La gestion des données de la recherche : retours d'expérience

Journée d'étude organisée par le secteur ADBS Recherche et InVisu (CNRS-INHA)
16 janvier 2015, Salle Vasari - INHA, 2 rue Vivienne 75002 Paris

Appel à communication

"Cette journée a pour objectif d'offrir aux participants une vision pratique des projets de gestion de données de la recherche qui peuvent être menés, en sciences dures ou en sciences humaines dans les laboratoires de recherche privés ou publics et quel que soit le type de données : corpus de texte, données de mesure, données d'essai, données d'enquêtes, images.

Les journées d'étude portant sur *les données de la recherche* sont de plus en plus nombreuses mais restent souvent très théoriques et/ou abordent un aspect particulier de cet objet, relativement nouveau, qui soulèvent de nombreuses problématiques techniques, pratiques, épistémologiques, logistiques. Nous souhaitons, dans cette journée d'étude, réellement « **voir » des projets de gestion de données en train de se faire,** aborder les questions que ces démarches posent, identifier et comprendre les freins et les difficultés, avoir des exemples de ce qui fonctionne et ce qui ne fonctionne pas.

Nous cherchons donc **des retours d'expériences** de projets de gestion des données de la recherche, de la part de personnels IST (ou de personnels ayant des fonctions de soutien à la recherche) et de chercheurs, qui puissent nous donner des pistes de réflexions sur :

- Le « qui fait quoi, quand et comment » dans un projet de recherche par rapport à la gestion des données : comment construire un plan de gestion de données avec la question centrale du cycle de vie des données, de la collecte puis la sélection des données (ce que l'on garde ou pas) à la pérennisation de ces données, en passant par la co-construction des corpus de données en collaboration entre les chercheurs et les ingénieurs

- Les processus, les démarches et les outils choisis ou construits pour une mise à disposition des données : il s'agit d'aborder notamment les questions d'interopérabilité et de réutilisation des données (travail sur les métadonnées) : standards, formats et normes de description des données ; contextualisation des données utilisation des référentiels, des ontologies…et avec en filigrane les questions d'ordre juridique."

Pour retrouver les tweets de la journée : #ADBSdonnees

Animation de la journée :
Jean-Luc Minel, MoDyCO, Université Paris Ouest Nanterre La Défense - Université Paris Descartes - CNRS

09:30 - 09:45 : Introduction et présentation de la journée

Odile Contat [INSHS - CNRS] et Juliette Hueber [InVisu, INHA - CNRS]

Remerciements aux organisateurs, ADBS et INHA et présentation de Jean-Luc Minel, animateur de la journée - http://www.jeanlucminel.fr/cv.html

L'organisation d'une journée dédiée aux données de la recherche a semblé aller de soi car c'est une question centrale aujourd'hui dans tous les laboratoires de recherche.
Si cette question est devenue centrale c'est notamment grâce ou à cause des directives du programme européen "Horizon 2020". Le document intitulé « Lignes directrices pour la gestion des données dans Horizon 2020 » de décembre 2013, commence ainsi « Le programme Horizon 2020 mettra en œuvre une action pilote sur le libre accès des données de recherche. Il sera exigé des projets participants qu'ils élaborent un plan de gestion des données (PGD) précisant les données qui seront libres »…
Si ce sujet des données s'est imposé c'est aussi parce que les professionnels de l'information scientifique et technique sont de plus en plus sollicités sur ces questions. On évoque souvent les plans de gestion de données ou la curation des données mais ces notions ne sont encore pas ou peu maîtrisées et pour beaucoup le besoin est de savoir ce que cela recouvre pratiquement. La question sous-jacente est comment, en tant que documentaliste, bibliothécaire et enfin professionnel de l'IST, l'on peut et l'on doit se positionner pour participer à la gestion des données de la recherche, apporter nos compétences en matière de référentiels, de normes et de métadonnées.
C'est pourquoi cette journée a été conçue dans l'objectif d'offrir une vision pratique des projets de gestion de données de la recherche qui peuvent être menés, en sciences dures ou en sciences humaines dans les laboratoires de recherche privés ou publics et quel que soit le type de données : corpus de texte, données de mesure, données d'essai, données d'enquêtes, images.
La volonté était de positionner cette journée d'étude sur les dimensions concrètes des projets de gestion de données de la recherche grâce à des retours d'expérience(s)
>> *Cf. questions formulées dans l'appel à communication*

