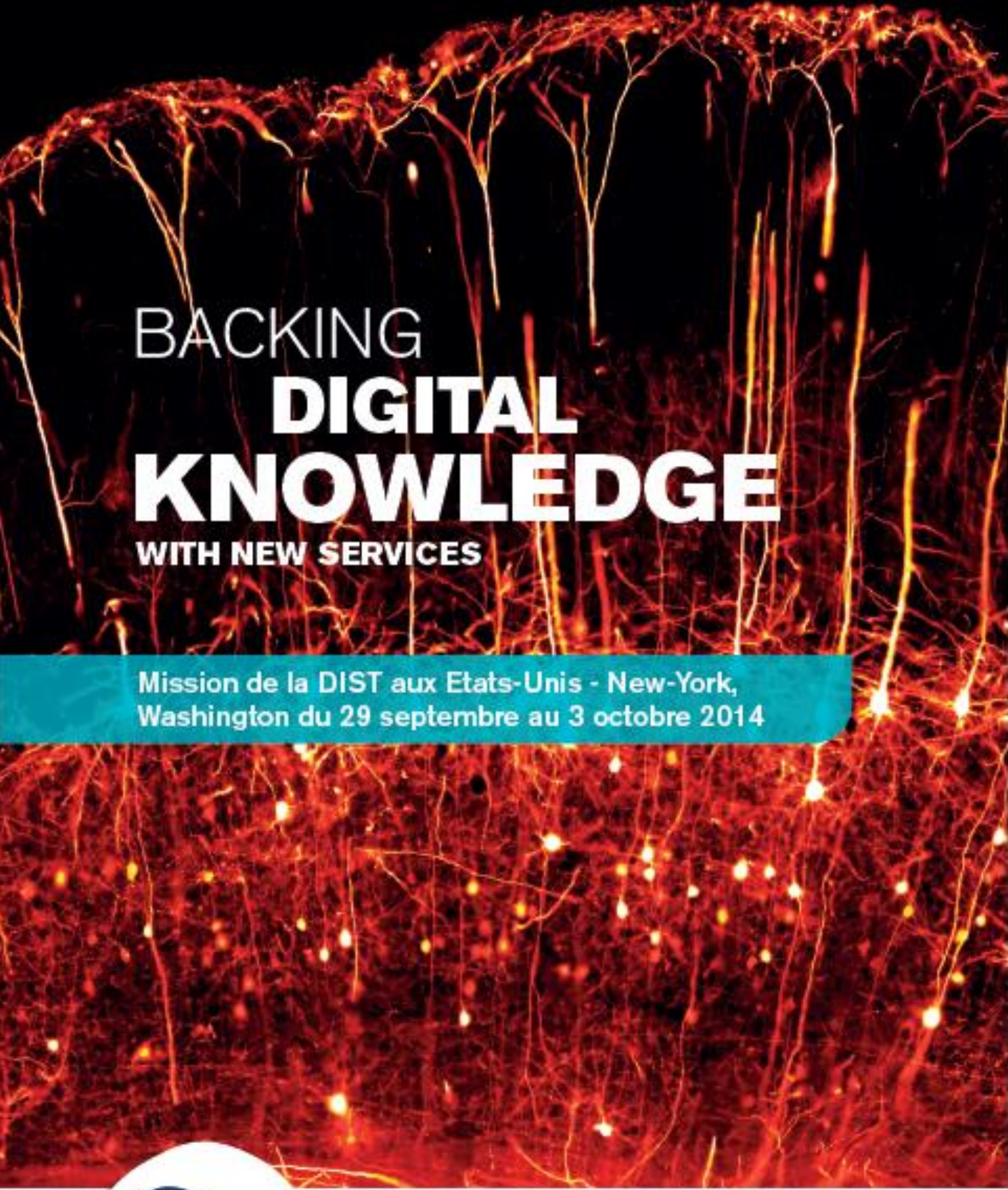


BACKING DIGITAL KNOWLEDGE

WITH NEW SERVICES

Mission de la DIST aux Etats-Unis - New-York, Washington du 29 septembre au 3 octobre 2014

www.cnrs.fr

Direction de l'information scientifique et technique

“Backing Digital Knowledge with New Services”

Mission de la DIST aux Etats-Unis

New-York, Washington du 29 septembre au 3 octobre 2014

Ce rapport de Mission «**Backing Digital Knowledge with New Services**" présente les résultats du suivi des actions initiées par la DIST en avril 2014. Ces actions ont été enrichies et précisées pour fournir les bases de plusieurs rencontres de travail qui se tiendront aux Etats-Unis dans la deuxième partie de l'année 2015. Le thème général de ces rencontres de travail est particulièrement stratégique pour le devenir des projets IST du CNRS : en mettant l'accent sur les services numériques d'accompagnement du travail de recherche ces rencontres permettront d'approfondir un ensemble de réponses à la question : « Comment cherchent les chercheurs ? ». Un certain nombre de jalons pourront donc être partagés sur la nouvelle génération de services d'appui à la recherche.

Je remercie le bureau Amérique du Nord du CNRS ainsi que nos amis et partenaires des organismes et associations américaines pour la qualité de l'accueil et des propositions qu'ils ont bien voulu faire à la mission de la DIST.

Renaud FABRE

Directeur de l'Information Scientifique et Technique

CNRS

Contenu

Tableau de synthèse		
NEW-YORK		
Partenaires	**Propositions**	**Dates envisagées**
Columbia University Libraries	Rebecca KENINSON suite à sa rencontre avec Phil Bourne,Associate Director for Data Science (ADDS) au NIH fera un retour à la DIST sur cette discussion. Amy NURNBERGER fait partie de la *Research Data Alliance* (RDA) et participera à la prochaine conférence en 2015 à San Diego. La DIST participe à RDA depuis sa création, représentée par Francis ANDRE, chargé de mission des données de la recherche. Ils seront mis en contact à cette occasion. Le *Libraries Information Services* de Columbia est invité à participer au séminaire organisé par le CNRS-DIST et NYU à New-York à l'automne 2015.	**Du 9 au 11 mars 2015** **Automne 2015**
Data Science Center (Yann LECUN – Facebook)	Yann LECUN a réaffirmé l'intérêt du Data Science Center pour le montage d'un séminaire commun entre le CNRS (DIST) et l'Université de New-York. Juliana FREIRE, professeur à NYU spécialisée dans les Sciences de l'ingénierie informatique (Computer Science and Engineering) sera l'interlocuteur de la DIST pour le montage de ce séminaire. D'autres institutions pourraient se joindre à ce séminaire, comme la fondation SLOAN ou encore la fondation MOORE.	**Automne 2015**
New-York University	La DIST et l'Université de New-York (*Data Science Center*) organiseront un séminaire commun à laquelle pourraient se joindre la *fondation* SLOAN la fondation MOORE et les Universités de Columbia et Washington. Une quinzaine d'experts seront invités à participer à ce séminaire. Un préprogramme sera transmis par la DIST à Juliana FREIRE qui procurera ensuite une liste Les questions de logistiques seront partagées par la DIST du CNRS et l'Université de New-York.	**Automne 2015**

