

# L'archivage numérique aux Archives fédérales - un rapport d'expérience

**Peter Keller-Marxer**

Archives fédérales Berne

Chef du Service ARELDA et

Directeur du projet e-Gouvernement ARELDA

E-Mail: [Peter.Keller@bar.admin.ch](mailto:Peter.Keller@bar.admin.ch)

*(Traduction française par Jean-Daniel Zeller et Myriam Erwin de l'article original allemand paru dans ARBIDO, no 3, 2003, pp. 13-15)*

Les Archives fédérales suisses archivent des documents numériques de l'administration fédérale depuis 1982. La quantité de données numériques archivées aujourd'hui aux Archives fédérales équivaut à sept Terabyte (TB ; un Terabyte représente environ 1'000 gigaoctets), ce qui correspond à la quantité de données que l'on peut sauvegarder sur 12'000 CD-R commerciaux. En 2003, la quantité de données augmentera approximativement d'environ neuf autres TB; et dès 2004, il faudra tabler sur un taux d'accroissement de 20 TB par année.

Pour la prise en charge courante, le traitement et la conservation des données numériques, le Service ARELDA (**A**rchivierung **e**lektronischer digitaler **D**aten und **A**kten [Archivage des données et documents numériques sur supports électroniques]) des Archives fédérales bénéficie d'une structure informatique autonome, particulièrement sécurisée, et gérée en collaboration avec le Centre de services informatiques du Département de l'intérieur. Les données archivées sont sauvegardées sur des bandes magnétiques du type AIT-2 (Advanced Intelligent Tape, de la firme Sony), qui sont gérées dans plusieurs robots échangeurs de bandes (tape library).

## **Le projet ARELDA: recherche de solutions à long terme**

Dans le processus de travail d'archivage quotidien et courant, des solutions "ad hoc" sont élaborées : pour chaque cas, les formats des données et les interfaces sont négociés avec les offices fédéraux versants, avec des coûts techniques et en personnel relativement importants. Le traitement, le contrôle de qualité et l'intégration des données dans les Archives suivent par analogie l'expérience et les "best practices" acquises lors des versements précédents. Il n'existe cependant pas d'architecture des données globale, ni de procédures standardisées, ni de normes claires ou des processus pouvant être automatisés.

Momentanément, cette manière de procéder peut certes empêcher des pertes de la mémoire patrimoniale à court terme; cependant, à long terme, avec la quantité et l'hétérogénéité fortement croissantes des fonds numériques, aucun dépôt d'archives numériques ne pourra se laisser planifier ou gérer de cette manière, tant sur le plan technique que sur le plan des ressources en personnel ou financières.

C'est pourquoi les Archives fédérales ont pris l'initiative en 2000, avec le projet ARELDA<sup>1</sup>, ambitieux tant sur le plan professionnel que sur le plan financier, de développer et d'instituer des solutions, qui devraient garantir l'archivage numérique à long terme aux Archives fédérales. Cela signifie que les méthodes, les normes, les concepts et les processus à implémenter devront alors être solidement fondés et garantir la planification et le financement de l'archivage ainsi que l'entretien continu des données, si la base technique supportant le système en place devient obsolète et doit être remplacée - comme on peut s'y attendre - tous les dix ans.

Le projet ARELDA est conjointement une unité spécialisée du même nom des Archives fédérales et un des cinq projets clé de la stratégie de cyberadministration (e-Government) de la

---

<sup>1</sup> [http://www.bar.admin.ch/webserver-static/docs/f/arelda\\_expose\\_0401\\_f.pdf](http://www.bar.admin.ch/webserver-static/docs/f/arelda_expose_0401_f.pdf)

Confédération<sup>2</sup>. La première étape 2001-2004 du projet ARELDA, qui court jusqu'en 2008, est principalement financée par le Groupe interdépartemental de coordination 'Société de l'information' (GCSI) du Conseil fédéral<sup>3</sup>.

L'analyse de la situation et des risques qui a précédé le lancement du projet ARELDA d'e-Government a conduit à une nouvelle orientation méthodologique à l'égard des anciens travaux, menés sous le même nom depuis 1992. Cela concerne avant tout le constat que les connaissances informatiques manquantes ou insuffisantes dans le domaine des archives se révèlent à un double égard l'obstacle essentiel à la recherche de solution. D'une part, l'absence de compétences professionnelles en informatique empêchait une analyse de l'efficacité au niveau conceptuel: les propositions de solution techniques des sociétés extérieures, prestataires de services, ne pouvaient pas être évaluées sur leur efficacité ou leur aptitude à répondre aux objectifs archivistiques, objectifs formulés avant tout en fonction de l'ancienne compréhension que l'on avait des archives, lesquelles étaient liées au support papier.

