

Standard d'échange de données pour l'archivage électronique

F. Banat-Berger
Présentation Groupe PIN
20 octobre 2005

Le standard d'échange

- Il vise à décrire un ensemble de métadonnées et son implémentation pour permettre le transfert d'archives électroniques d'un service versant à un service d'archives, avec les informations nécessaires à leur préservation, et la communication de ces archives au service versant et au public.
- Le standard défini est générique et adaptable à tous types de documents et de données électroniques.
- Il est utilisable pour les transmissions automatiques par voie électronique mais aussi en cas de transmission électronique déclenchée manuellement ou en cas de transmission par support amovible

Origine

- Les premiers travaux ont été menés avec la Caisse des dépôts et consignations
- Dans le cadre du projet du contrôle de légalité dématérialisé et de l'utilisation de la plate-forme FAST
- Poursuivi ensuite avec l'ADAE pour le rendre complètement générique

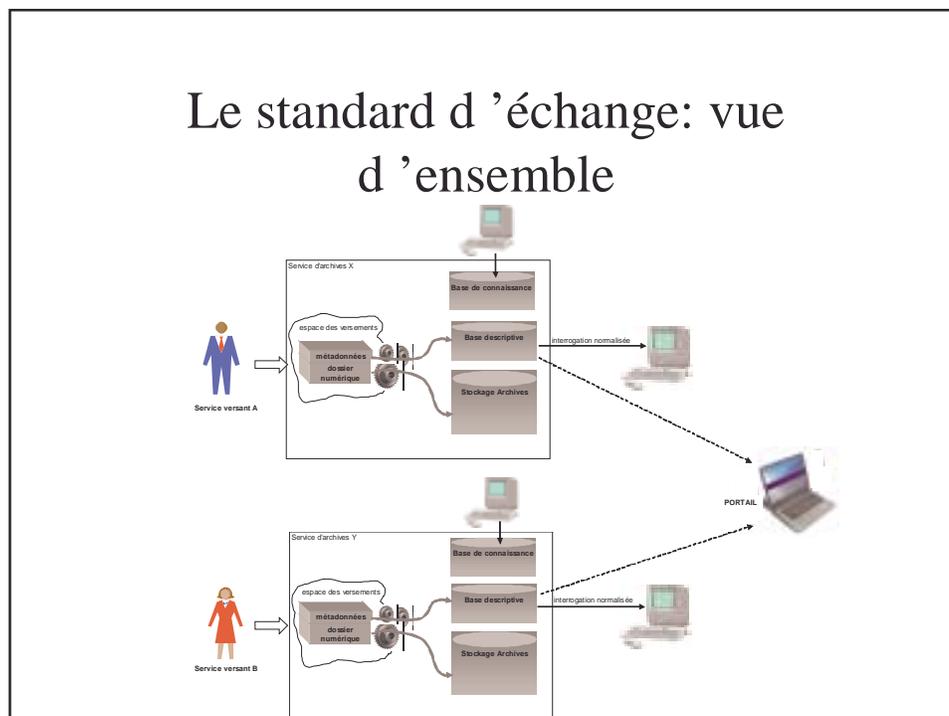
Public visé

- producteurs d'archives publiques
- services publics d'archives
- éditeurs de logiciels (développement de modules d'archivage)
- prestataires de services d'échanges
- sociétés privées d'archivage
- services d'archives étrangers

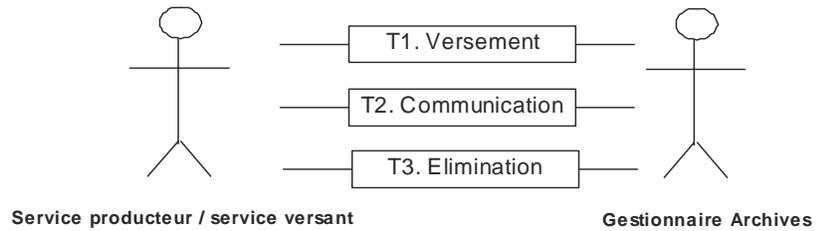
Le standard d'échange

- Il est basé autant que possible sur les normes :
 - XML
 - la norme ISO 14721 (modèle OAIS)
 - la DTD EAD (Encoded Archival Description)
 - le format d'encodage base 64 qui permet une encapsulation de fichiers binaires en relation les uns avec les autres dans un même versement

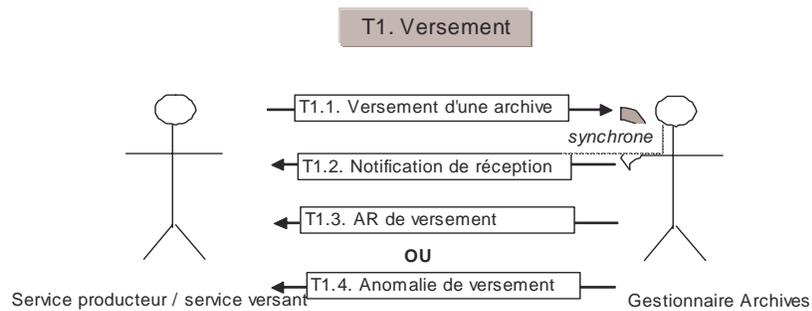
Le standard d'échange: vue d'ensemble



Vue d'ensemble des transactions



Vue d'ensemble des transactions « versement »