Tableau de synthèse		
WASHINGTON		
Partenaires	**Propositions**	**Dates envisagées**
Ambassade de France aux Etats-Unis	Marc DAUMAS reviendra vers la DIST pour l'inviter à participer à l'une des conférences annuelles du NIST sur les données. L'ambassade rappelle son intérêt de mieux connaitre les directions des organismes et leurs domaines d'actions et d'intérêts afin de pouvoir effectuer au mieux les relais.	
American Chemical Society	Organisation d'un séminaire conjoint CNRS / ACS. L'ACS reprendra contact avec le CNRS – DIST pour mettre en place les échanges nécessaires à l'organisation de ce séminaire. Le programme et les experts participants (environ une quinzaine) seront proposés par l'ACS et le CNRS. Le bureau de Washington propose d'héberger cette réunion d'experts à l'ambassade et éventuellement le *social dinner* qui s'en suit.	**Automne 2015**
National Endownment for Humanities (NEH)	Le CNRS pourrait participer au programme « Digging into Data Challenge ». Marin DACOS directeur d'Open Edition (CNRS) est invité à rencontrer les équipes du NEH. Les équipes américaines du projet de Maurice GODELIER « Genèse de l'Etat et diversité de ses formes »recontacteront le NEH.	
Corporation for National Research Initiatives	Rencontre avec P.BAPTISTE, DGDS CNRS Désignation d'un membre français pour la DONA	**Novembre 2014**

Lundi 29 septembre 2014

❖ Libraries Information Services (Columbia University)

COLUMBIA UNIVERSITY LIBRARIES

http://library.columbia.edu/index.html

Rencontre avec :

- **Robert CARTOLANO**, Associate Vice President, Digital Programs and Technology Services
- **Rebecca KENNISON**, Director, Center for Digital Research and Scholarship
- **Amy L.NURNBERGER**, Research Data Manager, Center for Digital Research and Scholarship
- **Mark NEWTON**, Production Manager, Center for Digital Research and Scholarship
- **Simone SACCHI,** Research and Scholarship Initiatives Manager, Center for Digital Research and Scholarship
- **Leyla WILLIAMS,** Communications Coordinator, Center for Digital Research and Scholarship

Contacts	
Robert CARTOLANO Associate Vice President Digital Programs and Technology Services Libraries/Information Services Email : rtc@columbia.edu Adresse : 508 Butler Library 535 West 114th Street, New-York, NY 10027	**Rebecca KENNISON** Director Center for Digital Research and Scholarship Libraries/Information Services Email : rkennison@columbia.edu Adresse : 201 Lehman Library, International Affairs Building 420 West 118th Street, New-York, NY 10027

Amy L.NURNBERGER Research Data Manager Center for Digital Research and Scholarship Libraries/Information Services Email : ANurnberger@columbia.edu Adresse : 201 Lehman Library, International Affairs Building 420 West 118th Street New-York, NY 10027	**Mark NEWTON** Production Manager Center for Digital Research and Scholarship Libraries/Information Services Email : mnewton@columbia.edu Adresse : 201 Lehman Library, International Affairs Building 420 West 118th Street New-York, NY 10027
Leyla WILLIAMS Communications Coordinator Center for Digital Research and Scholarship Libraries/Information Services Email: lwilliams@columbia.edu Adresse : 201 Lehman Library, International Affairs Building 420 West 118th Street New-York, NY 10027	**Simone SACCHI** Research and Scholarship Initiatives Manager Center for Digital Research and Scholarship Libraries/Information Services Email : ssachi@columbia.edu Adresse : 201 Lehman Library, International Affairs Building 420 West 118th Street New-York, NY 10027

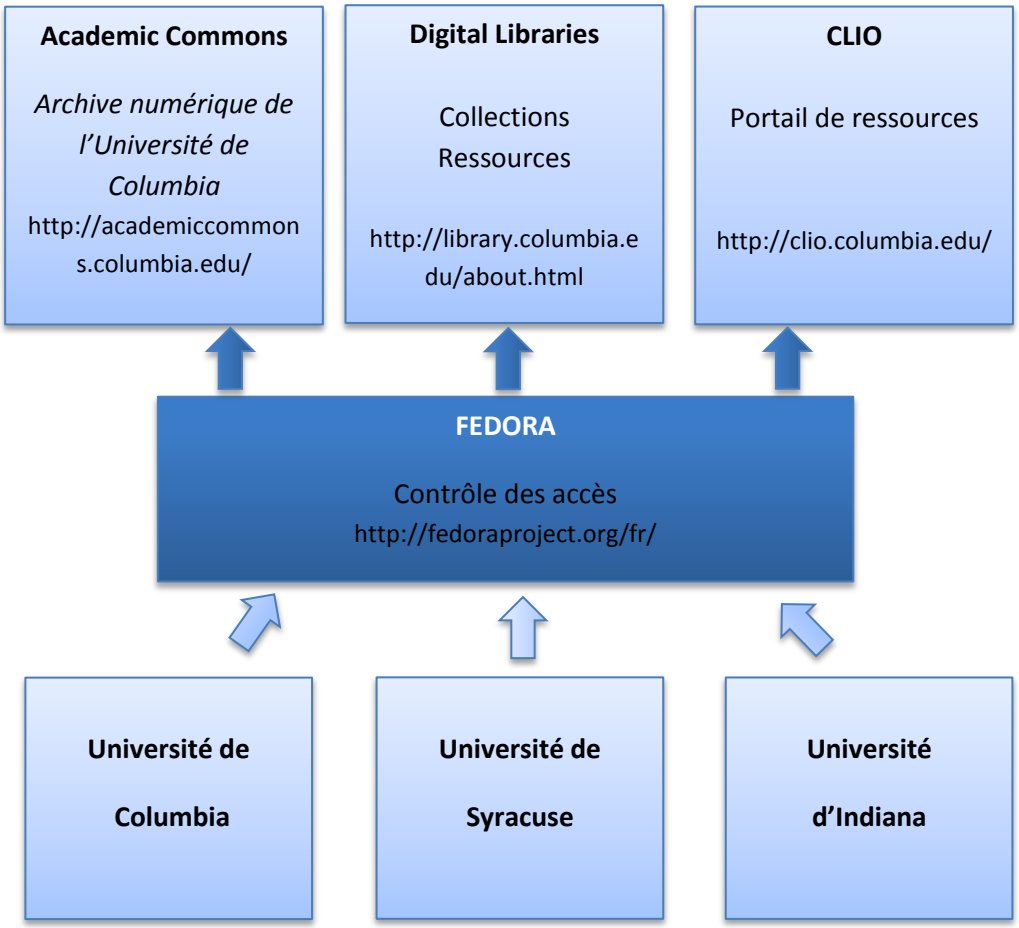
Le département de *Libraries Information Services* de l'Université de Columbia offre l'accès à une collection comprenant plus de 12 millions de volumes, plus de 160 000 revues et périodiques courants, et une vaste collection de ressources électroniques, des manuscrits, des livres rares, des microformes, cartes et graphiques et des documents audiovisuels.