Tout aussi frustrant fut l'échec de la collaboration 'commerciale typique' avec les prestataires externes de services informatiques. Nous entendons par le qualificatif 'commercial typique' qu'il n'est habituellement pas nécessaire que le commanditaire ait une compréhension de la solution technique de son problème, tant qu'il peut définir assez exactement ses besoins et exigences professionnels. C'était cependant exactement ce qui n'était pas possible dans le cas de l'archivage numérique à long terme, puisque les problèmes fondamentaux se trouvent exister précisément sur le plan technique et que les concepts habituels du domaine archivistique "conventionnel" que sont l'intelligibilité, l'accessibilité, le maintien et la conservation des fonds, l'intégrité, l'authenticité, la communicabilité, etc. doivent dans le contexte technico-numérique être en grande partie tout d'abord re-compris et redéfinis.

A côté du manque *d'analyse de l'efficacité (des solutions)*, l'incapacité *d'analyser le problème* sur le plan fondamental s'est ainsi révélée être le second facteur de blocage. Cela a donc été une décision stratégique, que d'orienter la méthodologie de processus d'ARELDA d'une façon nouvelle telle, qu'un point essentiel de cette méthodologie repose sur l'élaboration et l'institutionnalisation d'une compétence professionnelle « d'informatique d'archives », propre aux Archives fédérales. Ceci a eu comme conséquence que la moitié des huit personnes occupées aujourd'hui au Service ARELDA sont des informaticiens de profession.

Les expériences issues de cette politique sont globalement positives. Dans le domaine de *l'analyse du problème*, c'est avant tout la possibilité d'un prototypage expérimental propre aux Archives (logiciel et système pilote) qui s'est révélée comme un instrument essentiel pour la réduction des risques et des coûts inhérents au projet, du fait que grâce aux prototypes propres, la faisabilité et la plausibilité de parties critiques du système ont pu être examinées, les principales fonctionnalités ont pu être graduellement développées dans des systèmes pilotes, des exigences consistantes et complètes ont pu être définies et des "solutions pré-machées" pour les prestataires externes ont pu être spécifiées.

Dans le domaine *de l'analyse d'efficacité*, l'expérience a été faite qu'il est parfaitement justifié de parler de la nécessité absolue d'une informatique d'archives professionnelle, spécialisée dans ce domaine. Aujourd'hui, un segment croissant du marché de l'informatique propose des solutions d'archivage intégrées, qui n'ont cependant rien à voir avec la tâche d'archivage numérique à long terme, telle que l'entendent des Archives nationales. Ces solutions s'orientent au contraire majoritairement sur un horizon temporel de dix ans (le délai légal de conservation typique pour les documents comptables des sociétés privées) et supposent que l'acheteur possède le contrôle complet sur la production de ses documents, et qu'il peut donc déterminer les types et les paramètres des documents archivables déjà dès leur production, sur la base des fonctionnalités du système d'archivage à acquérir.

Il est généralement impossible d'établir plausiblement quant aux produits du marché comment les fonds d'archives hétérogènes – le volume de données est estimé raisonnablement aujourd'hui

<sup>2</sup> <http://www.admin.ch/ch/f/egov/egov/strategie/strategie.html>

<sup>3</sup> <http://www.admin.ch/ch/f/egov/egov/kig/kig.html> et <http://www.infosociety.ch>

d'environ 100 TB – pourraient être transférés dans de nouveaux systèmes sans perte d'authenticité ni d'information *et* avec un coût financier supportable, *et ce* après la "durée de vie" typique d'un produit propre à un fabricant (soit 15 ans).

De plus, il ne doit pas être oublié que "l'obsolescence technologique" représente *le* facteur de revenu essentiel du marché informatique et que dans ce domaine, la réflexion sur des espaces temporels dépassant dix ans s'avère par expérience des plus spéculatives et est de ce fait considérée comme sans intérêt.

C'est cependant justement le fait de s'occuper de ces questions qui fait la spécialisation professionnelle de l'informatique d'archives et qui permet d'évaluer l'efficacité des solutions, de les mettre en pratique de façon continue, qui rend possible enfin la mise en place d'une veille technologique et d'une analyse de risque pour la gestion à long terme d'un dépôt d'archives numériques.