Cette problématique, ces projets demandent de s'arrêter un instant sur la question de la définition de la donnée, des données de la recherche : qu'est-ce qu'une donnée de recherche ? On constate en effet que ce terme est de plus en plus employé et recouvre différentes "choses", différents "états" de la donnée, et pose à son tour d'autres questions de nature autant épistémologique que documentaire et technique.
Pour chaque personne qui en parle ce que recouvre le terme semble évident mais chacun a sa propre définition, selon les métiers et les disciplines les définitions divergent fortement… Les concepts s'entrechoquent, se superposent... Et si l'on y regarde de plus près on trouve des données personnelles, des données publiques, des données de la recherche, du big data, de l'open data, des métadonnées…
Référence au mémoire de fin d'étude de Rémi Gaillard* pour appréhender cette notion - une première distinction est opérée entre données primaires ou brutes et données secondaires ou issues d'un traitement.
* De l'open data à l'open research data :
http://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche).
Urfist Info nov 2013 "Données de la recherche les mal nommées" - http://urfistinfo.hypotheses.org/2581

Selon les champs disciplinaires et les métiers, la réalité que représentent les données de la recherche est très diverse, allant des relevés effectués par les satellites au prises de vue de bâtiments en passant par la collecte de zooplancton, les enquêtes de terrain ou des jeux de tweets.
Il n'y a pas que des données quantitatives de même nature qui constitueraient des séries homogènes relativement faciles à manipuler, échanger et compiler.
Bien entendu, la nature des données a des conséquences sur le traitement et la gestion des données, ou plutôt ajoute des questions spécifiques, par exemple les données d'enquêtes nécessitent une

anonymisation, un corpus de texte ne se traite pas comme des données sismographiques. Mais des questions fondamentales, qui vont nous intéresser tout au long de cette journée, restent communes : elles concernent le cycle de vie et la pérennité des données (que garde-t-on?, combien de temps?), la description des données (métadonnées, référentiels et vocabulaires contrôlés), et la diffusion des données (Interopérabilité, réutilisation, open data).

*** *** *** *** *** ***

09:45 - 10:30 : Construire des outils pour la gestion des données de la recherche dans une communauté d'universités

Aurore Cartier, Magalie Moysan, ~~Nathalie~~ ~~Reymonet~~ [Université Paris-Descartes, Université Paris-Diderot, Sorbonne Paris Cité]

"Produire un data management plan (i.e. plan de gestion de données)" : cette intervention est centrée sur la présentation d'un outil concret - le plan de gestion de données

L'intervention rappelle dans un premier temps le contexte européen de la recherche :
L'Open Access dans Horizon 2020 : la communauté européenne encourage désormais fortement à rendre accessibles les données de la recherche (selon leurs spécificités) et à publier les résultats de la recherche dans des revues modèles Gold ou Green
>> Deux documents à rédiger : un plan de dissémination et un plan de gestion de données

Dans le cadre de la COMUE, qui favorise (et encourage) toute forme de coopération entre universités, création de postes d'ingénieurs projets européens pour accompagner les chercheurs dans cette démarche :

L'ingénieur projet est là pour coordonner les projets européens - besoin de monter en compétence pour répondre aux nouvelles exigences > rédaction d'un guide
Cf. Schéma des différents acteurs diapo 4
Formation organisée pour ces ingénieurs projets européens sur l'open access et les données.

Que dit l'*Open Research Data Pilot* ?

- Il faut rédiger un Data Management Plan (DMP), lequel doit être remis dans les 6 mois suivant le dépôt d'un projet européen.
  Voir : https://www.fosteropenscience.eu/content/horizon-2020-open-research-data-pilot
- L'exigence du *data management plan* concerne les données publiées et les données non publiées.
- Le premier périmètre des projets appelés à présenter un DMP concerne surtout des champs de recherche des nouvelles technologies (STIC), les projets relevant du champ "sciences et société, climat, etc.", des sujets qui présentent *a priori* un intérêt pour le partage des données.

Dissémination : Un entrepôt de données scientifiques issu du projet européen OpenAIRE (Zenodo)
Préservation : Projet EUDAT (http://www.eudat.eu/) participation du CINES (comme centre de ressources).

Traduction française des deux guides pour Horizon 2020 :
http://www.donneesdelarecherche.fr/spip.php?article531

L'élaboration d'un plan de gestion de données : il s'agit d'un **document prospectif et évolutif destiné à être enrichi à mesure de l'avancement du projet.**
Première étape : sensibiliser les chercheurs : pourquoi gérer les données ?