L'Université de Columbia gère les numérisations et les acquisitions de fonds pour agrandir cette collection. La numérisation des ressources papier ou de certains enregistrements audio et vidéo peut être prise en charge par des sous-traitants ou par d'autres services de proximité. Le département de *Libraries Information Services* offre la possibilité à des personnes extérieures à l'Université d'archiver leurs contenus dans des archives de recherche en Open Access ou dans des archives fermées *(« black » archives*).

Par exemple, elle héberge l'archive de la collection FORD, pour laquelle elle est financée depuis 2011 par un programme de la Fondation Ford à hauteur d'un millions de dollars par an pendant 7 ans[1].

Un journal en Open Access est géré par l'Université de Columbia : le « *Columbia Undergratuaded Science Journal*[2] ». Ce journal est disponible sur la base de données institutionnelle.

Accès aux ressources de Columbia



Constats

Le constat partagé aujourd'hui est le fossé qui se crée entre la quantité de données produites et les moyens mis à disposition pour les stocker, les traiter et les archiver. L'Université de Columbia finance l'acquisition de nouvelles ressources pour la bibliothèque Butler mais l'analyse des données et leur conservation sont financées par subventions. Que faire lorsque les subventions arrivent à leur fin ?

[1] https://library.columbia.edu/bts/ford-ifp.html
[2] http://cusj.columbia.edu/about.shtml

Cette réflexion doit être collective : il faudrait créer des consortia pour que le travail se fasse dans le partage.

La préservation et la pérennité des données

Columbia participe à l'"Academic preservation trust"[3], projet initié par l'Université de Virginie associant plusieurs grandes universités américaines (Michigan, Stanford, Columbia, ...) qui s'engagent à la création et la gestion d'un environnement durable pour la préservation des données académiques et culturelles numériques.

L'évolution rapide des techniques (des logiciels,..) fait que les publications avec des données associées et des programmes, peuvent ne plus être reproductibles au bout d'un an, par disparition de l'environnement numérique associé (obsolescence). Cet état pose la question de la définition du périmètre de l'environnement à préserver. Or, il n'existe pas de budget pérenne fléché "gestion des donnée de la recherche", et la taille de ce sujet n'est pas encore évaluée. Les budgets qui traitent des données sont ceux des projets et ils s'arrêtent à la fin des projets. La perte potentielle d'information et de données qui s'en suit pourra induire un "digital dark age" une période noire pour le numérique, comme commencent à l'appeler les spécialistes du web de données faisant référence à la chute de l'empire romain période après laquelle très peu d'écrits permettent de relater l'histoire[4].

Le rôle moteur des bibliothèques dans la transformation numérique des institutions de recherche

Les évolutions des dernières années nécessitent de nouvelles politiques, règles et pratiques pour faire face aux nouveaux défis.

Dans les communautés scientifiques, le temps et la connaissance manquent aux chercheurs pour se pencher sur ces problématiques. Il existe un réel besoin de ligne budgétaire dédiée à la gestion des données : leur conservation, leur analyse, leur archivage. Les chercheurs doivent faire l'effort de documenter leurs travaux et données et ainsi faciliter l'interopérabilité, l'interdisciplinarité, l'échange des connaissances. Les documentalistes encouragent « la réutilisation et la reconversion » "*re-use and re-purpose*" des données. Certaines actions nécessiteront de définir un cadre légal, d'autres un cadre technique. Les bibliothèques se sont engagées dans ces directions. C'est le défi des 10 prochaines années.

---

[3] http://aptrust.org/

[4] http://rt.com/usa/232243-digital-development-documents-lost/

Auparavant l'information résidait dans les livres, sur papier dans les bibliothèques. Les bibliothèques sont toujours le lieu qui abrite l'information passée, présente et future. Elles doivent assurer l'accès à la connaissance. Aujourd'hui la connaissance se développe à grande vitesse et dépasse les frontières physiques.

Où doivent être construites ces bases de connaissances ?

Cadre juridique autour des données

Le numérique pose la question du droit d'auteur. Le nombre de données échangées et la facilité de création de nouvelles données numériques sont immenses et leur traçabilité difficile.

Le département de Robert CARTOLANO travaille en étroite collaboration avec le bureau consultatif des droits d'auteur de l'Université de Columbia (*Copyright Advisory Office*) pour définir des lignes directrices et des règles pour l'auto-dépôt dans les archives de l'Université de Columbia. Le cadre de travail au niveau du numérique évolue tellement vite que les textes sur le Copyright produits en 2000 doivent être mis à jour.

Le projet ISTEX[5]

L'Université de Columbia s'est intéressée à la prise en charge des différentes versions (Version Control) sur la plateforme ISTEX. Quelles sont les données acquises ? Sont-elles pertinentes ? Cette question trouvera des réponses avec le développement des services de la plateforme à échéance d'avril 2017 donnant la possibilité aux bénéficiaires du projet (chercheurs, enseignants chercheurs, …) d'effectuer des traitements sur les ressources acquises (Fouille de texte, enrichissements, …).

L'équipe du *Libraries Information Services* souhaite qu'à terme, les données des chercheurs ne soient plus stockées sur les plateformes privées, les bases de données institutionnelles mais sur des plateformes nationales voir internationales de type ISTEX.

La question de la succession des données en cas de disparition de l'institution qui les héberge (Succesion right planning) est soulevée. Le *Libraries Information Services* met actuellement en place son plan de succession. Columbia salue l'initiative Istex qui permet de rendre l'accès pérenne aux données de publication aux chercheurs. Effectivement, les bibliothèques ont longtemps sous-traité l'édition aux éditeurs qui n'avaient pas pour mission de gérer sa pérennité.