### Caractéristiques de l'archivage numérique

La discussion autour de la "disponibilité et de la fiabilité" de l'information mise à disposition par la technologie informatique (information availability & reliability), qui est très en vogue dans le domaine informatique, se laisse poursuivre sans discontinuer du monde de "l'information vivante" dans le domaine des archives : l'archivage signifie la disponibilité illimitée dans le temps des informations, c.-à-d. devoir satisfaire à quatre exigences fondamentales :

- *La persistance* : capacité d'un objet d'archives numérique d'exister plus longtemps que chaque équipement technique le rendant accessible. "Exister" signifie également ici implicitement l'accessibilité (la lisibilité) de l'objet.
- *L'intégrité physique*: la conservation sûre et indemne, c.-à-d. l'intégrité et le non-endommagement d'un objet d'archives numérique dans le temps, au niveau du bit.
- *L'authenticité*: «intégrité intellectuelle»; l'authentification (de l'auteur et de la provenance) et la fiabilité (de la valeur d'évidence contenue). Cela implique aussi comme condition préalable la compréhensibilité intellectuelle.
- *La continuité*: la présence simultanée, corrélée (c.-à-d. mise en relation mutuelle) et continue des caractéristiques mentionnées ci-dessus, dans un processus paramétré (c.-à-d. défini qualitativement et quantitativement et de ce fait mesurable) dans la durée.

Cependant, la tentative de livrer une interprétation de ces caractéristiques déterministe, mesurable grâce à des paramètres quantitatifs, conduit, pour les documents numériques, à des questions non encore résolues à ce jour <sup>4,5,6,7</sup>. Les problèmes surgissent pour une raison déterminante, le fait que ces concepts ne sont pas indépendants l'un de l'autre : la persistance est une condition de l'intégrité (non l'inverse), mais pas de l'authenticité; l'authenticité est une condition de l'intégrité, mais pas de la persistance, etc.

Le concept de "continuité" implique un autre degré de complexité de par sa composante temporelle, comme l'exemple trivial qui suit le montre. Contrairement à un document papier authentique signé manuellement, gardé quelques 400 ans dans les archives, l'authenticité d'un document numérique, muni d'une signature numérique valable aujourd'hui, n'est pas constante : une signature numérique (dont la validité légale est de toute façon limitée à quelques années) et son encryptage sont éliminés lors de l'archivage ou sont pour le moins détruits à la première

<sup>4</sup> Authenticity in a Digital Environment; Council on Library and Information Resources, Washington, DC. (Ed.), 2000, <http://www.clir.org/pubs/reports/pub92/pub92.pdf>

<sup>5</sup> Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust; Clifford Lynch, CLIR, 2000, <http://www.clir.org/pubs/reports/pub92/lynch.html>

<sup>6</sup> Preserving the Authenticity of Contingent Digital Objects; Anne J. Gilliland-Swetland and Philip B. Eppard, D-Lib Magazine Volume 6, Number 7/8, 2000, <http://www.dlib.org/dlib/july00/eppard/O7eppard.html>

<sup>7</sup> Trusted Digital Repositories: Attributes and Responsibilities, An RLG-OCLC Report, Research Libraries Group, Mountain View, CA, May 2002, <http://www.rlg.org/longterm/repositories.pdf>

migration/conversion du support. Dans ce cas, l'archiviste doit remplacer l'authenticité "originale" par une authenticité de qualité équivalente, de laquelle l'archiviste doit se porter garant lui-même: les services d'archives reprennent le rôle d'un substitut, lequel doit toujours redéfinir, produire et garantir l'authenticité originale des documents numériques.

L'interprétation de tous ces concepts doit fondamentalement prendre en considération l'obsolescence technologique:

- Les formats des supports, des fichiers et des données passent de mode en peu d'années et deviennent obsolètes.
- Les logiciels propriétaires d'un fabricant pour le traitement et la lecture des données ne sont plus disponibles après quelques années. Les nouvelles versions du même logiciel ne peuvent plus lire correctement les données produites avec les versions précédentes.
- Les technologies des supports de données, grâce auxquelles les données sont stockées, et les lecteurs correspondants disparaissent rapidement du marché.

Dans tous les cas, l'obsolescence empêche que les documents numériques soient maintenus sous une forme originale complètement identique plus de quelques années, puisque ce sont aussi les environnements logiciels qui les produisent, et fréquemment aussi les environnements matériels originaux, qui devraient être conservés et entretenus de pair. L'obsolescence force au contraire au reconditionnement avant l'archivage d'une manière ou d'une autre des documents en vue d'un archivage de longue durée, opération engendrant nécessairement une perte de l'authenticité et de l'information. Ceci pose la question de la définition des concepts de persistance, d'intégrité, d'authenticité et de continuité dans ces environnements reconditionnés d'archives, et ce fondamentalement à des degrés différents pour chaque type de documents.

