ANF MATHRICE 2018
Problématiques de stockage et protection des données dans le laboratoire

Philippe Depouilly - Joël Marchand

04 décembre 2018

mathrice

Catégoriser, cartographier les données dans le laboratoire
- Etat des lieux de la donnée dans un laboratoire aujourd'hui
- Définir le périmètre de l'ASR

Plan de sauvegarde de la donnée

Expériences personnelles, retours d'expérience, et conseils

Des élements de réponse face à ces risques

Eléments de conclusion

mathrice

Dans le cas particulier du laboratoire de mathématiques

- Hier, peu de données et une typologie simple, gérée par l'ASR :
  - la messagerie électronique (mbox)
  - les articles scientifiques (essentiellement LaTeX)
  - les sources d'un logiciel
  - les résultats (sur LE serveur de calcul)
- Aujourd'hui, la plupart des données échappe à l'ASR
  - la messagerie est externalisée
  - autour du calcul : SourceSup, GitHub, Mésocentres, GENCI, etc.
  - la donnée semi-professionnelle/personnelle (BYOD) est massive
  - les documents d'enseignement : supports, examens, rendus des étudiants, enseignements en ligne
  - les dossiers de conventions et de projets : COMUE, ANR, ERC, etc.
  - la multiplication des hébergements cloud : données, sites Web, forges, services IAAS, PAAS et SAAS

mathrice

Ce mouvement de dispersion technologique, typologique et organisationnelle semble inéluctable

- dispositifs techniques : ordinateur, tablette et smartphone professionnel et personnel, serveurs de nature variée et souvent inconnue
- donnée non atomique : relation forte entre la donnée, un ou des logiciels, des outils de gestion de versions, des bases de données, des méta-données, des plateformes
- disponibilité et intégrité : sécurisation mise en oeuvre hétérogène et souvent inconnue
- échelles : individu, laboratoire, établissement, régional, national, européen, communautaire, projets

Statut de la donnée

- Professionnelle individuelle
- Professionnelle interne à l'unité
- Professionnelle transverse (collaboratif)
- Institutionnelle - unité - établissement
- Personnelle sur poste professionnel... ou non...

Différentes phases de la vie de la donnée peuvent être identifiées, même si elles varient suivant les usages et les disciplines

- Collecte brute
- Organisation collaborative
- Filtrages et post-traitements
- Documentation, indexation
- Publication et diffusion
- Archivage éventuel

Lors d'une panne, d'une perte, d'un accident de toutes natures, l'ASR est en première ligne

- Il se pose alors les questions suivantes... ou non...
- Quelle donnée précisément ?
- Est-elle chez moi ?
- Si oui
  - Sous quelle forme ? (working-copy svn/git, copie DropBox, etc)
  - Comment et quelle version puis-je restaurer ?
- Si elle est distante
  - Comment fonctionne le service distant ? (jamais utilisé DropBox...)
  - Comment puis-je contacter le service distant ?
  - Quel processus, interface, processus va-t-il me proposer ?
- Ai-je bien anticipé ?
- L'hébergeur a-t-il bien anticipé ?

mathrice

Lors d'une panne, d'une perte, d'un accident, l'ASR est en première ligne

- Quelle urgence ?
- Quelle volume ? quel impact sur mes ressources ?
- Quel processus pour y accéder : contact, méthode, protocole ?

Dans le pire des cas, il peut devoir récupérer très vite ou un volume importante

- Dois-je prévoir de savoir récupérer tout très vite ?
- Dois-je prévoir le stockage et la sauvegarde d'un volume inconnu de données, car hébergées à l'extérieur ?

Il est donc indispensable de définir un plan de sauvegarde de la donnée

- Connaître les usages
- Sensibiliser sur les risques
  - physiques
  - sur les supports techniques
  - liés aux logiciels
  - liés aux processus
  - liés aux organisations
  - liés à l'externalisation : DropBox, GoogleDrive, etc,
- Identifier les alternatives techniques et organisationnelles
- Autant que possible
  - être vigilant continûment sur toute la chaîne
  - notamment au niveau humain : soi-même, les utilisateurs, les tiers
  - opérer des restaurations
  - prendre le temps de faire un retour d'expérience de chaque incident

mathrice

- Classifier la donnée
  - messagerie - fichier - base de données
  - scientifique - pédagogique - administrative
  - donnée reproductible facilement - difficilement - pas du tout
  - sensible - non
  - à durée de vie courte - longue - patrimoniale
  - ...
- Associer une localisation de la donnée et sous quelle forme
  - locale totalement
  - locale partiellement (copie imparfaite ou version antérieure)
  - distante totalement
  - format brut
  - format dégradé
  - ...
- Les réponses à ces questions permettent au moins face à un événement donné
  - de ne pas être pris de court
  - d'avoir une meilleure idée de sa capacité ou non à faire face