[5] http://www.istex.fr

Recommandations de H2020 en Europe et de l'OSTP pour les Etats-Unis

Aujourd'hui l'Europe et les Etats-Unis font face aux mêmes problématiques. Il faut à présent créer des moyens, des structures pour la préservation des données. Aujourd'hui c'est le rôle des institutions de les préserver, (IT institutions) mais comment les préserver ? Quand les préserver ? Comment faire les choix de données ?

Des discussions avec les chercheurs seront nécessaires pour documenter les méthodes et les procédés de recherche afin d'évaluer au mieux quel est l'environnement à mettre en place autour des communautés de recherche pour permettre la préservation de leurs données.

Identification du chercheur

Tout chercheur qui dépose et interagit avec une institution ou agence de financement, le fait à travers son « **identifiant recherche** ». Pour faciliter la cartographie de la recherche il faudrait normaliser le système d'identifiants, mettre en place un modèle national d'identification. Pour le moment aux Etats-Unis deux systèmes coexistent, celui d'ORCID[6] qui fournit des identifiants aux chercheurs et l'« Authority record[7] » qui permet la création d'un historique de la carrière des auteurs/chercheurs (publications scientifiques).

L'objectif est de parvenir à suivre le chercheur en ce qui concerne sa production scientifique mais aussi son institution de rattachement. Ainsi, toute institution pourrait connaitre sa production à tout instant.

Le **« Linked data for libraries » (LD4L)**[8] est un projet des universités de Cornell, Harvard et Stanford qui permet de lier les métadonnées de l'"Authority record" à celles des bibliothèques, et ainsi lier le chercheur aux référentiels des bibliothèques.

6 http://orcid.org/

7 http://www.loc.gov/marc/uma/

8 wiki.duraspace.org/pages/viewpage.action?pageId=4135402

Outils, projets et études

Le Projet ODIN[9] initié par *ORCID*[10] et le *DataCite Interoperability Network*

Ce projet de deux ans a débuté en septembre 2012. Il est financé par l'action de coordination et de soutien de la Commission européenne au titre du 7ème PCRD. Il se base sur le succès de ces 2 projets pour sensibiliser à l'usage d'identifiants pérennes pour les auteurs et les objets.

Ceci permettant : 1 La référence à une donnée. 2. Le suivi des usages et ré-usages. 3 Le lien entre la donnée (en tant qu'objet et ses constituants, articles, droits et toute personne impliquée dans son cycle de vie (créateur, éditeur, évaluateur, etc).

SIdora

SIdora est un logiciel conçu pour recueillir toute la production des chercheurs de la Smithsonian[11]. L'objectif de ce logiciel est de soutenir activement le processus de la recherche en permettant de conserver de manière pérenne, l'ensemble des contenus numériques issus d'un projet de recherche de façon à permettre leur curage si nécessaire.

Taverna tool kit[12]

Taverna est un système de gestion de flux de travail en open source, qui se décline en une suite d'outils utilisés pour concevoir et exécuter les workflows scientifiques. En effet, la recherche est de plus en plus productive en données et tributaire de l'utilisation de logiciels de simulation scientifique de plus en plus complexes. Les formats de données incompatibles produits ou utilisés engendrent des difficultés. Les Workflows scientifiques permettent de résoudre ces problèmes d'interopérabilité et d'homogénéisation et précisent les tâches qui doivent être effectuées au cours d'une expérience spécifique.

---

9 http://odin-project.eu/

10 http://orcid.org/

11 http://www.si.edu/

12 http://www.taverna.org.uk/download/workbench/2-5/digital-preservation/

Conclusion

Les équipes de la DIST et de Columbia resteront en contact pour partager leurs avancés sur les thèmes abordés et notamment :

- L'écart entre les données produites et les moyens en place pour les traiter, les analyser, les conserver
- Le cadre juridique, les règles, les chartes de bonnes pratiques autour des données
- Le management des données

Rebecca KENINSON suite à sa rencontre avec Phil Bourne,Associate Director for Data Science (ADDS) au NIH fera un retour à la DIST sur cette discussion.

Amy NURNBERGER fait partie de la *Research Data Alliance* (RDA) et participera à la prochaine conférence en 2015 à San Diego. La DIST participe à RDA depuis sa création, représentée par Francis ANDRE, chargé de mission des données de la recherche. Ils seront mis en contact à cette occasion.

Le *Libraries Information Services* de Columbia est invité à participer au séminaire organisé par le CNRS-DIST et NYU à New-York à l'automne 2015.

❖ **Facebook**

https://www.facebook.com/

Rencontre avec :

- **Yann LECUN,** Director of AI Research

Yann LECUN est le fondateur du ***Data Science Centre*** de l'Université de New-York.

Contact
Yann LECUN Director of AI Research Email : yann@fb.com Adresse: 770 Broadway, 8th floor New-York, NY 10003, USA

Lors de la 1ère mission CNRS-DIST aux Etats-Unis du 24 mars au 4 avril 2014 « *A Better Sharing of Knowledge* », l'équipe avait rencontré Yann LECUN alors directeur du *Data Science Center* de l'Université de New-York. Lors de cet entretien l'organisation d'un séminaire d'experts communs sur les principes et règles de régulation des plateformes de science avait été jugée intéressante.

Un séminaire regroupant le CNRS, le CERN et la DG Connect (Commission Européenne) se tiendra à Paris le 28 novembre 2014[13] et sera le premier d'une série de 3 séminaires qu'organisera la DIST du CNRS autour des problématiques des plateformes de science. Les séminaires suivants se tiendront aux Etats-Unis à New-York et Washington.

[13] http://www.cnrs.fr/dist/z-outils/documents/CERN_CNRS_DG-CONNECT%20Workshop_nov2014.pdf

Lors de la rencontre avec Yann LECUN, une discussion s'est tenue sur les problématiques des données sur les plateformes de science qui seront évoquées lors du séminaire du 28 novembre à Paris.

Le fossé entre la production et l'analyse des données de la Science

Les plateformes de science traitent et analysent des données. Des exemples de la croissance de ces données sont visibles à l'échelle mondiale et révèlent, entre autres, les raisons de ce développement et les conditions de stockage et de réutilisation des résultats dans les archives des laboratoires. Les preuves de l'écart croissant entre la production de données, la capacité et la qualité de leurs traitements, leurs analyses sont recueillies par le biais de divers canaux (calcul intensif, visualisation, infrastructure d'analyses, …). L'idée est de partager sur l'écart grandissant entre la production et le traitement des données, avec l'appui d'hypothèses pertinentes pour le partage de résultats.

Questionnement autour : de la gestion du flux des données sur les plateformes de Sciences pour traiter les données dans un processus de valeur ajoutée pour la Recherche.