La recherche produit des données, une partie produite, une partie traitée, une partie retenue - ce qui est invisible et ce qui est amené à être visible (de plus en plus)
Il convient par ailleurs de considérer les conséquences juridiques et techniques (qui sont les propriétaires des données ? garantir leur accès… Quels sont les formats des données et les dispositifs de lecture associés ? Garantir la conservation de ces derniers est aussi important.

DMP : document contractuel qui décrit le cycle de gestion de toutes les données qui seront collectées, traitées ou générées par un projet de recherche.
Principe d'anticipation : on retrouve aussi l'importance accordée au cycle de données.

Schéma des étapes de rédaction et de validation du DMP pour H2020.
Proposition de règles de nommage et de plan de classement.

Développement d'un guide de rédaction d'un DMP
Documents qui puissent être remis aux chercheurs, plan de gestion type demandé par les ingénieurs projets européens.
"Ce guide devait pouvoir répondre à un projet de recherche dans diverses disciplines. L'aspect universitaire a amené une réflexion sur la mise en application de ce plan de gestion sachant que les directions de la recherche ne sont pas orientées également selon les établissements."
Ce guide rassemble ainsi sur un document unique des informations qui auparavant existaient dans les projets dans divers documents.
Impliquer un acteur (dans le projet de recherche) dans la rédaction et la mise à jour du DMP (qui ait une vision globale du projet)

La question de la définition d'un jeu de données s'est posée : groupe de données qui a une homogénéité technique ou groupe de données hétérogènes techniquement mais ayant une homogénéité intellectuelle.

Identification des données et des jeux de données avec un identifiant unique.

Embargo de 6 mois pour les sciences dures et de 12 mois pour les sciences humaines.

Compétences et acteurs dans la mise en oeuvre d'un DMP : c'est la réunion de plusieurs acteurs // compétences
Schéma des acteurs (personnes ressources) du plan de gestion de données :

- Chercheurs (coordinateur DMP), ingénieurs projet, service informatique, référents IST, archivistes.
- C'est le chercheur qui est responsable du DMP. A l'échelle du projet il est possible de projeter le financement de formations pour faire monter en compétence des acteurs ou recruter.

  Les professionnels de l'IST / bibibliothécaires / archivistes vont intervenir sur la partie métadonnées / description des données et sur la partie dissémination / diffusion / publication / archivage

La rédaction du guide s'est accompagnée d'une démarche de formation des ingénieurs projets sur sa mise en application.
La question de l'informatisation du DMP s'est également posée, afin d'en faciliter l'accès et la diffusion.
Les intervenantes travaillent avec l'ensemble de la COMUE et les réseaux professionnels (ADBU et AAF)

Formation prévue en juin 2015 (à destination des ingénieurs projet) : retenue dans le cadre du projet européen FOSTER (https://www.fosteropenscience.eu/)

Il est primordial de réfléchir aux acteurs impliqués : il y a des spécificités françaises et il n'est pas possible de se contenter d'une traduction de DMP anglo-saxons.

Mise en ligne du document :"Guide pour la rédaction d'un DMP" sur le site de Paris-Diderot et sur les réseaux professionnels, en licence Creative Commons.
Objectif de décloisonner des compétences.

Questions :
Nécessité de se positionner aussi sur ces questions pour les archivistes et les bibliothécaires
Nécessité de recruter des archivistes et sinon, qui prend en charge l'archivage ? (CINES)
Le guide pour le moment n'a été que préparé pour les ingénieurs projet Europe et pas avec les chercheurs.
Les universités poussent de plus en plus à remporter des projets européens mais les chercheurs souhaitent de moins en moins y répondre en raison de la lourdeurs administrative de ces projets. Les ingénieurs Europe sont eux largement favorables et poussent à la rédaction de ce document. C'était une demande de leur part.

*** *** *** *** *** ***

10:30 - 11:15 : Diffuser pour mieux préserver : l'expérience de beQuali (banque d'enquêtes qualitatives)

Sarah Cadorel et Emilie Groshens [CDSP, IEP Paris - CNRS]

Site du projet beQuali : http://www.bequali.fr/bequali/
Retour d'expérience. Cette intervention présente le projet beQuali ou l'articulation entre la mise à disposition des données et les bénéfices qu'il est possible d'en tirer en terme de préservation.

Le périmètre du projet : enquêtes menées à partir de méthodes qualitatives.
Le site met à disposition des enquêtes en SHS et vise à offrir des outils d'enrichissement et d'exploitation de ces données.
>> Inspiration du projet : Qualidata (Royaume-Uni)

"Il s'agit de développer un outil expérimental qui rende visible tout le processus de l'analyse, qui permette de montrer aux chercheurs la diversité des démarches qualitatives et rende possible la compréhension des matériaux bruts."

