MINISTÈRE DE L'EUROPE
ET DES AFFAIRES ÉTRANGÈRES

RETOUR D'EXPÉRIENCE DE LA COLLECTE DES ARCHIVES ÉLECTRONIQUES DE LA COP21

23/11/2017

Isabelle Josse – Chef de projet MOA Saphir

SOMMAIRE

- Collecte des dossiers de la COP21 : un cas particulier pour le ministère
- Une première expérience de traitement d'archives électroniques (bureautique, courriels)
- De nouvelles perspectives avec la mise en place d'un Système d'Archivage Electronique

COLLECTE DES DOSSIERS DE LA COP21 : UN CAS PARTICULIER POUR LE MINISTÈRE

UN CADRE HABITUEL DE « *RECORD MANAGEMENT* » AU MEAE

- Un réseau de Centres d'Archives et de Documentation (CAD) en relation avec les référents du pôle collecte de la direction des archives du MEAE
  - Elaboration avec la direction des archives de plan de classement pour chaque service producteur
  - Classement au fil de l'eau des documents papiers et électroniques
  - Collecte et gestion de l'information au sein des entités du MEAE

- Des outils et des procédures pour la collecte des archives électroniques
  - Une procédure normale de collecte de l'information numérique (documents bureautiques) ou circulant au sein du MEAE et avec ses partenaires basée sur le portail Diplomatie (outil de production de la correspondance diplomatique et outil d'enregistrement)
  - Des procédures exceptionnelles pour la collecte de cabinets ou la collecte de données classées en collection (cartes, photos, vidéos, rapports…) nécessitant une intervention d'un archiviste.

COP21 : UNE COLLECTE SUR 6 MOIS AUPRÈS DE 2 ÉQUIPES INTERMINISTÉRIELLES

1/ L'équipe négociation composée d'experts sous la responsabilité de Laurence Tubiana, ambassadrice chargée des négociations sur les changements climatiques.

- ⇒ Collecte non exhaustive : organisée sur la base du volontariat sous forme d'entretien de 1 à 2h (recueil d'information de contexte)
- ⇒ Collecte disparate et peu ordonnée : l'équipe interministérielle ne disposait ni espace bureautique, ni messagerie commune et dont chaque membre avait la charge d'un champ d'expertise ou d'une zone géographique spécifique

COP21 : UNE COLLECTE SUR 6 MOIS AUPRÈS DE 2 ÉQUIPES INTERMINISTÉRIELLES

2/ Le secrétariat général chargé de la préparation et de l'organisation de la conférence sous la responsabilité de M. Pierre-Henri Guignard, ministre plénipotentiaire

- ⇒ Collecte coordonnée : une réunion de coordination initiale et un suivi global assuré la chef de cabinet tout le temps de la collecte
- ⇒ Elaboration préalable d'un plan de classement permettant aux rédacteurs de trier leurs documents papier et électroniques à partir d'un cadre identique

ARCHIVAGE DES DONNÉES DU WEB PAR LA BNF

- Demande de confirmation auprès de la Bibliothèque nationale de France de l'archivage des données du web relative à la conférence.
  - ⇒ Dans le cadre du dépôt légal du web la BnF a collecté le site www.cop21.gouv.fr à plusieurs reprises entre février 2015 et mars 2016.
  - ⇒ Les archives les plus complètes sont celles de décembre 2015, avec plus 8 300 URL collectées et celles de novembre 2015, représentant plus de 10 700 URL collectées.

COMPOSITION DES FONDS COLLECTÉS

		Equipe de négociation	Equipe du Secrétariat général	Total
Nombre de personnes versantes		21	9	**30**
Archives papier	Nombre de cartons	1	35	**36**
Archives électroniques	Poids (Mo)	46 800	78 100	**12 4900**
	Nombre de fichiers	46 091	18 009	**64 100**
	Nombre de dossiers	7 676	2 949	**10 625**
	Nombre de boîtes de messageries électroniques	13	4	**17**

