Bulletin des Archives de France sur la conservation à long terme des documents électroniques
N° 10 janvier 2003

La valorisation et la pérennisation des données scientifiques et techniques. Le colloque tenu du 5 au 7 novembre 2002 à Toulouse

Nous publions ci-dessous le compte rendu de cette manifestation, rédigé par Isabelle Rouge-Ducos, conservateur au Ministère de la Défense.

Le congrès, organisé par le Centre National d'Études Spatiales, situé à Toulouse, avait pour but de dresser un panorama des travaux passés ou en cours conduits par différents organismes ayant en charge la pérennisation des données scientifiques, avec une présentation de leurs systèmes d'information. Mais il ne se bornait pas à une description de l'existant, puisque toute un session était consacrée au retour d'expérience et aux leçons que l'on pouvait tirer des opérations d'archivage initiées par le passé. Des interventions étaient spécifiquement consacrées aux technologies et aux normes existantes. Enfin, un pan original des communications était celui de la valorisation des données et des connaissances qui y sont attachées.

La plupart des intervenants appartenaient à des organismes de recherche spécialisés dans les sciences de la terre, ou de l'espace, dont voici un échantillon :

L'Agence spatiale européenne (ESA)

Le Centre de données astronomiques de Strasbourg (CDS, France)

Le Centre de données de la physique des plasmas (CDPP)

Le Centre national d'études spatiales (CNES)

Le Centre national de la recherche scientifique (CNRS, France)

L'Institut français de recherche pour l'exploitation de la mer (IFREMER)

Le Max-Planck-Institut für extraterrestrische Physic (Allemagne)

Météo France

Le National Aeronautics and Space Administration (NASA)

Le National Oceanic and Atmospheric Administration (NOAA, USA)

L'Office national d'études et de recherches aérospatiales (ONERA, France)

Le Service hydrographique et océanographique de la marine (SHOM, France)

Le monde occidental est bien sûr très représenté : des savants de pays comme l'Ukraine ou la Russie avaient été conviés mais n'ont pas pu venir.

Des entreprises privées présentaient également leur expérience au service des administrations, comme European Aeronautic, Defence and Space company (EADS) pour le ministère de la défense français, CS-systèmes d'information (Toulouse) expliquait la technologie EAST (description de

données binaires associant le XML).

Le public qui participait à ce colloque était nombreux, et représentatif de catégories professionnelles encore plus larges, venant à la fois du domaine public et privé, ce qui a rendu les débats très stimulants.

L'ensemble des actes du colloque est disponible sur : http://sads.cnes.fr:8010/pvdst/welcome.html

Pour d'autres informations sur ce colloque, on peut contacter Claude.Huc@cnes.fr

Ce colloque international rassemblait plusieurs communautés professionnelles scientifiques ayant développé des méthodes de gestion, de conservation et de mise en valeur des données scientifiques et techniques. Si la prise de conscience de ces enjeux n'est pas une nouveauté pour les scientifiques, ce type de manifestation commune l'est beaucoup plus, ce qui a été souligné par le Centre national d'études spatiales (CNES), organisateur de ces journées.

Consacrer un congrès à la problématique de la conservation et à la mise en valeur des données est en effet l'outil indispensable à l'échange des connaissances et des meilleures pratiques en cours, et l'occasion de susciter de nouveaux projets réalisés en coopération.

C'est également une condition essentielle pour optimiser les opérations de pérennisation. Il n'existe pas de panacée en matière de conservation sur le long terme, mais des méthodes et des préconisations peuvent être dégagées et mises en commun entre professionnels grâce à ce type de réunion. Le directeur du CNES a d'ailleurs souhaité une "pérennisation de la pérennisation" : des colloques périodiques devraient être organisés afin d'évaluer les progrès et l'évolution des techniques de conservation et de mises à disposition des chercheurs.

La pérennité des données est devenue pour les scientifiques un enjeu aussi important que leur création et leur interprétation. Quant aux services spécifiquement collecteurs, ils n'ont pas toujours mis en commun leurs informations et leurs ressources : la dispersion et l'isolement même de ces données augmentent le risque de les perdre.

Si les questions méthodologiques ont bien été mises en évidence, les aspects techniques ont été abondamment abordés, notamment pour le format XML, qui a suscité des controverses liées, selon certains, à son effet de mode mais aussi des réflexions intéressantes sur ses potentialités réelles.