- Collecte de tous les matériaux (projet de recherche, devis fiche de synthèse et transcription des entretien) qui sont classés selon un plan de classement déterminé.
- Le plan de classement reprend le processus de la recherche : préparation, collecte, analyse.

Les données d'entretiens ou d'observation ethnographique sont collectées au cours d'une interaction humaine, ce qui pose des questions juridiques.
beQuali est encadré par un dispositif juridique rédigé par un cabinet juridique extérieur et validé par le service juridique de Sciences Po et le CIL CNRS.
Deux types de contrats ont été définis : avec le chercheur et avec l'utilisateur.
L'objectif est principalement la protection des enquêtés.

Accompagnement des chercheurs dans la collecte (ces derniers ont besoin d'être rassurés)
Interface prochaine avec le portail Quetelet.
Association de compétences et d'interlocuteurs : archiviste, documentaliste, chargé d'études SHS et développeurs et réseau d'interlocuteurs (qu'on dirait facilitateurs), et un comité scientifique et technique.

Fonctionnalités :

- Site en accès restreint sur justification et une fois signé le contrat
- Pas d'accès pour les journalistes
- Procédures calquées sur celles mises en places par le portail Quetelet
- La documentation des corpus est en libre accès. Travail important sur les métadonnées.
- Utilisation de standards :
    - Standard DDI, développé pour les données SHS, bien adapté : http://fr.wikipedia.org/wiki/Data_Documentation_Initiative
    - EAD et Dublin Core
    - METS

Standards :

- Objectif d'interopérabilité, dialogue entre les machines (DDI permet par exemple de dialoguer avec le portail Quetelet)
- EAD pour dialoguer avec les services d'archives
- Travail de mapping entre les différents standards de métadonnées utilisées

>> Recours à des technologies open source afin que l'outil soit évolutif et qu'il ne soit pas étanche.
>> Travail sur la question des identifiants pérennes.
>> Offrir à l'utilisateur des données contextualisées.

Construction d'outils d'exploration (échelle de la base) :

- Critères géographiques (intéressant par exemple pour les enquêtes multi-terrains)
- Critères chronologiques
- Outil d'exploration à l'échelle du document : édition en XML TEI (encodage des tours de paroles) - tentative d'automatisation des procédures

Une enquête, en moyenne c'est 2000-3000 "documents", d'où la nécessité d'une sélection.
Évaluation pertinence / coût numérisation (pas suffisamment anticipé).
Travail de contrôle qualité important (et lourd).

Pérennisation des données et du dispositif
Intégrer le "circuit patrimonial"

*** *** *** *** *** ***

11:15 - 12:00 : Pérennisation et mise à disposition des données de l'Observatoire de recherche méditerranéen de l'environnement (OSU-OREME)

Juliette Fabre et Olivier Lobry [OSU-OREME - CNRS]

Équipe constituée d'un ingénieur de recherche et 2 ingénieurs d'études en informatique

L'objectif est de pérenniser les informations d'un point de vue technique
Spécificité des données d'observation : impossible à reproduire d'où la nécessité de les conserver.
Problématique : conservation, traitement et description des données

Problème d'interopérabilité (entre bases de données et entre logiciels) = problématique des droits notamment pour l'utilisation de croisement de données.
Besoin de visualisation et un manque de partage d'outils entre observatoires.
Ces données sont hétérogènes : format, type, volume, disciplines d'origine variés

En terme de méthodologie, essaient de mettre les données issues des observations dans des bases de données relationnelles. Permet de structurer les données par la mise en place d'un modèle. Travail fait avec les chercheurs qui ainsi se posent la question des relations entre les différentes entités qu'ils sont amenés à gérer.
Utilisation d'applications robustes offrant des garanties de pérennité. Également intérêt pour l'interrogation des données, plus simple que quand l'information est dispersée dans une série de fichiers.
Faire en sorte que toutes les problématiques qui sont abordées et toutes les expériences soient capitalisées.
Capitaliser les savoir-faire, les compétences… et les outils si c'est possible!
Mutualisation des développements et des outils web de visualisation = réutilisation

Comment cela fonctionne ?