SECRÉTARIAT GÉNÉRAL

- 2 949 dossiers
- Formats de fichiers rencontrés:
  - eml, pdf, odt, ods, xls, doc, jpg, db, zip, msg, rtf, pst, png, HTML, css, js, ppt, gif, mp4, pps, msf, dwg, ink, bak, txt, dwl, dsd, tmp, bmp, mov, ai, mxf, eps, xlsb, tiff (soit 35 formats différents)
- Présence de fichiers corrompus et sans extensions
- Présence de fichiers cachés
- Nombre de fichiers temporaires : 1119
- Nombre de doublons : 1696 (soit 4,54 Go)
- Nombre de dossiers/sous-dossiers vides : 115

Données issues du mémoire de Cédric Lefebvre

EQUIPE NÉGOCIATION

- 7 676 dossiers
- Formats de fichiers rencontrés:
  - eml, pdf, odt, ods, xls, doc, jpg, db, zip, msg, rtf, pst, png, HTML, css, js, ppt, gif, mp4, pps, msf, ink, txt, tmp, bmp, mov, ai, eps, xlsm, vdf (soit 30 formats différents)
- Présence de fichiers corrompus et sans extensions
- Présence de fichiers cachés
- Nombre de fichiers temporaires : 1 970
- Nombre de doublons : 10239 (soit 2,38 Go)
- Nombre de dossiers/sous-dossiers vides : 693

Données issues du mémoire de Cédric Lefebvre

PREMIERS CONSTATS

- Peu d'appréhension du fonds lors de la collecte des archives numériques (vs la collecte de papier)
  - Collecte de masse en un temps très court , sans possibilité de prise de connaissance et de tri des contenus.

- Intérêt des informations collectées incontestable mais traitement en vue de leur communication encore difficile à estimer
  - Volume important de données, compliqué par des niveaux d'arborescence parfois très profonds
  - Présence de documents en plusieurs exemplaires, non finalisés, souvent non datés, mélange de versions intermédiaires ou finales, de brouillons ou de versions validées
  - Des documents « de travail » dont les métadonnées ne permettent pas de garantir leur authenticité (auteur, date de création, date de modification, statut…)

UNE PREMIÈRE EXPÉRIENCE DE TRAITEMENT D'ARCHIVES ÉLECTRONIQUES

OBJECTIFS ET ENJEUX

1/ Campagne de tests des outils d'aide au traitement de collecte numérique

- Prise en main, évaluation…

2/ Traitement appliqué au sous-fonds de l'équipe interministérielle de négociation de la COP21

- Traitement des documents bureautiques
- Traitement des boîtes de messagerie électronique

3/ Alimentation des réflexions autour des problématiques liées aux archives numériques

- Elaboration de procédures de traitement et dotation d'outils adaptés

1/ CAMPAGNE DE TESTS : SUR LA BASE DE LA LISTE IDENTIFIÉE PAR LE SIAF*

Fonctionnalité(s)	Outil identifié par le SIAF*	Système d'exploitation	Testé	Outil(s) alternatif(s)	Système d'exploitation	Testé
Récolement « technique »	/	/	/	**WinDirStat**	Windows	Oui
	/	/	/	**DiskBoss**	Windows	Oui
Détection et décompression des dossiers compressés	/	/	/	**Extract Now**	Windows	Oui
Détection et suppression des dossiers vides	/	/	/	**Remove Empty Directories**	Windows	Oui
Capture, récolement et calcul d'empreinte	**DataAccessioner**	Windows	Oui	/	/	/
Détection des doublons	FSlint Janitor	Linux	Non	**Duplicate File Finder**	Windows	Oui
				DupeGuru		
Analyse anti-virus	ClamTK	Linux	Non	/	/	/
Formatage et renommage	**VRenamer**	Windows	Oui	/	/	/
Classement / empaquetage / Identification de format	**DocuTeam Packer**	Windows	Oui	/	/	/
Extraction de métadonnées	**PyExifTool GUI**	Windows	Non	NLNZ Metadata Extraction Tool	Windows	Oui
	MediaInfo	Windows	Non			

** Cycle de publications sur la thématique des outils de pré-versement d'archives numériques :* http://siaf.hypotheses.org/tag/pre-versement