Le partage de connaissance sur l'analyse des données

Données et publications sont de plus en plus en plus liées aux "objets numériques". Elles sont traitées par des outils d'analyse de différents types (nécessitant des logiciels de base et spécifiques) et font partie des processus de découverte scientifique qui conduisent à l'avancement de la connaissance partagée et de l'excellence. Les processus analytiques qui peuvent être menées sur des plateformes numériques avancée de Science constituent un vaste paysage comprenant des centaines d'outils de haute qualité (publications, données, modèles, logiciels, etc.)

Il est aujourd'hui nécessaire de développer des idées sur la façon dont les approches européennes et mondiales peuvent être coordonnées pour favoriser l'émergence de "collaboratoires" et des réseaux d'outils analytiques. Cet aspect est crucial pour le développement de la science interdisciplinaire qui n'est pas conduite uniquement par la création de données, mais par la capacité à les comprendre et à les traiter pleinement.

Questionnement autour : des architectures des plateformes numériques de pointe pour la science en Europe et au niveau mondial et comment celles-ci pourraient être liées entre elles pour favoriser la collaboration scientifique et le partage des connaissances est donc posée.

Le partage de bonnes pratiques et de règles pour le management des données

La science est en transformation pour intégrer de nouveaux paradigmes basés sur l'usage intensif de données numériques. Toutefois, les transformations doivent prendre en compte l'héritage en termes de technologie, de services et des pratiques établis dans les des communautés scientifiques et éducatives.

Les données de la recherche et les outils associés sont utilisés et gérés par des équipes scientifiques possédant leurs propres habitudes, leurs propres règles d'éthique et leurs propres capacités de partage. Les règles, les pratiques, les lois, sont dispersées aujourd'hui dans la mosaïque des utilisations.

Un besoin important existe également de cartographier et de mieux comprendre les principales tendances et les bonnes pratiques qui pourraient se développer au niveau européen pour favoriser l'ambition de l'excellence scientifique.

Il faut de même apprendre des nouvelles expériences disponibles dans les différents domaines de la science dans lesquelles les scientifiques ont pu appréhender la dynamique unique des plateformes numériques de pointe pour la Science.

Questionnement autour : des nouveaux cadres juridiques (droit mou ou droit dur) pour permettre aux plateformes numériques de pointe pour la Science de se développer et de réduire les barrières de partage de l'information primaire ainsi que permettre la confiance dans la Science pour un partage transparent des connaissances.

Conclusion

Suite à cette discussion Yann LECUN a réaffirmé l'intérêt du Data Science Center pour le montage d'un séminaire commun entre le CNRS (DIST) et l'Université de New-York. Juliana FREIRE, professeur à NYU spécialisée dans les Sciences de l'ingénierie informatique (Computer Science and Engineering) sera l'interlocuteur de la DIST pour le montage de ce séminaire. D'autres institutions pourraient se joindre à ce séminaire, comme la fondation SLOAN ou encore la fondation MOORE.

Mardi 30 septembre 2014

❖ New-York University (NYU)

NEW YORK UNIVERSITY

http://cds.nyu.edu/

Rencontre avec :

- **Juliana FREIRE**, Professor, Computer Science and Engineering

Contact
Juliana FREIRE Professor Computer Science and Engineering Email : Juliana.freire@nyu.edu Adresse : 2 MetroTech Center, 10th floor Room # 10.097 Brooklyn, NY 11201

Data Science Initiative

Depuis 2013, l'Université de New-York, l'Université de Californie et l'Université de Washington travaillent en collaboration sur un projet financé à hauteur de 37 millions de dollars soutenu par la *Gordon and Betty MOORE Foundation* et l'*Alfred P.SLOAN Foundation* avec pour objectif le développement de partenariats innovants pour l'évolution des technologies qui supportent le management des données et l'analyse technique des données.

Management des plateformes analyse et conservation des données

Constats partagés :

De parts et d'autres sont constatés :

1. Le fossé entre les grands volumes de données, et les capacités de stockage et d'analyse
2. La nécessaire élaboration de guides des bonnes pratiques, de chartes d'éthique et d'un cadre légal pour la gestion des plateformes
3. La nécessité de partager l'analyse et d'investir dans la science des données
4. La nécessité de former des trios, chercheur (expert scientifique), informaticien, documentaliste

Il n'y aura pas une seule réponse à ces constats mais différentes réponses en fonction des domaines de science.

L'Université de New York (NYU) intègre dans les cours qu'elle propose des modules consacrés aux données (gestion, analyse, conservation). Dans ces modules, les professeurs communiquent à leurs étudiants les « bonnes pratiques » pour l'utilisation des données produites dans les cours. L'objectif de NYU est de mettre en place pour ses étudiants, professeurs, chercheurs, un guide des bonnes pratiques, notamment pour la reproductibilité des données. Pour cela, les experts du domaine scientifique, les informaticiens, les experts en bases de données interagissent.

A ce jour, il est possible de mettre en place ces bonnes pratiques et les outils dans des domaines spécifiques mais ces processus sont difficiles à implémenter à un niveau global et interdisciplinaire.

L'Université de New-York cherche également à développer une politique forte de conservation des données dans des archives (fermées, en libre accès, ...).

Dans le futur, des cours consacrés à la Science des données (management, analyse, conservation) pourraient voir le jour.

Il existe aujourd'hui de nombreuses initiatives des communautés : par exemple Github, très utilisé pour les codes, De la même manière Git Data[14] est une initiative pour la gestion et l'échange de données Ces sites mettent à disposition des API permettant l'accès à la lecture et l'écriture d'objets Git dans les bases de données des utilisateurs.

[14] https://github.com/

Center for Urban research Science Program - CUSP[15]

Le *Center for Urban research Science Program*, qui a été créé par NYU, sous l'impulsion du Maire de New-York Michael BLOOMBERG, est un centre de recherche public-privé unique qui utilise la ville de New York comme laboratoire de recherche pour aider les villes à travers le monde à devenir plus productives, vivables, équitables, etc. Le CUSP observe, analyse et modélise les villes pour optimiser des résultats, de nouvelles solutions, formaliser de nouveaux outils et processus, et développer de nouvelles expertises. L'objectif du CUSP est de devenir leader mondiale dans le domaine émergent de l' «*Urban informatique* ».

Cadre juridique autour des données

NYU reconnait l'importance de mettre en place un cadre juridique (chartes de bonnes pratiques, règles) autour de l'utilisation des données et des réflexions sont en cours autour de ces aspects. Ces questions soulèvent des problèmes de propriétés privées. A qui appartiennent les données ?

Conclusion

La DIST et l'Université de New-York (*Data Science Center*) organiseront pour l'automne 2015 un séminaire commun à laquelle pourraient se joindre la *fondation* SLOAN la fondation MOORE et les Universités de Columbia et Washington.