- Identifier les contacts
- Identifier les données intéressantes dans le cadre de l'Observatoire, les types, les formats utilisés et les standards du domaine
- Choix des traitements : ce qui peut être automatisé, ce qui doit être corrigé.
- Identification des méthodes de diffusion et les conditions de partage des données (ouvert à tous, mise en place de protocoles d'accès).
- Au moment de la conception des bases de données, échange avec les chercheurs afin de bien comprendre leurs besoins

Pour la construction de base de données : PostgreSQL, PostGIS

Pour garantir la ré-utilisabilité et l'interopérabilité des données de ces bases, important de faire attention aux référentiels (vocabulaires contrôlés, ex : taxons), aux formats, aux standards

Quelles données sont stockées : différents niveaux de données

- données brutes ou niveau zéro
- données supérieures

Pour les données brutes, le public visé est celui des experts et pour les autres cela va des personnes intéressées au grand public selon l'enrichissement.
Toutes les données sont traitées en base de données.

Décrire : qui, pourquoi, où, comment, quoi ? + Enrichir en utilisant des standards de données, vocabulaires contrôlés ou thésaurus = cf. tableau (GEMET / WoRMS, NCBI….)

Linked Open Vocabularies (LOV) - http://lov.okfn.org/dataset/lov/
Ontologies sur l'observation scientifique, mais semble plus complexe à exploiter (?)
Standards simples à mettre en oeuvre mais plus de difficultés avec les ontologies.

Exemple d'utilisation de référentiels dans les bases de données :

- WoRMS (Registre mondial des espèces marines - http://www.marinespecies.org/aphia.php) = un identifiant par espèce marine, cela permet de récupérer l'identifiant et de lier les données de la base avec d'autres bases pour le croisement. Permet d'avoir une information plus riche et faire de la "recherche intelligente" sur les données précise l'intervenante.

Référentiel Carthage (réseau hydrographique, IGN)
Interopérabilité avec d'autres bdd cartographiques

Outils qui permettent de diffuser les descripteurs au public

Bilan au terme de trois années :
Freins : pendant que les producteurs de données sont sur le terrain ou pendant les périodes d'enseignement difficulté de dialoguer avec eux. Disponibilité et motivation fluctuent beaucoup.
Les degrés de motivation varient aussi selon les personnes : certains respectent juste une obligation d'autres souhaitent valoriser leurs travaux ou disposer d'outils, avoir accès à des données.
Difficulté de retrouver les données et d'avoir des formats homogènes.

Parfois difficultés pour se comprendre entre les différents métiers : chercheurs, informaticiens. Gros effort à faire de ce côté.

Difficulté liée à l'apprentissage d'un sujet, à la bonne compréhension d'une thématique.

Temps moyen de constitution d'un système d'information : 2 ans.

Les clés :
Motiver les producteurs : outils, valorisation.
Récupérer les données au plus vite pour aider à formater les données des chercheurs.
Capitaliser les savoir-faire.

Question sur les données brutes :
voir : https://halshs.archives-ouvertes.fr/halshs-00990771/document

*** *** *** *** *** ***

13:30 - 14:15 : Problématique du devenir des données au Centre de Calcul de l'IN2P3

Pascal Calvat [Centre de Calcul de l'IN2P3 - CNRS]

"Avalanche numérique dans tous les domaines" > les instruments de recherche produisent de plus en plus de données : 15 Po de données brutes par an répartis sur une grille de calcul mondiale.
Astrophysique surtout, qui produit énormément de données : 150 Po de données sur 15 ans, avec l'arrivée notamment d'un microscope LSST (Chili).
En biologie : arrivée des séquenceurs haut débit

// BnF : production de 100 To par an, problématique d'archivage pérenne.

Pour y faire face, les chercheurs doivent avoir accès à des ressources informatiques mutualisées : laboratoire, centre de calcul, grille de calcul (> plusieurs centres de calcul)
La gestion des données numériques devient un point incontournable pour la conduite de projet scientifique

Fait d'utiliser des ressources mutualisées oblige à travailler en équipe….

Importance de la conservation à long terme : si on ne peut pas relire les données en cours de projet cela pose problème.
Aussi problème pour des projets plus petits où le partage des données n'a pas été pris en compte.
Attention aussi à la perte de données uniques.

Beaucoup de données orphelines : exemple du thésard qui produit des données puis il part et laisse ses données.