Que sont les données scientifiques et techniques ?

Elles sont des résultats d'expériences ou d'observations fournis par différents instruments : des satellites, des radars, des bateaux, des bouées dérivantes, des mouillages installés dans les profondeurs des océans, des télescopes, des capteurs de stations météorologiques etc.

Les données scientifiques peuvent être des magnitudes, des rayonnements, des températures, des précipitations, des altitudes …

Ces données sont de véritables événements historiques, car les observations et les mesures qu'elles représentent ne pourront plus jamais être renouvelées dans le temps : l'instrument de mesure, les conditions naturelles, les techniques informatiques … et l'inexorable composante temporelle

conditionnent tout résultat et le rendent pour ainsi dire unique, et donc digne d'être préservé et accessible dans le futur.

Elles correspondent à un "patrimoine scientifique" qu'il faut protéger, expression qui est revenue à plusieurs reprises au cours des différents exposés, nous montrant une nouvelle fois l'acception large et tentaculaire du mot patrimoine.

Pérenniser ces données c'est également capitaliser et faire fructifier les résultats des programmes de recherches permis par d'importants financements publics ou privés.

Les agences de recherches scientifiques doivent faire face à une incroyable quantité de données à archiver : celles-ci sont ingérées par centaines de giga-octets par jour dans les systèmes informatiques. La pérennisation est donc un véritable défi pour nos sociétés modernes : la perte des données créées à grands frais et avec les avancées technologiques les plus à la pointe du progrès serait une sorte de régression, une perte irrémédiable de ce capital scientifique, lui-même à l'origine de nouvelles découvertes.

L'échange des données

Celui-ci dépend de la restitution et de la mise à disposition des données dans des environnements techniques hétérogènes. La possibilité de créer des catalogues partagés et d'échanger des données est une autre préoccupation des scientifiques, qui peut être englobée par le terme de "valorisation".

La science devient en effet de plus en plus globale, et a besoin de faire des prévisions sur le plan planétaire, en ayant recours au maximum de données sur le long terme et sur la plus vaste échelle géographique. De nombreux centres de données existent par discipline, par pays, ou par universités, mais il existe également des centres internationaux, rassemblant plusieurs disciplines scientifiques.

L'exemple du Conseil international pour la science (International Council for Science, ICSU) est à ce titre parlant puisqu'il gère environ 49 centres de données mondiaux, où toutes les disciplines sont couvertes. Les centres de données mondiaux (World data center), mis en place au début des années cinquante, sont des organismes non gouvernementaux, accueillis de façon volontaire par une administration nationale (aux Etats-Unis, il s'agit par exemple du National Oceanic and Atmospheric Agency, NOAA). Ces centres ont développé un catalogue général couvrant toutes les disciplines relatives à la terre, le Global Science Data Network (GSDN). Les centres de données mondiaux ont pour objectif de renforcer leur partenariat avec de nouveaux pays où ils étaient jusque-là absents, afin de créer des structures régionales et thématiques. Ils ont vocation à les aider dans la mise en place de l'infrastructure et du réseau, et des technologies les plus innovantes.

Les techniques utilisées sont celles des sites-miroirs entre les différents centres : cela permet un meilleur accès aux données par discipline, encourage les échanges entre communautés très éloignées, et distribue l'archivage en fonction des régions et des domaines.

Pour effectuer ces échanges, on doit disposer de la même description des données. La normalisation de cette description est un point commun à toutes les disciplines, les archivistes et les bibliothécaires le savent bien.

En astronomie, l'usage d'un standard de description depuis plus de 30 ans (Flexible Image Transport System, FITS) permet à n'importe quel astronome d'utiliser les observations de n'importe quel

instrument. D'autres normes liées à l'information bibliographique astronomique existent également : le bibcode, développé par le Centre de Données astronomiques de Strasbourg. Il s'agit d'une suite de 19 caractères intelligibles par l'homme, décrivant une référence publiée, par exemple 1990 A& A… 246..24 M signifie qu'en 1990, dans Astronomy and Astrophysics, volume 246, page 24, un auteur comme Muller (M) a publié des informations sur tel ou tel sujet. Cette référence a été utilisée très largement, notamment par la base de données bibliographique et astronomique de la NASA, ainsi que par les éditeurs de publications électroniques. Ainsi, chaque référence des articles d'une publication est en lien avec la base de données et inversement. La base de données permet également de pallier les indisponibilités des publications, en faisant le lien avec d'anciens articles numérisés à dessein. Cette base de données fait le lien avec des centres et des observatoires archivant des données qu'elle récolte pour créer un lien entre les publications et les données d'observation qui y sont mentionnées.