1/ CAMPAGNE DE TESTS : CONSTITUTION D'UN JEU DE TESTS REPRÉSENTATIF

Echantillon constitué de plusieurs dossiers et sous-dossiers du sous-fonds de l'équipe négociation

	Dossier_Test1
Taille	1,24 Go
Nombre de sous-dossiers	804
Nombre de documents	1695
Formats rencontrés	• Bureautique : .odt ; .ods ; .doc ; .docx ; .xls ; .xlsx ; .ppt ; .pptx ; .pdf ; .txt • Mail : .pst ; .eml • Image : .jpg ; .png ; .bmp • Vidéo : .mp4 • Formats compressés : .zip • Autres formats : .tmp ; .db ; .ink ; .ini
Niveau d'arborescence	Très profond (10 niveaux au total)
Notes	Plusieurs fichiers sans extension et impossible à lire, y compris des fichiers

Données issues du mémoire de Cédric Lefebvre

2/ TRAITEMENT DU SOUS-FONDS DE L'ÉQUIPE NÉGOCIATION

- Les questions qui se sont posées
  - Comment gérer l'hybridité du fonds ?
  - Comment traiter la redondance et la non-exhaustivité des dossiers des versements par rédacteurs ?
  - Comment traiter les boîtes de messagerie électronique ?
  - Les documents bureautiques valides ne sont-ils pas présents en doublon dans les boites de messagerie des rédacteurs ?
  - Doit-on aller jusqu'au traitement à la pièce ?

2/ TRAITEMENT DU FONDS DE L'ÉQUIPE NÉGOCIATION

Traitement des documents bureautiques : Etat des lieux

- Appréhender la composition du fonds (en termes de structure, de volumétrie et de formats) avant tout type de traitement

- Approche technique (DiskBoss):recensement par typologie
  - Regroupement des fichiers en grandes catégories artificielles (Internet, Bureautique, Multimédias, etc.)
  - Regroupement des fichiers par extensions : permet de repérer les fichiers systèmes, temporaires et corrompus

- Approche intellectuelle (TreeSize Professional):vision d'ensemble de la structure interne d'un fonds
  - Retranscription de la structure interne du fonds en une arborescence dynamique avec possibilité de développer ou de réduire tout ou partie de l'arborescence
  - Affichage d'informations comme la longueur des chemins d'accès des fichiers

Données issues du mémoire de Cédric Lefebvre

2/ TRAITEMENT DU SOUS-FONDS DE L'ÉQUIPE NÉGOCIATION

Traitement des documents bureautiques : Nettoyage du fonds

- Détection et suppression des fichiers systèmes et des fichiers temporaires
  - Fichiers désignés sous le terme anglais de « junk files »,
  - Fichiers généralement cachés et inexploitables

  ⇒ **Total : 2 242 sur 46 091 fichiers (4,9 %)**

- Détection et suppression des dossiers et sous-dossiers vides
  - Nombre important au sein du sous-fonds
  - Complexifie davantage la structure interne du sous-fonds mais dans certains cas porteurs d'informations

  => Total : 693 sur 7676 dossiers, soit 9 %

Données issues du mémoire de Cédric Lefebvre

2/ TRAITEMENT DU SOUS-FONDS DE L'ÉQUIPE NÉGOCIATION

Traitement des documents bureautiques : Outils d'aide au classement

- Détection et décompression des fichiers compressés
  - Complexifie la structure interne du sous-fonds
  - Perte d'informations importantes pour la suite du classement

  => traitement au « cas par cas » pendant le classement

- Détection des doublons
  - Rassembler les collections de documents pour en faire des dossiers exhaustifs et éviter la disparité et la redondance des éléments
  - Possible perte d'informations pertinentes pour le classement si suppression

  => opérations réalisées tout au long du classement en comparant des parties ayant déjà fait l'objet d'un classement

Données issues du mémoire de Cédric Lefebvre

2/ TRAITEMENT DU SOUS-FONDS DE L'ÉQUIPE NÉGOCIATION

- Composition des dossiers classés (COPIL et CCNUCC)

	Fichiers	Courriels (.eml)	Dossiers	Poids (Mo)
Dossier COPIL	975	656	222	197
Dossier CCNUCC	6252	25280	770	2390
TOTAL	7227	25936	992	2587

