recherche suivants liés à la qualité :

- Identification et formalisation de facteurs de qualité et des métriques associées ;
- Analyse statistique et profilage des grandes masses de données, détection d'anomalies ;
- Modèles formels de qualité (estimation, prédiction, cohérence, optimisation...) ;
- Nettoyage et transformation de données hétérogènes, outils et workflows ;
- Intégration de données hétérogènes, de différentes natures et à différentes résolutions ;
- Benchmarking et procédures de mesure et de test ;
- Analyse d'impact sur les usages et la décision ;
- Perception de la qualité selon les domaines et les usages ;
- Résistance des données à des modèles approximatifs ou multi-échelles ;
- Evaluation de la perte d'information, techniques de compensation, fidélité de la représentation ;
- Gains et risques sociétaux de la combinaison de données détaillées sur les comportements humains, géolocalisées et résolues temporellement…

Les consortiums intéressés peuvent déposer un projet scientifique (**3 à 5 pages maximum**) comportant les éléments suivants :

- Vision scientifique de l'équipe/consortium sur les thèmes du défi ;
- Les verrous scientifiques et les axes de recherche à moyen terme, avec un focus particulier sur la première année ;
- Les acquis scientifiques dans le domaine ou dans un domaine connexe susceptible de contribuer aux problèmes scientifiques ou sociétaux posés (publications significatives, projets passés ou en cours, applications réalisées, logiciels, brevets…) ;
- Les différentes disciplines impliquées et leurs contributions respectives au projet ;
- Une liste de 3 à 5 chercheurs seniors impliqués de façon significative dans le projet.
- L'originalité du projet par rapport aux recherches actuellement en cours dans les équipes impliquées.

Le montant du financement des projets retenus sera évalué par le comité d'évaluation du défi selon la nature du projet, son ambition, les résultats visés et le nombre de partenaires. Le soutien financier concernera uniquement le fonctionnement de projets. Les frais de personnels (CDD, salaires doctorants, post doctorants…) ne sont pas éligibles. A titre exceptionnel, quelques gratifications de stages pourront être autorisées sous réserve mais leur nombre est limité pour l'ensemble du défi. La gratification de stage explicitement justifiée, pourra être sollicitée et allouée uniquement aux seules structures CNRS (UMR, UPR…). La convention de stage sera établie par la Délégation régionale du CNRS sur les crédits correspondants notifiés. La totalité du budget alloué à chaque projet sera versée à l'unité de rattachement du ou de la porteur-e du projet, qui sera en charge de son utilisation selon les besoins du projet. Les reliquats de budget non utilisés en 2016 ne seront pas reportés en 2017. Bien que les projets soumis soient pluriannuels, leur financement se fait de façon annuelle en tenant compte des résultats scientifiques, de la synergie créée dans le projet et de l'évolution du budget de la Mission pour l'interdisciplinarité du CNRS.

L'appel à projets est ouvert aux chercheur(e)s et enseignant(e)s-chercheurs relevant ou non d'une unité CNRS, mais le porteur du projet doit appartenir à une unité CNRS (UPR, UMR, USR…). Chaque porteur-e de projet s'engage à rédiger un rapport d'activité annuel et à présenter ses résultats dans le colloque de restitution annuel du défi.

==========

Le formulaire de candidature est disponible à l'URL : http://www.cnrs.fr/mi/spip.php?article819

Il doit être déposé par le/la porteur-e du projet sur l'application SIGAP *(Obtenir de l'aide sur l'application SIGAP)*

Date limite de dépôt des projets : le 26janvier 2016 à minuit.

Pour plus d'informations : mi.contact@cnrs.fr