Les informations indispensable à la pérennisation et à la valorisation : les métadonnées (= données décrivant d'autres données)

Les instruments fournissant les données scientifiques sont eux-mêmes, par définition, des produits de la recherche souvent appelés à être modifiés et améliorés dans le temps. Ainsi, les données n'ont de sens que par rapport à l'état de l'art à un moment donné de l'histoire des technologies.

Elles doivent donc en permanence être accompagnées d'informations concernant les instruments qui les ont produites, les programmes de recherche dans lesquels elles s'inscrivent, et les scientifiques qui en sont responsables.

Dans une problématique de pérennisation, les données en tant que telles ne sont pas suffisantes, puisque leur bonne interprétation est impossible sans les métadonnées, ce qui est valable quelle que soit la manière d'y accéder (papier, informatique), et quelles que soient ces données. L'intérêt de cette préservation est bien sûr de pouvoir rendre l'information utilisable et d'obtenir des résultats plus pertinents lors des recherches : la valorisation des données en dépend comme l'a montré l'exemple de l'astronomie.

L'exemple des recherches de Météo-France était particulièrement révélateur de l'importance des métadonnées. Le programme de recherche en données anciennes y a été entamé depuis 1994 afin d'enrichir les observations consistant en moyennes mensuelles de températures minimales et maximales, et en cumuls mensuels de précipitations. Il s'agit de saisir des données pour la période 1880-1950 jusque-là beaucoup moins documentée, que les années 1950 à 2002.

Afin d'étudier les changements climatiques, il est nécessaire de disposer de séries climatiques sur le long terme ; or cette étude, à partir de séries brutes, est hasardeuse en raison des nombreux points de rupture dus au déplacement de postes, aux modifications de l'environnement de mesures, ou à la non conformité du lieu choisi pour effectuer l'expérience.

Il faut donc procéder à une homogénéisation des séries pour pouvoir les comparer avec les séries actuelles, afin d'établir des évolutions sur le long terme. Des outils d'homogénéisation sont établis à partir de méthodes statistiques, développées à la direction de la climatologie, et permettent la détection et la correction des ruptures. Actuellement, Météo France dispose de 70 séries de températures commençant avant 1900, et de 226 séries mensuelles de cumuls de précipitations. On s'aperçoit donc qu'on collecte des données anciennes et que l'on modifie les données brutes ou

"originales" afin de pouvoir améliorer la pertinence de ces mesures et l'usage qui peut en être fait . Cependant, cette méthode n'entraîne rien d'irréversible pour les données brutes, qui sont conservées par ailleurs, l'homogénéisation n'étant qu'un processus appliqué aux données brutes, lui-même appelé à être perfectionné par d'autres découvertes statistiques. La procédure d'homogénéisation n'est pas automatique, elle est validée par un expert, qui utilise les métadonnées pour confirmer une rupture ou la localiser précisément, en fonction du lieu et du type de capteur. Ainsi les métadonnées de chaque expérience permettent de procéder à l'homogénéisation, qui permet à son tour de valoriser les données en les rendant exploitables, avec les plus récentes, dans des séries homogènes. Les métadonnées concernant la procédure d'homogénéisation doivent être conservées avec les données transformées, afin de conserver la trace de cette opération.

La frontière entre données et métadonnées peut paraître arbitraire (les données homogénéisées ne peuvent-elles pas être considérées à la fois comme des données, et des métadonnées des observations initiales ?)

Les difficultés de la pérennisation : les migrations

Il est patent que les instruments de mesure, mais aussi les formats et les supports entravent la pérennisation des informations archivées, qui ne sont pas indépendantes des systèmes technologiques. Une politique de migration des supports, (éventuellement des formats), doit donc être mise au point. Celle-ci est en réalité moins une conséquence de l'obsolescence, que le fait du marché, dictant la disparition d'une technologie et rendant difficile l'acquisition d'un support.