## **Les principes du développement de la solution ARELDA**

Le choix des principes (stratégiques) essentiels pour le développement de solution dans le projet ARELDA s'est orienté à partir des questions et constatations suivantes :

- Dans quelles conditions et jusqu'où, les quatre concepts mentionnés se laissent définir au mieux, et implémenter de manière mesurable et contrôlable à long terme ?
- Progresser avec la conscience de transmettre un problème incomplètement résolu à la prochaine génération, mais sous une forme plus manipulable (du point vue actuel !) et finançable. ("principe d'étapes", échelonnement)
- Où peuvent être trouvées les synergies professionnelles avec les ébauches de solution existantes dans les domaines apparentés ?
- Dans quelles conditions, une gestion des risques<sup>8</sup> peut-elle être effectuée de manière plausible?

Sur ce dernier point, il faut remarquer qu'il y a fondamentalement trois manières possibles de perdre les fonds d'archives numériques :

- *La perte physique* par un risque élevé de fausses manipulations ou de défauts;
- *La perte logique* par une obsolescence irréversible;
- *La perte opérationnelle* par des dépenses inflationnistes, ne pouvant plus être financées, pour les interventions manuelles lors de la gestion et des migrations, de fonds de données de plus en plus hétérogènes et documentés techniquement de façon lacunaire.

Par expérience, la perte opérationnelle se trouve clairement au premier plan. La vraisemblance de son occurrence dépend directement du budget à disposition, elle reste cependant considérable dans tous les cas.

Après l'analyse de ces données du problème, le principe choisi se laisse reformuler comme "archivage indépendant de l'application" : soit l'archivage numérique de longue durée au-delà des générations technologiques, dont la condition est que les documents :

- soient détachés des environnements spécifiques les ayant produits (logiciel, matériel, formats de

<sup>8</sup> Risk Management of Digital Information, Gregory W. Lawrence et al., Council on Library and Information Resources, Washington, D.C, 2001, <http://www.clir.org/pubs/reports/pub93/pub93.pdf>

mémoire, de données et de fichiers) (en s'accommodant des pertes en information et en authenticité),

- et transférés dans des formats de données et des environnements ouverts, normalisés, aussi génériques que possible et avant tout complètement documentés,
- puissent y être entretenus dans des cycles de migration aussi longs que possible (au moins 15 ans) ;
- et que les fonctionnalités (logicielles, matérielles) ne soient par principe pas archivées (mais seulement documentées).

Aujourd'hui, ce point de départ est préconisé universellement par beaucoup d'institutions et est considéré comme le meilleur compromis entre un processus pragmatique et une disponibilité technologique à long terme. Mais l'équilibre entre des pertes à subir et un environnement générique pose de hautes exigences.

Cette approche, présentée seulement superficiellement ici, suppose cependant une préservation soignée de la qualité archivistique des données<sup>9</sup> du document d'archives potentiel par:

- la mise en relief de la valeur à long terme des informations : le développement d'instruments d'évaluation prospective concrets et maniables, selon des critères pertinents (légaux, économiques, scientifiques, historiques, militaires, etc.);
- l'identification des éléments essentiels, porteurs de signification;
- la collecte de toutes les métadonnées nécessaires à la compréhension intellectuelle de longue durée et à la gestion technique des données.

Sans oublier ce point essentiel : la qualité archivistique des données ne peut être assurée dans le cas des documents numériques qu'au début, et uniquement au début du cycle de vie des documents.

## **Les "stakeholders" de l'archivage numérique**

Différents groupements d'intérêts ont établi le manque de solutions et l'urgence du problème de l'archivage numérique à long terme. Aujourd'hui, les "stakeholders" importants de l'archivage numérique à long terme, à côté des Archives publiques et des bibliothèques, appartiennent avant tout à l'industrie pharmaceutique et aux Science Data Centers des sciences naturelles. L'industrie pharmaceutique se voit soumise à une pression massive du fait de la pratique d'autorisation universellement adoptée de la Food and Drug Administration (FDA) américaine: la définition légale du "21 CFR Part 11"<sup>10</sup> de 1997, fixe des exigences très vastes sur la production, le Records management, la présentation et l'archivage authentique à long terme des documents électroniques pour les autorisations dans le domaine des sciences de la vie.