Une quinzaine d'experts seront invités à participer à ce séminaire.

Un préprogramme sera transmis par la DIST à Juliana FREIRE qui procurera ensuite une liste d'experts dont par exemple, Josh GREENBERG (directeur à la Sloan du programme technologies de l'information numérique, l'un des fondateurs du projet ZOTERO. Les questions de logistiques seront partagées par la DIST du CNRS et l'Université de New-York.

[15] http://cusp.nyu.edu/about/

❖ **Ambassade de France aux Etats-Unis**

http://france-science.org

Rencontre avec :

- **Marc DAUMAS,** Attaché pour la Science et la Technologie, Ambassade de France aux Etats-Unis

Présentation de la Mission Scientifique française aux Etats-Unis et du Service Scientifique (CNRS).

Contact
Marc DAUMAS Professeur des universités Attaché pour la Science et la Technologie Email : marc.daumas@diplomatie.gouv.fr Adresse : Ambassade de France aux Etats-Unis 4101 Reservoir Road NW Washington, DC 20007

Leurs programmes, leurs missions, leurs évènements

- STEM Chateaubriand Fellowship[16]

Le *STEM Chateaubriand Fellowship* est un programme de recherche qui regroupe environ 30 étudiants par an sur 4 à 9 mois. Ce programme permet de créer des liens entre des doctorants américains (3ème ou 4ème année de thèse – 5 ans de thèse pour les Etats-Unis) et les laboratoires français. Le programme existe depuis une quinzaine d'années et permet de mettre en place des cotutelles et de renforcer les partenariats

[16] http://stem.chateaubriand-fellowship.org/

entre chercheurs américains et chercheurs français. La NSF possède un programme similaire destiné à des collaborations avec l'Europe du Nord.

- New Technology Venture Accelerator (NETVA)[17]

Ce programme est destiné aux entreprises françaises qui souhaitent s'établir aux Etats-Unis ou travailler en collaboration avec les Etats-Unis. Il s'agit d'un programme léger qui permet une immersion des PDG d'entreprises dans des sociétés américaines sur une période d'une semaine. Ce programme concerne les entreprises d'innovation et les petites entreprises. Le coût est d'environ 1700€. Ce programme s'étend également au Canada.

- Young Entrepreneur Initiative (YEI)[18]

A l'inverse de NETVA, ce programme s'adresse aux entrepreneurs américains qui souhaitent venir s'établir en France.

D'autres missions / programmes / évènements sont également menés par la Mission Scientifique française aux Etats-Unis comme les Cafés de la science, France Atlanta, *The French American Innovation Day,* …

Horizon2020

Horizon2020 a demandé à la Mission Scientifique française aux Etats-Unis la mise en place de projets en coopération avec les Etats-Unis mais, dans le cadre de ces projets, seuls les partenaires européens sont financés et non les équipes américaines. Des partenariats bi-financés pour les projets Europe/Etats-Unis dans le cadre d'Horizon2020 n'existent pas encore.

Par ailleurs, le NIH serait prêt à financer le partenaire américain dans des projets Europe / Etats-Unis sur le budget de défense américain.

[17] http://www.netvafrance.com/

[18] http://www.yeifrance.com/

National Institute of Standards and Technology (NIST)

Le *National Institute of Standards and Technology* organise des évènements ouverts tout au long de l'année notamment sur le sujet des BIG DATA : http://bigdatawg.nist.gov/home.php.

La Mission Scientifique Française aux Etats-Unis est en relation avec le NIST qui a constitué, par exemple, le groupe de travail « Data Science ». Une réflexion conjointe est en cours sur les stratégies autour des données. Les documents liés à ce groupe de travail sont accessibles en ligne http://www.nist.gov/itl/iad/data-science-symposium-2014.cfm

Conclusion

Marc DAUMAS reviendra vers la DIST pour l'inviter à participer à l'une des conférences annuelles du NIST sur les données.

L'ambassade rappelle son intérêt de mieux connaitre les directions des organismes et leurs domaines d'actions et d'intérêts afin de pouvoir effectuer au mieux les relais.

Mercredi 1er octobre 2014

❖ American Chemical Society (ACS)

http://www.acs.org/international

Rencontre avec

- **Bradley D.MILLER**, Director, Office of International Activities
- **Brandon NORDIN**, Vice president , Sales Marketing & Web Strategy-Publications

Contacts	
Bradley D.MILLER, Phd Director Office of International activities Email : b_miller@acs.org Adresse: American Chemical Society 1155 Sixteenth Street, NW Washington, DC 20036 USA	**Brandon NORDIN** Vice President Sales, Marketing & Web Strategy Publications Email : b_nordin@acs.org Adresse: American Chemical Society 1155 Sixteenth Street, NW Washington, DC 20036 USA

Séminaire commun ACS/CNRS

Lors de la 1ère mission DIST au Etats-Unis du 24 mars au 4 avril 2014 l'organisation d'un séminaire commun CNRS / ACS, sur les modèles économiques de la publication sur les plateformes de chimie, avait été envisagée.

En effet de parts et d'autres sont constatés :

1. Le fossé entre les grands volumes de données, et les capacités de stockage et d'analyse
2. La nécessaire élaboration de guides des bonnes pratiques, de chartes d'éthique et d'un cadre légal pour la gestion des plateformes
3. La nécessité de partager l'analyse et d'investir dans la science des données

La définition des modalités de ce séminaire ont été définis lors de cette mission de suivi:

- Thèmes abordés
- Mise en place des échanges pour l'organisation
- Date

Le séminaire avec l'ACS se basera sur le même modèle que le séminaire qui se tiendra à Paris avec le CNRS le CERN et la DG Connect (Commission Européenne). Le programme final de ce séminaire ainsi que les actes seront transmis à l'ACS au début de l'année 2015 pour déterminer les thèmes communs à aborder lors du séminaire à Washington

http://www.cnrs.fr/dist/z-outils/documents/CERN_CNRS_DG-CONNECT%20Workshop_nov2014.pdf.

Conclusion

Le séminaire conjoint CNRS / ACS se tiendra à l'automne 2015.

L'ACS reprendra contact avec le CNRS – DIST pour mettre en place les échanges nécessaires à l'organisation de ce séminaire. Le programme et les experts participants (environ une quinzaine) seront proposés par l'ACS et le CNRS.

Le bureau de Washington propose d'héberger cette réunion d'experts à l'ambassade et éventuellement le *social dinner* qui s'en suit.