IN2P3 a commencé à réfléchir à la mise en place d'un plan de gestion de données
Intensifier la collaboration entre chercheurs (en accédant à des données structurées en ligne)
Souligner coût de la reproduction des données (impensable), et certaines sont trop rares (impossible)

Centre de calcul de l'IN2P3 installé à Lyon depuis 1986 (physique des particules notamment), ouverture à d'autres champs disciplinaires

Centre de Calcul stocke sur de la bande magnétique et sur du disque.
20 000 coeurs de calcul pour l'analyse des données et les simulations

Difficulté de passer d'une technologie à une autre : prend du temps car même si un nouvelle technologie apparaît, l'ancienne reste en vigueur.

Type de formats très variés : dès qu'on passe à une normalisation, on considère qu'il y a déjà un niveau d'analyse.

Faut-il conserver tout et comment gérer efficacement une telle diversité de données?
Aujourd'hui ils ont décidé de tout garder, mais ne savent pas toujours ce qu'il y a dedans.

Dans l'idéal il faudrait que chaque projet formalise un plan de gestion des données.
Que va-t-on y trouver?

- description des données
- métadonnées
- description du cycle de vie des données y compris après le projet
- détails de la politique associée aux données
- aspects budgétaires

Un peu comme un cahier de manip qui décrirait les paramètres contextuels dans lesquels les données ont été collectées

Pas de campagne systématique d'effacement des données. C'est au responsable des données de prendre les décisions. Et pas de sauvegarde sauf si c'est demandé, mais en fait pas de demande car il est difficile de comprendre ce qui est important.

Dans le plan de gestion : désigner un responsable ou correspondant du projet en question (identifier notamment quelles sont les données critiques à conserver)

Perspective: réaliser un inventaire des données stockées au centre de calcul
Faire un point chaque année entre les utilisateurs et les responsables des données
Inventaire aujourd'hui : 40 Po

Cet inventaire est à destination des ingénieurs du centre pour avoir une vue détaillée des données et des responsables des projets. Cet inventaire a été mis en place sur 2 ans.

Rencontre avec les chercheurs, là où ils travaillent, dans leur laboratoire, pour appréhender davantage leurs problématiques et leurs besoins et les sensibiliser à la gestion de données..
Téléphone marche plutôt bien, par mail c'est plus laborieux

Devenir des données au centre de calcul : préférable que les données soient supprimées en fin de projet (pas si évident car il faut savoir qui a les droits sur les données pour décider de la suppression et cela peut pendre des mois).
La réflexion sur la pérennisation des données et leur mise à disposition doit être faire au départ.

Gestion des données devient indispensable. DMP doit être fait en amont du projet. Peut prendre la forme d'une simple description de workflow.

Le responsable des données qui interagit avec l'IN2P3 est forcément quelqu'un qui a des notions d'informatique.

Importance d'anticiper le départ d'un gestionnaire de données dans les plan de gestion des données. Il faut une passation des savoirs par exemple entre deux thésards qui se succèdent sur un projet.

L'IN2P3 conserve les données mais ne fournit pas les services permettant aux données d'être toujours lisibles, et ne fait pas évoluer les formats des données enregistrées il y a quelques années. C'est aux producteurs de s'en occuper. Le CINES n'est pas forcément approprié pour archiver les données en question car il archive à partir de formats standard. Ici il y a des données de formats très spécifiques et surtout un volume énorme.

Le stockage dépend vraiment des budgets !!!

*** *** *** *** *** ***

14:15 - 15:00 : Illustration des questions de dialogue et coordination entre les acteurs à travers des projets de gestion de données

Christine Plumejeaud-Perreau [LIENSs, Université de la Rochelle - CNRS]
Responsable de la plateforme DISA.

Dans un projet de gestion de données le plus important est le dialogue. Il faut anticiper les questions de gestion de projet et de communication.
Les différents acteurs n'ont pas les mêmes notions de l'utilité de la gestion de données ni les mêmes connaissances.

1er exemple sur les données de l'environnement : très fort enjeu sur la conservation, l'interopérabilité et valorisation auprès du grand public.
Problème de gestion de projet et de synchronisation...en fait pas d'anticipation...problème de financement.
Pas de documentation claire sur les logiciels et les données.

Intérêt de l'utilisation d'un serveur centralisé qui permet de synchroniser les équipes qui sont sur des sites différents. Permet aussi d'éviter que des développeurs développent dans leur coin des extensions qui ne sont pas compatibles. Travail de pédagogie nécessaire.