De plus la pérennité théorique des supports (notamment optiques) ne résout pas la rapide disparition des machines de lecture.

Le problème des migrations de supports a surtout été évoqué, entre autres, par l'Agence Spatiale Européenne (ESA), après vingt-cinq années d'expérience. Cette intervention mettait en évidence la coexistence de données historiques, stables en volume, et de données provenant de missions toujours en cours.

Elles sont donc actuellement stockées sur des médias de différentes natures appartenant à la famille des supports magnétiques, allant des cassettes digitales à haute densité (support des données historiques) aux cassettes Sony Dir1000, DLT et IBM Magstar. Leur méthode a consisté à migrer une partie des données des cassettes à hautes densités sur des nouveaux supports, déterminés en fonction des catégories de données : les données à taux élevés de bits sont stockées sur des cassettes Sony DIR1000, et celles à bas taux de bits sur des DLT. Les différents supports de stockage ont permis à l’ESA de faire des statistiques de coût en les comparant. Une des raisons de la migration a été l’économie importante que permettaient de nouveaux supports. Le coût de stockage sur des cassettes HDDT revenait à 3 millions de dollars par an, tandis qu’après une première migration, le prix des consommables diminua selon un facteur estimé à 10. Le nombre des cassettes HDDT transférées est actuellement de 155 000.

L’ESA a mis au point un système d’archivage, proche de celui du CNES, permettant des migrations automatiques, depuis l’espace de stockage en ligne vers une bibliothèque et vers des archives hors ligne composées de supports plus modernes et plus compacts. Le système peut être programmé pour effectuer une migration en arrière-plan, complètement transparente, laissant accessible la consultation des données aux utilisateurs. La délimitation de différents niveaux de stockage (off-ligne, near-line, one line), selon différents supports est un moyen de promouvoir les données les

plus utilisées à la catégorie d'accès la plus rapide, et d'éloigner les données les moins consultées vers les catégories les moins rapides, et par conséquent vers des supports appropriés (stockage hiérarchisé). De plus, ces technologies permettent de faire une maintenance des lecteurs avec des nettoyages automatiques et des copies de supports en cas de détérioration. Ce système rend également possibles les duplications d'archives, stockées dans d'autres bibliothèques, différentes de celles vouées à la consultation. Les niveaux de supports les plus performants, pour un accès rapide, sont conçus à partir des technologies magnétiques, dont les potentialités ont connu une croissance sans comparaison dans les années 1990.

Dans le cas du CNES, on a choisi de ne pas gérer de médias "off-line" pour faciliter les migrations de supports. Cependant, ce système n'empêche pas l'existence et l'accès aux copies de sauvegarde, en plus des données originelles.

Ainsi, les opérations de migration, propres au stockage, et la procédure d'accès aux données font partie d'une même chaîne de traitement. La configuration de ces systèmes met en évidence la nature dynamique de l'archivage, qui n'est pas une simple sauvegarde inerte, mais bien un processus modélisé selon trois fonctions : l'ingestion, le stockage proprement dit, et la restitution pour le chercheur sur des interfaces du web, éventuellement accompagnées d'autres services. Selon ce schéma, l'archivage et la valorisation des données sont des notions intimement liées, qui nécessitent de réunir trois conditions : les formats normalisés de métadonnées, les standards de métadonnées et des supports.

Ce colloque a permis des échanges interdisciplinaires de très haut niveau, montrant combien les méthodes de pérennisation avaient progressé depuis trente ans, et combien les architectures d'archivage étaient performantes. Cependant la complexité de l'information scientifique et ses volumes importants incitent à la plus grande prudence, et nécessitent des moyens humains toujours plus nombreux et spécialisés. Les tendances actuelles vont vers un développement des sites partagés, vers une multiplication des échanges d'informations, en les rendant accessibles dans des environnements techniques différents.

Les expériences des scientifiques, et leur expertise, doivent permettre aux autres acteurs de faire progresser la conservation numérique : celle-ci est devenue légitime pour les savants, elle doit aussi le devenir dans les autres secteurs producteurs d'informations.