Jeudi 2 octobre 2014

❖ National Endowment for Humanities (NEH)

http://www.neh.gov/

Rencontre avec :

- **Brett BOBBLEY**
- **Jennifer SERVENTI**

Contacts	
Brett BOBBLEY Chief Information Officer Director, Office of Digital Humanities Email : bbobley@neh.gov Adresse : Constitution Center 400 7th street SW Washington DC 20024	**Jennifer SERVENTI** Senior Program Officer Office of Digital Humanities Email : jserventi@neh.gov Adresse : Constitution Center 400 7th street SW Washington DC 20024

Open Access

Le NEH incite les chercheurs à déposer en Open Access sur les plateformes qu'il subventionne telles que « Open Context » ou « the Perseus Digital Library » qui gagnent en popularité grâce à l'appui du NEH.

Digging into Data Challenges[19]

Le *Digging into Data Challenge,* est un programme de subventions parrainé par plusieurs grands organismes de recherche internationaux, et a pour but d'analyser comment les "Big Data" changent le paysage de la recherche en Sciences Humaines et Sociales.

[19] http://diggingintodata.org/

Aujourd'hui les bases de données massives disponibles pour la recherche dans les sciences humaines et les sciences sociales sont de toutes sortes : livres numérisés, journaux, musique, information générée par les activités Internet et les communications mobiles, les données administratives des organismes publics, les bases de données clients des organisations du secteur privé, etc. Face à toutes ces données et aux nouvelles méthodes de recherche, et alors que le monde devient de plus en plus numérique, de nouvelles techniques seront nécessaires pour rechercher, analyser et comprendre ces matériaux.

Le *Digging into Data Challenge* incite la communauté de la recherche à développer et accompagner les nouvelles infrastructures de recherche du 21ème siècle à travers des financements et subventions.

Première vague (2009)
A son lancement en 2009, le *Digging into Data Challenge* a été parrainé par quatre financeurs de recherche (NEH, NSF, CRSH, JISC) représentant les États-Unis, le Canada, le Royaume-Uni. Finalement, 8 projets internationaux ont été retenus. Plusieurs de ces projets ont fait l'objet de publications dans des journaux tels que le New York Times. Ces projets ont également fait l'objet d'un important rapport de recherche publié par le CLIR (Council on Library and Information Resources). Les 8 projets ont été présentés lors d'une conférence à Washington en juin 2011.

Deuxième vague (2011)
Quatre bailleurs de fonds supplémentaires ont rejoint le *Digging into Data Challenge* en 2011. Lors de cette deuxième vague, 14 projets ont été primés, choisis par des experts internationaux. Les 14 projets ont été présentés lors d'une conférence tenue à Montréal au Canada le 12 octobre 2013.

Troisième vague (2013)
Deux autres financeurs ont rejoint le projet en 2013. 10 pays participent à présent au *Digging into Data Challenge*. A la fin de cette troisième manche 14 projets ont été primés.

Les bibliothèques numériques, les archives et les musées représentent une part importante du *Digging into Data Challenge.* Ce sont les organisations qui créent, regroupent, curent, et préservent les données numériques étudiées par les chercheurs. Une liste des référentiels de données créés lors des projets du *Digging into Data Challenge* est mise à disposition des chercheurs.

Brett BOBBLEY invite le CNRS à intégrer ce programme.

Trans-Atlantic Platform [20]

Le *Trans-Atlantic Platform* est un partenariat entre 15 organismes financeurs de recherche d'Europe et d'Amérique. Cette plateforme a pour objectif d'améliorer la collaboration dans la recherche transatlantique dans des domaines clés d'engagements mutuels répondant aux défis de la société du 21e siècle impliquant les Sciences Humaines et Sociales (SHS). L'Agence National pour la Recherche (ANR) en France participe à cette collaboration. L'apport de chaque participant est de 1 millions d'euros pour le financement de ses équipes sur environ deux ans et demi.

Présentation du projet de Maurice GODELIER « Genèse de l'Etat et diversité de ses formes »

Le NEH propose des subventions pour des projets SHS comme celui proposé par Maurice GODELIER « Genèse de l'Etat et diversité de ses formes » et est prêt à rencontrer les membres américains du consortium proposés afin d'évaluer les subventions auxquelles ce projet pourraient candidater.

Conclusion

Le CNRS pourrait participer au programme « Digging into Data Challenge ».

Marin DACOS directeur d'Open Edition (CNRS) est invité à rencontrer les équipes du NEH.

Les équipes américains du projet de Maurice GODELIER « Genèse de l'Etat et diversité de ses formes »recontacteront le NEH.

[20] www.transatlanticplatform.com

Vendredi 3 octobre 2014

❖ Corporation for National Research Initiatives (CNRI)

http://www.cnri.reston.va.us/

Rencontre avec :

- **Robert KAHN**, President, CNRI
- **Patrice LYONS**, General Counsel, CNRI

Robert Kahn, rencontré lors de la première mission de la DIST aux Etats-Unis, a indiqué la création en janvier 2014 de la fondation DONA (Digital Object Numbering Authority), organisation à but non lucratif www.dona.net. Elle est basée dans le canton de Genève et hébergée à l'Université de Genève. Christophe BLANCHI a été élu Directeur Exécutif en juillet 2014 et sera salarié dès janvier 2015.

Ses objectifs

La DONA a pour mission de favoriser l'interopérabilité entre les systèmes d'information hétérogènes. Elle permet et étend l'utilisation des architectures d'objets numériques en :

1. Fournissant **une gestion**, le **développement de logiciels** et d'autres services stratégiques pour la **coordination technique**, l'évolution, l'application et d'autres utilisations d'intérêt public autour de l'architecture des objets numériques.

2. Dans le cadre de ces prérogatives, **administrant et maintenant stable** le fonctionnement **du GHR** (Global Handle Registry), un élément essentiel de l'architecture des objets numériques, et **autorisant et coordonnant** l'administration du GHR avec les MPA (Multi-Primary Adminstrator).

Philippe BAPTISTE, DGDS du CNRS, a donné son accord pour que la France soit représentée au Conseil d'Administration de la DONA par le CNRS, si cela était possible.

Les personnalités membres de ce CA doivent maintenir la diversité présente à ce CA et **l'équilibre géographique** dans la mesure du possible. Elles doivent être **engagées**

dans le domaine de la libre architecture des objets numériques et posséder **le bagage technique et l'expertise appropriée** pour réaliser les objectifs de la DONA.

De plus, il a été retenu que les candidats potentiels devraient être des organisations multipartites et avec une **certaine indépendance gouvernementale**.

La création de la DONA permet à des parties autres que le CNRI d'administrer le GHR et de répartir ainsi cette responsabilité sur plusieurs organisations. Dès la fin de l'année ou dès la fin des contrats en cours avec le CNRI (anciennement le seul habilité à cette gestion), les établissements, organisation, entités souhaitant utiliser le système Handle pourront s'adresser à l'une des parties présente à la DONA.