Les Science Data Centers gèrent et archivent les résultats des mesures scientifiques issues des sciences naturelles, ainsi de la navigation spatiale, des missions des géosatellites, de la physique expérimentale, de l'océanographie, de la météorologie etc., typiquement dans des ordres de grandeurs de plusieurs Petabytes (plusieurs 1'000 TB), avec des taux d'accroissement qui vont jusqu'à plusieurs Terabytes par jour. Il existe un intérêt vital à maintenir de telles données disponibles à long terme : leur production est liée à des investissements de milliards de dollars et leur acquisition ne peut habituellement pas être répétée (par exemple les données climatiques). En outre, de telles données brutes originales doivent aussi pouvoir être ré-exploitées dans 50 ans ou plus, sur la base de nouveaux modèles théoriques. Cependant, au cours des dernières années, les organisations comme la NASA ont remarqué que beaucoup de leurs anciens fonds contenant des données irremplaçables ne sont plus compréhensibles et de ce fait, ne sont plus utilisables. Cela a eu pour conséquence que quelques-unes de ces institutions se sont occupées intensivement des besoins d'un archivage à long terme des données des sciences naturelles.

---

<sup>9</sup> L'ouvrage suivant, accessible également à des non-informaticiens, offre une introduction utile sur le concept des techniques de l'information de „qualité des données“: Enterprise Knowledge Management: The Data Quality Approach; David Loshin; Morgan Kaufmann / Academic Press, San Diego, 2001, ISBN 0-12-455840-2

<sup>10</sup> [http://www21cfrpart11.com/pages/fda\\_jlocs/](http://www21cfrpart11.com/pages/fda_jlocs/)

Les comptes rendus du symposium international "La pérennisation et la valorisation des données scientifiques et techniques"<sup>11</sup>, organisé l'année dernière par le Centre National d'Etudes Spatiales (CNFS) à Toulouse, montraient que les principes de traitement définis pour la compréhensibilité à long terme de ce genre de données sont formellement singulièrement semblables aux principes de mise en valeur établis pour celles de certaines Archives nationales d'ici. Cela concerne avant tout une description systématique, et procédant par niveau, des « unités de description » usuelles dans ce domaine : les "missions" qui contiennent plusieurs "expériences", lesquelles se composent chacune de différents "instruments", lesquels à leur tour produisent plusieurs "séries de résultats de mesures". A chaque étape, on décrit les connaissances contextuelles, qui étaient actuelles au moment de la production des données, et nécessaires à la compréhension à long terme : l'état des connaissances et des technologies, les théories et procédés utilisés, les termes techniques, les unités de mesure, les formats de données, etc.

L'intérêt vital éveillé par l'archivage à long terme a incité des organisations des domaines de la navigation spatiale, de l'industrie aéronautique, de la physique expérimentale, de l'océanographie, de la météorologie, etc., à produire un modèle de référence pour les archives numériques : le "Reference Model for an Open Archival Information System" (OAIS)<sup>12</sup> est apparu en 1999 comme une recommandation du Consultative Committee for Space Data Systems, lequel est supporté par plus de 30 organisations internationales, dont la NASA, l'agence spatiale européenne ESA, le Centre national d'Etudes Spatiales français (CNES), la National Oceanic & Atmospheric Administration (NOAA) américaine et le World Data Center Panel (WDCP).

En 2002, l'OAIS a été accepté comme norme internationale ISO 1472 1:2002. Entre-Temps, il a également éveillé un grand intérêt dans le domaine des bibliothèques et des services d'archives. Il représente un modèle fonctionnel et conceptuel pur, qui établit plus de 50 fonctions ainsi qu'une ontologie grossière pour les archives numériques et qui devrait permettre de rendre les futurs systèmes comparables. À moyen terme, une certification ISO devrait également être recherchée pour les systèmes conformes au modèle OAIS.

L'OAIS développe un effet avant tout parce qu'il propose une terminologie unifiée et interdisciplinaire pour la discussion sur les archives numériques et qu'il contribue ainsi à ce qu'une compréhension interdisciplinaire de la problématique se développe lentement. De ce fait, des compétences complémentaires contribueront dans l'avenir à maintes synergies : d'une part l'expérience technologique de plusieurs années dans le stockage de longue durée de très grandes quantités de données (Science Data Center), d'autre part la compétence de plusieurs années en matière de contextualisation, de mise en valeur, et de métadonnées (Archives).

---

<sup>11</sup> <http://www.cnes.fr/pvdst/>

<sup>12</sup> <http://ssdoo.gsfc.nasa.gov/nost/isoas/overview.html> et <http://www.rlg.org/longterm/oais.html>