- Comparaison avec le reste du sous-fonds

	Nombre de fichiers	Nombre de doublons	Poids des doublons
COPIL	1631 (dont 656 courriels)	1782	Environ 300 Mo
CCNUCC	31352 (dont 25280 courriels)	8789	Environ 2000 Mo
TOTAL	32983	10571 (soit 22,9 %)	Environ 2300 Mo

Données issues du mémoire de Cédric Lefebvre

2/ TRAITEMENT DU SOUS-FONDS DE L'ÉQUIPE NÉGOCIATION

Traitement des documents bureautiques : normalisation des noms de fichiers

- Normalisation ou renommage ?
  - renommage va à l'encontre du principe du « respect du fonds »
  - noms de fichiers peuvent être porteurs d'informations
  - noms de fichiers peuvent être trop longs ou contenir des signes particuliers qui peuvent poser problèmes
- Normalisation des noms de fichiers :
  - remplacement des caractères accentués par leurs équivalents non-accentués
  - suppression des caractères et symboles particuliers
  - remplacement des espaces vides par des underscores (« _ »)

  => Opération réalisée à la fin du traitement intellectuel

Données issues du mémoire de Cédric Lefebvre

2/ TRAITEMENT DU SOUS-FONDS DE L'ÉQUIPE NÉGOCIATION

Traitement des boîtes de messagerie électronique : Etat des lieux des fichiers de messagerie

Format	Nombre de fichiers	Poids	Client de messagerie associé
.pst	17	20224,00 Mo	Microsoft Outlook
.msg	605	310,89 Mo	Microsoft Exchange Server
.msf (+ fichiers sans extension)	312	17377,50 Mo	Mozilla Thunderbird
.eml	993	245,34 Mo	Format générique utilisé par beaucoup de clients (par exemple Windows Live Mail)

Données issues du mémoire de Cédric Lefebvre

2/ TRAITEMENT DU SOUS-FONDS DE L'ÉQUIPE NÉGOCIATION

Traitement des boîtes de messagerie électronique : Test de l'outil Aid4Mail

- Extraction des messages électroniques
  - Extraction des données des conteners PST en conservant leur arborescence d'origine
  - Conversion des boites et fichiers en un format unique (eml) au total 156 867 messages

- Extraction des métadonnées
  - Extraction des métadonnées de l'ensemble des boites mails dans un fichier unique (fichier .tab) qui peut être ouvert à partir d'un tableur et qui peut servir d'instrument de recherche

Données issues du mémoire de Cédric Lefebvre

DE NOUVELLES PERSPECTIVES AVEC LA MISE EN PLACE D'UN SYSTÈME D'ARCHIVAGE ELECTRONIQUE

CHANTIERS EN COURS ET A MENER

- Aujourd'hui une cartographie des applications pour un états des lieux
- …demain aller vers une Politique d'archivage plus intégrée au projet SI pour les nouvelles applications versantes

- Mettre en place des procédures et des recommandations pour maitriser les versements

UN SYSTÈME D'ARCHIVAGE ELECTRONIQUE EN COURS DE DÉVELOPPEMENT

- Implémentation de la brique logicielle VITAM dans le Système d'Archivage électronique SAPHIR

- Prise en compte de la problématique de la collecte dans son périmètre
  - Pouvoir transférer les archives de manière sécurisé en préservant leur intégrité
  - Se doter d'outils adaptés au traitement des archives numériques
    - Intégrer OCTAVE, l'outil de préparation des versements de fichiers bureautiques en cours de développement par le SIAF
    - Des espaces serveurs dédiés et d'une capacité suffisante (salle de tri virtuelle)

PROBLÉMATIQUE DE LA COMMUNICATION DES ARCHIVES

- Un réflexion encore en cours :
  - Quel objet communiquer : le courriel ou la boîte de messagerie ?
  - Comment alerter le futur chercheur des documents non valides (document de travail, version intermédiaire…) ?

MERCI DE VOTRE ATTENTION

WWW.DIPLOMATIE.GOUV.FR

@francediplo