2ème exemple : enquête population d'un territoire littoral
Problème du pilotage multiple recherche, communes et entreprises? Pas d'entente sur les méthodes et notamment d'analyse… et utilisation d'outils différents qui met en jeu l'interopérabilté...
Enquête courrier, téléphonique et focus groupe.
Données protégées par la CNIL et donc pas de possibilité d'envoi par mail.
Ces données sont recueillies par des étudiants et également projet d'harmonisation de ces fichiers.
Recruter quelqu'un qui gère le relationnel et travaille en groupe de travail

Problèmes rencontrés :
Hétérogénéité des modes d'acquisition donc importance de règles communes.
Questions liées à la confidentialité des données
Gestion des mises à jour
Dispersion géographique : mettre en place des outils pour travailler ensemble.

Question : peut-on rester au niveau des "Proof of Concept" quand on cherche à fabriquer des corpus de données, stables, homogènes, et exploitables ?
Grand écart entre innovation et production.
Question : est-ce le métier des chercheurs que de travailler sur les données de la recherche alors que rien n'est publiable derrière ? La recherche, c'est plutôt l'innovation. Pour les données de la recherche, ce sont d'autres métiers, plutôt des métiers de support. Une autre intervenante plaide le fait que ce travail en amont sur les données, leur description, leur recueil homogène, permet aussi une meilleure qualité des données de la recherche, et donc des productions de recherche. Reste à faire la démonstration du gain de ce travail pour que la recherche reste innovante.

Idée que certains projets ne devraient pas être menés si les moyens (humains et techniques) ne sont pas disponibles.

Intervention de la salle : dans certains contrats d'édition, mentionné qu'il faut mettre les données à disposition pendant plusieurs années comme preuve de ce qui est démontré dans article.

*** *** *** *** *** ***

15:30 - 16:15 : Indigeo, une infrastructure scientifique de données et d'informations géospatialisées sur l'environnement

Mathias Rouan [LETG - Brest Géomer, CNRS]

Présente le contexte de création de l'infrastruture Indigeo avec son ancêtre Menir. Pendant cette période, prise de conscience de l'utilité du catalogage, développement de l'expertise, évolution de la structure.
Indigeo regroupe dans un visualiseur des données provenant de différentes sources et donne accès à leurs métadonnées. Pas de possibilités de télécharger des données venant d'autres plateformes par Indigeo.
Puis 2007 directive INSPIRE et mise en place de différents sites.
juin 2013 : ouverture d'Indigeo en ligne.

http://www.indigeo.fr/
http://www-iuem.univ-brest.fr/fr/observation/indigeo/osuiuem.jpg/view

Infrastructure de données spatiales avec catalogue de métadonnées et fait pour le stockage et la diffusion de l'information.
Georchestra : infrastructure pour être interopérable, et moissonner différents portails http://www.georchestra.org/fr/
Besoin en visualisation donc outil qu'ils développent eux-mêmes GeoCMS (disponible sur Github) = avec accès à un catalogue de données géoréférencées.
Accès aux données du BRGM, Ifremer, IGN ou d'autres données présentes sur d'autres portails d'institutions
Catalogue : possibilité de faire une recherche multisource.
Navigation dans un jeu de données spatio-temporelles, accès au jeu de données également (directive Inspire)

Outils indépendants et libres = geonetwork et geoserver
Interopérabilité grâce à des standards

Metadata party : aide à publier les données. Une organisée en 2013. Prévoient de le refaire.

Aller vers standard WPS (Web Processing service) : mettre à disposition à travers une infrastructure des traitements, des informations.

Deux facettes : catalogage de l'information et… visualisation des données
Il faut que les chercheurs soient intéressés par la valorisation des données
Évocation de peurs des chercheurs : comment leurs données vont être utilisées?

*** *** *** *** *** ***

16:15 - 17:00 : Accompagnement actif des chercheurs à la gestion et au partage des données de la recherche

Marie-Christine Jacquemot-Perbal [INIST-CNRS] et Thierry Beguiristain [LIEC Université de Lorraine - CNRS]

Adaptation aux besoins des chercheurs = gestion et valorisation des données de la recherche
Expérimentation Inist-OTElo = partenariat autour des données de la recherche
Rôle des documentalistes en fonction du cycle de vie des données

Inculquer les bonnes pratiques aux chercheurs pour permettre l'exploitation, la curation des données.
Encourager les chercheurs à publier dans des data journal.

Quelles connaissances et compétences à développer pour la gestion des données?
Passer du rôle de documentaliste scientifique à celui de **data librarian** : méthodes actives de formation et participation active dans des projets. Méthode constructiviste par questionnement.
Etape fondamentale de la sensibilisation = module de sensibilisation sur le site de l'Inist = http://www.inist.fr/donnees/index.html

http://www.dcc.ac.uk/ : data curation center du JISC. Cité comme exemple ici car premier du genre mais on en trouve de plus en plus.