Expériences personnelles, retours d'expérience, et conseils

- /bin/rm -fr / *.log
  - Est-on organisé pour assumer ses propres erreurs ? celles de ses utilisateurs ?
  - A chaque erreur humaine, surtout les siennes, faire l'effort d'en tirer les conséquences organisationnelles
- tar cvf /dev/exabyte dir1 dir2 dir3
  - Est-on organisé pour garantir l'exhaustivité de son processus ?
  - Privilégier l'exhaustivité par construction - être systématique - pouvoir vérifier la complétude
- Réplication et synchronisation
  - Est-on organisé pour remonter le temps ? jusqu'où aller en la matière ? combien cela coûte-t-il au final ?
  - Ne pas confondre disponibilité et sécurité - quelle profondeur temporelle est justifiée ?
  - Compter le nombre de copies d'une même donnée et faire des choix appropriés

- Grande tempête de décembre 1999
  - Est-on organisé pour assumer les risques physiques : électricité, incendie, eau ?
  - Ces risques sont permanents - aucune réponse technique ponctuelle n'est fiable
  - Seule réponse possible = découplage spatial
- Baie RAID perdue au CEA
  - Est-on organisé pour être averti des incidents de sécurité ?
  - Systématisation de la supervision depuis le matériel
  - Vérification périodique : du visuel jusqu'aux journaux
- Thésard à l'IMJ en 2004 à quelques heures du dépôt de sa thèse
  - Est-on organisé pour assumer l'urgence ?
  - Donner le maximum d'autonomie à l'usager pour assumer lui-même ses accidents et erreurs
- Baie RAID perdue à l'Observatoire de Paris en 2007
  - Est-on organisé pour assumer seul tous les risques ?
  - Ne pas jouer perso et utiliser les offres extérieures

mathrice

- Vol ciblé au CNRS - post-doc UPEM en 2013
  - Est-on organisé pour assumer les risques de vols ?
  - Ces risques sont permanents et sont accrus par l'aspect BYOD
  - Seule réponse possible = découplage spatial PERMANENT
- Accident psychiatrique à l'IMJ en 2004 - mouvement étudiant de 2018
  - Est-on organisé pour assumer les risques sociaux (vandalisme, sabotage) ?
  - Ces risques sont permanents et ne devraient pas baisser
  - Seule réponse possible = découplage spatial PERMANENT
- Bug sur les firmwares - entreprise attaquée sur tous ses MBR
  - Est-on organisé pour assumer la perte complète d'un ensemble technique ?
  - Autant que possible : avoir plusieurs supports distincts

- Données sur supports externes disparus ou ordinateurs obsolètes
  - Est-on organisé pour faire face à l'obsolescence technique ?
  - Eliminer les vieux supports avant qu'il ne soit trop tard
- Documents Word 2 ou données ou bases de données liées à des versions très anciennes
  - Est-on organisé pour faire face à l'obsolescence des formats ?
  - Problème encore plus difficile, car nécessite une cartographie fine
  - Autant que possible, forcer l'évolution logicielle autour des données
- Jeu de données dans iRods en 2018
  - Est-on organisé pour relier les données à des personnes, des usages ?
  - Besoin d'avoir des fiches de renseignements actualisés sur chaque jeu de données
  - Sinon, risque d'encombrement et de charges inutiles
  - Quel processus de suppression ?

mathrice

Des élements de réponse face à ces risques

- Principes à l'esprit et mis en oeuvre autant que possible
  - Systématisation du découplage spatial et organisationnel
  - Tout n'est pas dans le même espace spatial, électrique, technique, humain
  - Soin et choix explicite sur la profondeur temporelle
  - Auto-vigilance récurrente
  - Organiser, ranger, lister, compter régulièrement : pour mieux et plus facilement savoir et agir
- Espaces de stockage pour le calcul à l'Observatoire de Paris
  - Différentiation suivant l'importance, la nature, le volume, le coût
  - Compromis social, organisationnel, financier

mathrice

Stockage sécurisé distribué avec Active-Circle à Huma-Num

- Données très peu accédées, mais dans l'ensemble uniques et non reproductibles
- Choix radical d'un outil adapté centrée sur la sécurité, et pas la disponibilité
- Tout en maitrisant explicitement et dynamiquement les coûts pour chaque jeu de données
- Réplication, historisation, disques et bandes
- Verrouillage des médias, jeux de données WORM
- Vérification des empreintes, journalisation des opérations
- Fiche descriptive pour chaque jeu de données
- Inventaire régulier
- Usage de l'outil FACILE du CINES

mathrice

Que faire vis à vis des données externes à l'unité ?

- Faut-il anticiper une éventuelle défaillance des tiers ?
- Notamment pour les données finalisées ou critiques et peu volumineuses : messagerie, documents administratifs, publications, logiciels
- Inversement besoin de faire confiance pour les volumes importants : demander de la visibilité sur l'administration du système et si possible y participer autant que possible pour acquérir de la confiance et de l'autonomie

mathrice

La sécurisation des données est une question multi-couches

- environnement physique
- technologies matérielles
- systèmes d'exploitation et systèmes de fichiers
- protocoles et intergiciels
- logiciels et applications clientes
- processus et gestion au sens documentataire des données
- utilisateurs et leurs usages
- projets scientifiques

L'ASR est invité à prendre en compte et être vigilan autant que possible sur ces différentes couches

mathrice