Les MPAs initiaux désignés en juillet 2014

- Corporation for National Research Initiatives (CNRI)
- Coalition for Handle Services -- China (ETIRI, CHC and CDI)
- International Telecommunication Union (ITU- Union internationale des télécommunications)
- Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG)

Les organisations pourront à nouveau comme précédemment s'adresser au CNRI (basé aux Etats-Unis) ou au GWDG (basé en Europe). Les ONG à l'ITU et ceux qui ne souhaitent s'adresser à aucune de ces parties pourront s'adresser à la Chine.

Les membres du Conseil d'Administration de la DONA - Réunion juillet 2014

Dr Robert Kahn, (Président), CNRI
Dr Stephen Wolff – Internet2. Internet international cooperations.
Dr. Peter Wittenburg - The Language Archive & Max Planck Institute for Psycholinguistics, First to lead the RDA (Research Data Alliance)
Dr. Norman Paskin International DOI Foundation,
Dr. Antoine Geissbuhler Division of eHealth and Telemedicine (standard for medical records and medical information), Genève, University & Hospitals
Mr.Stefan Eberhard (Secrétaire) ABELS Avocats Geneva (legal interface with the government in Switzerland : tax exemption, corporate stiff). He will step out the board as soon as he is not needed.
It could become an international organization.
Mr. Adama Samassekou Président du CIPSH, Président du Réseau MAAYA, Ancien Secrétaire Exécutif de l'ACALAN BP E 214 Bamako – MALI, he was the president of the 2 world security summit.
Mr. Gao Xinmin - Deputy Head of the Internet Society of China.

Recommandation UIT-T X.1255

La DONA a produit la Recommandation UIT-T X.1255 qui a pour objet de définir un cadre d'architecture ouverte dans lequel il est possible de découvrir des informations relatives à la gestion d'identité (IdM Identity Management).

L'Union internationale des télécommunications (UIT) est une institution spécialisée des Nations Unies dans le domaine des télécommunications et des technologies de l'information et de la communication (ICT).

Ces informations seront nécessairement représentées de différentes manières et seront prises en charge par divers cadres de confiance ou d'autres systèmes IdM utilisant différents schémas de métadonnées. Ce cadre permettra par exemple à des entités fonctionnant dans le contexte d'un système IdM de résoudre avec précision des identificateurs provenant d'autres systèmes IdM. Les utilisateurs ou les organisations (ou les programmes exploités pour leur compte) qui ne sont pas en mesure de découvrir ces informations n'ont d'autre choix que de déterminer la meilleure façon d'établir la crédibilité et de vérifier l'authenticité d'une identité adéquate, que ce soit pour un utilisateur, une ressource de système, une entité d'information, etc.

A la lumière de ces informations, il revient à l'utilisateur ou à l'organisation de déterminer si, aux fins considérées, il peut ou non se fier à un cadre de confiance donné ou à un autre système IdM.

Les éléments principaux du cadre présenté dans cette recommandation sont notamment:
1) un modèle de données d'entité numérique ;
2) un protocole d'interface d'entité numérique ;
3) un ou plusieurs système(s) de résolution/d'identificateur ;
4) un ou plusieurs registres de métadonnées.
Ces éléments constituent la base du cadre d'architecture ouverte.

LE CNRI

Dr. Robert KAHN a développé le concept d'architecture des objets numériques. Cette notion fournit un cadre pour l'interopérabilité des systèmes d'information hétérogènes et permet de nombreuses applications telles que le Digital Object Identifier (DOI). Il est également le co-inventeur des programmes de Knowbot, agents logiciels mobiles dans l'environnement réseau

Parmi les programmes

CNRI développe une infrastructure pour les **programmes Knowbot** Ce sont des agents mobiles destinés à être utilisés dans des systèmes largement distribués comme Internet. Ils fournissent une architecture simple d'utilisation pour le développement de systèmes sécurisés d'agents distribués. Les agents peuvent être tout type de logiciels, et l'architecture permet plusieurs langages de programmation. Les agents sont clairement définis, ce sont des entités autonomes qui interagissent avec leur environnement dit " stations-service ", selon des règles précises.

Un logiciel de Knowbot pourrait être utilisé dans diverses applications, par exemple dans l'extraction de données, pour appuyer la négociation de protocole entre les systèmes distribués géographiquement, et servir de médiateur d'accès à l'information dans un environnement de réseau. Une application cible pour ces systèmes sont les bibliothèques numériques qui contiennent des documents sensibles. L'utilisation d'un système à base d'agents d'accès pourrait permettre à toute entité à voir un document dans la bibliothèque, mais de restreindre la modification ou la suppression d'un document à des entités avec une cote de sécurité requise. L'ouverture des bibliothèques numériques aux agents de confiance permettra des recherches plus souples sur de plus grands ensembles de données.

Le système Handle est une offre efficace, extensible, et permet des services de règlement sécurisés pour les identificateurs uniques et persistants d'objets numériques, et est une composante du Digital Object Architecture du CNRI. L'architecture des objets numériques fournit un moyen de gestion de l'information numérique dans un environnement de réseau. Un objet numérique présente une structure indépendante (machine et plate-forme) qui lui permet d'être identifié, accessible et protégé, le cas échéant. Un objet numérique peut intégrer non seulement des éléments d'information, à savoir, une version numérisée d'un document, une vidéo ou un enregistrement sonore, mais également l'identifiant unique de l'objet numérique et d'autres métadonnées sur l'objet numérique. Les métadonnées peuvent inclure des restrictions sur l'accès aux objets numériques, des avis de propriété, et des identifiants pour des accords de licence, le cas échéant.

Le système Handle comprend un ensemble de protocoles ouverts, un espace de noms, et une implémentation de références des protocoles. Les protocoles permettent un système informatique distribué pour stocker des identifiants de ressources arbitraires, appelés Handle, et former ces Handle dans les informations nécessaires pour localiser, accéder, contacter, authentifier, ou faire tout autre usage des ressources. Cette information peut être modifiée au besoin pour refléter l'état actuel de la ressource identifiée sans changer son identifiant, permettant ainsi la persistance du nom de l'élément lors de changements de lieu et d'autres informations d'état connexes. Quelques exemples d'applications qui utilisent les services d'identification et de résolution HDL® que les infrastructures sont des applications de gestion des droits , les registres et dépôts d'objets numériques, et institutionnel conservation des données et l'archivage .

Conclusion

Philippe BAPTISTE, DGDS du CNRS, rencontrera Robert KAHN afin d'étudier l'opportunité pour le CNRS de faire partie du Conseil d'Administration de la DONA.