Isabelle Rouge-Ducos, conservateur du patrimoine, chargée de mission pour les archives électroniques
Ministère de la défense
Direction de la mémoire, du patrimoine et des archives
Sous-direction des archives et des bibliothèques
isabelle.ducos@defense.gouv.fr
MINDEF/SGA/DMPA/SDAB

L'archivage des courriers électroniques : lancement d'une étude par la direction des Archives de France

Les archivistes sont de plus en plus sollicités par des questions sur la conservation à long terme des courriers électroniques. La direction des Archives de France a donc décidé de consacrer une

attention particulière à cette question. Une première étude préalable visant à définir les grandes lignes d'un système d'archivage vient de commencer. L'étude est en cours, mais il apparaît d'ores et déjà clairement que les solutions relèvent autant, sinon plus, de la mise en place de procédures rigoureuses de gestion que de choix purement techniques.

La deuxième étape de cette réflexion sera, en 2003, un test en vrai grandeur, test pour lequel le département de l'innovation technologique et de la normalisation sera le "cobaye".

Le développement de "l' E-administration" dans les collectivités locales, et ses conséquences archivistiques

Les initiatives en faveur de "l'E-administration" ne cessent de se développer. Par ce terme, on désigne notamment la mise en place de "téléprocédures" entre le citoyen et l'administration et entre les administrations elles-mêmes. Il ne s'agit pas seulement de transmettre électroniquement des données et des documents mais aussi de créer des outils informatique de gestion complète d'une procédure administrative. Pour ne citer qu'un exemple, le contrôle de légalité des actes des collectivités locales fait actuellement l'objet d'une expérience de télétransmission menée en commun par les services du Ministère de l'Intérieur et les représentants des associations d'élus.

Cette expérience n'est pas isolée et aura évidemment des conséquences tangibles en matière d'archivage avec le passage sur support électronique de documents traditionnels. C'est pourquoi la direction des Archives de France vient de créer, en commun avec l'Association des Maires de France, un groupe de travail sur le volet "archivage" de l'E-administration. Ce groupe, qui devrait s'élargir à l'Assemblée des Départements de France, s'est fixé comme objectif premier de fournir aux producteurs et utilisateurs des documents issus des téléprocédures des recommandations et normes techniques propres à faciliter la conservation à moyen et long terme des documents concernés.

Lu pour vous : le débat sur les méthodes de conservation des documents électroniques

La revue Courrier international publie dans son numéro 634-635 (26 décembre 2002-8 janvier 2003), sous le titre Pourrons-nous lire nos archives dans dix ans ? la traduction française (partielle) d'un article paru initialement dans le numéro d'octobre 2002 de Technnology review, la revue du prestigieux Massachusetts Institut of Technology (titre original : Preserve your data for ever).

L'article souligne la fragilité de la mémoire numérique et distingue quatre méthodes pour en permettre la conservation durable :

la migration,

l'émulation,

l'encapsulation,

l'ordinateur virtuel universel.

Les deux premières méthodes sont bien connues des lecteurs du Bulletin. L'encapsulation consiste à accompagner l'objet numérique à conserver d'une double couche : une couche de données numériques "simples" qui explique comment réexploiter le document (en incluant tous les aspects nécessaires: logiciel, système d'exploitation, type d'encodage etc.) et une couche en clair (texte écrit sur une étiquette par exemple) exposant ce qu'est le document et comment l'utiliser.

Le principe de l'ordinateur virtuel universel, défendu par Raymond Lorie, chercheur chez IBM, consiste à créer un programme volontairement très simple, qui serait capable, par sa simplicité même, de traiter tous les types de données numériques et de fonctionner avec n'importe quelle plate-forme technique.

Ces articles relèvent par ailleurs, en la déplorant, la faiblesse des financements consacrés à la recherche dans ce domaine.

Pour en savoir plus : http://www.technologyreview.com/forums/forum.asp?forumid=97

(cette adresse correspond au forum de discussion de la Technology review consacré à cet article, qui comprend des échanges particulièrement intéressants avec des liens utiles. L'accès au texte intégral de l'article n'est possible que sur abonnement)

Prochain Bulletin

Le prochain Bulletin paraîtra vers la mi-mars. Merci de faire parvenir à Joël Poivre (joel.poivre@culture.gouv.fr) vos informations ou suggestions.