Dans certaines disciplines, on commence d'abord par la gestion des matériels d'étude, car les chercheurs ne sont pas prêts à ce qu'on gère leurs données.

La stratégie d'OTELo est passée par de la sensibilisation, avec une évaluation par questionnaire. Puis en fonction des résultats du questionnaire, un "cookbook" fait des recommandations sur ce qu'il faut mettre en oeuvre en fonction du niveau de maturité.
Résultat : peu de différences entre les disciplines. On est au niveau basique : phase de sensibilisation.

Inist infrastructure de soutien à la recherche. Former les chercheurs dès le master.

De nouveau affirmation que la réflexion sur les données doit se faire en amont.
Encore une fois : partager des données oui mais avec un contrôle de la part du chercheur
Si les chercheurs s'investissent, il faut qu'ils en tirent une plus-value, qu'ils en retirent quelque chose : qualité des données stockées, analyse multidisciplinaire des données, valorisation de la base.

S'appuyer sur les usages des chercheurs pour faire évoluer leurs pratiques et non pas leur imposer quelque chose qui conduirait forcément à l'échec.

Les différents acteurs permettant la réutilisabilité des données :
- les chercheurs (intelligibilité des données),
- les informaticiens (accessibilité),
- les documentalistes (interopérabilité).

A partir de l'exemple de la base mise en place pendant ANR MultipolSite (ANR CESA 2008), diagnostic des pratiques des chercheurs avant la phase des recommandations qui s'appuient sur ces expériences pour apporter ces bonnes pratiques.
Faire passer des bonnes pratiques…
Recommandations : structuration des données, documentation du contexte de production, normalisation/standardisation des métadonnées, convention de nommage des fichiers, formats des fichiers, redéfinition collective des fonctionnalités attendues.

Amorcer une réflexion autour du DMP pour se servir d'un projet terminé qui pourra servir d'exemple pour la suite.
Il est important de réagir vite car les nouveaux projets sont déjà en route (ne pas faire les mêmes erreurs).

*** *** *** *** *** ***

17:00 : Clôture et synthèse de la journée par Jean-Luc Minel

Une journée très riche avec des projets concrets...que ce soit en sciences dures ou en SHS nous sommes bien dans le même questionnement sur la gestion des données.

La posture des documentalistes, professionnels IST : passer de “travailler ***pour*** les chercheurs” à “travailler ***avec*** les chercheurs”. Envisager davantage le partenariat, les chercheurs comme les professionnels IST sont au service de la recherche - mais véritable tradition des documentalistes d'être au service des chercheurs… >> Cela doit changer et se positionner comme des partenaires véritables.

Question de Jean-Luc Minel assez provocatrice sur notre rôle qui devrait être plus actif et de type partenariat avec les chercheurs = H2020 est une occasion unique pour les documentalistes d’être indispensables aux chercheurs pour répondre au DMP dans leurs projets européens...c’est un levier qu’il faut utiliser pour “s’imposer”...

Il faut se positionner au plus près de la recherche, la gestion des données commence dès le début d’un programme de recherche.
Il faut faire valoir l’importance des compétences documentaires dans la gestion des données que ce soit au niveau des référentiels, des normes ou des métadonnées. Mais nous avons vu aussi dans les différentes interventions qu’il fallait se former également au niveau technique (XML, TEI, etc.)

La journée a mis en lumière des projets où les informaticiens travaillent seuls et le dialogue a permis de montrer à quel point les compétences documentaires leur seraient utiles. Il est fondamental de mettre en place une véritable collaboration avec les informaticiens mais aussi avec d’autres ingénieurs accompagnant la recherche… Les documentalistes doivent travailler avec les informaticiens et les chercheurs !
On peut noter également que plusieurs interventions étaient faites en collaboration : documentalistes, bibliothécaires et archivistes..

Point non abordé au cours de la journée : quel statut juridique pour les données ?
Groupe de travail juridique dans le réseau base de données du CNRS

Surprise : la question du web de données et des ontologies n’a pas été abordée.
Réponse : le thème de la journée se positionnait en amont, étape suivante : le web de données

On a pu voir une tentation dans la présentation des projets d'insister sur la partie conservation et archivage ce qui n'est pas forcément ce qui compte le plus pour les chercheurs. Pour négocier avec les chercheurs, il faut leur montrer que la plus-value de la gestion des données est *pour maintenant*, leur montrer que cela compte pour la visibilité de leurs recherches aujourd’hui.