Standard d'échange

Le versement d'une Archive

Le versement : modélisation

- `<Versement>`
- `<ServiceVersant Identifiant = " Identifiant du service versant " Nom= " Nom du service versant " />`
- `<Archiver Identifiant = " Identifiant de l'archiver " Nom= " Nom de l'archiver " />`
- `<IDVersementServiceVersant>` Identifiant du versement fourni par le service versant
`</IDVersementServiceVersant>`
- `<Date>` Date et heure de l'échange `</Date>`
- `<Empreinte Algorithme="SHA-1" MiseSousFormeCanonique="URI Identifiant l'algorithme">` Empreinte du versement `</Empreinte>`
- `<RefIDArchiver>` (Elément optionnel utilisé dans le cas d'un versement précédent ayant levé une anomalie) Identifiant d'un versement précédent `</RefIDArchiver>`
- `<Archive MiseSousFormeCanonique="URI Identifiant l'algorithme"/>`
`<Description>` Description de l'objet (format XML selon la DTD EAD) `</Description>`
`<Empreinte Algorithme="xxx" MiseSousFormeCanonique="xxx">`HJHG`</Empreinte>`
`<Objet MediaType="xxx" Taille="xxx">` Objet archivé (format XML) `</Objet>`
- `</Archive>`
- `</Versement>`

Le versement : modélisation (suite)

Les principes suivants ont été adoptés :

- un versement contient une archive et une seule ;
- une archive contient une description et une seule ;
- une archive contient au plus un objet ;
- un objet peut être vide ou correspond à un ou plusieurs documents numériques.

Le standard d'échange permet l'atomisation des versements.

Par exemple, pour le contrôle de légalité dématérialisé, on peut faire correspondre le versement à une transmission d'une collectivité territoriale à une préfecture. Dans ce cas, il y aura autant de versements aux archives que de télétransmissions.

L'atomisation des versements

- Elle comporte des avantages :
 - en cas d'anomalie, renvoi d'une petite unité
 - transfert par réseau facilité (taille limitée des versements)
- Une application permettra d'automatiser de traitements, par exemple valider tout un lot de versements (après avoir procédé à des sondages)
- Mais les versements de petit volume ne sont pas une obligation pour le service versant (possibilité de versements volumineux à faible fréquence, sur supports amovibles ; découpage possible par le logiciel de gestion des archives (transparent pour le service versant))

Unicité de la description

- Plus simple que d 'imaginer plusieurs descriptions d 'un même objet exprimées dans des systèmes descriptifs différents
- Et... moins volumineux

Unicité de l 'objet à l 'intérieur de l 'Archive

- Ne signifie pas qu 'un versement n 'inclut qu 'un seul document numérique : plusieurs documents peuvent être encapsulés dans l 'objet
- Non dissociation entre l 'Archive et la description (encapsulation XML)

Correspondance avec l 'OAIS

- Élément Archives : constitue un paquet d 'informations
- Élément Objet : contenu d 'information, soit l 'objet contenu et l 'information de représentation à travers les attributs de l 'élément
- La description (identification, provenance, contexte) et l 'empreinte (intégrité) correspondent à l 'information de pérennisation
- Répartition des données à archiver et des informations à propos de ces données, entre les éléments description, empreinte d 'une part, et objet, d 'autre part, est une information d 'empaquetage
- L 'information descriptive est obtenue en dupliquant la description dans une base de données extérieure à l 'archive
- Correspondances non absolues notamment pour l 'information de représentation

L 'objet

- Le contenu de l'élément <Objet> peut inclure un ou plusieurs fichiers binaires préalablement convertis au format texte par encodage en Base 64.
 - Les actes soumis au contrôle de légalité doivent être transmis à la préfecture au format PDF ou PNG. Dans le cadre des expériences de télétransmission menées dans le département des Yvelines, ils sont intégrés, au format Base 64, dans un fichier XML (le fichier XML pourra être intégré tel quel dans l'élément <Objet> de l'archive) :
 - <ascl:Fichiers id="00">
 - <ascl:Fichier Encoding="Base64" Format="PDF 1.5" Id="_999-999777444-20050126-BAT160000-AR_1_1_1">JVBERi0xLjMKJcfsj6lKNiAwIG9iago8PC9MZ W5ndGggNyAwIFlvRmlsdGVyIC9GbGF0ZURlY29kZT4+C nN0cmVhbQp4nHWQPQ7CMAyF95zCY1IMkrZOOoIEO5 ALoPlj UAsqhfty</ascl:Fichier>
 - </ascl:Fichiers>

L 'objet

- Si plusieurs fichiers physiquement indépendants forment néanmoins un ensemble logique à regrouper dans la même archive, il est nécessaire de définir une structure XML élémentaire capable d'encapsuler ces différents fichiers.
- Il est conseillé de prévoir des attributs ID dans les différents éléments de la structure XML, de manière à faciliter les liens vers des parties de l'objet.

L 'objet

- **Un dossier papier de dix pages a été numérisé sous forme de dix images au format JPEG.**
 - Pour former un objet unique, ces dix images doivent être encodées en Base 64 et intégrées dans une structure XML de ce type :
 - `<dossier numero="2005-1256" id="00">`
 - `<page1 id="01" encoding="Base64">Q2VjaSBlc3QgdW4gZXhlcXBsZSAh</page1 >`
 - `<page2 id="02" encoding="Base64">Q2VjaSBlc3QgdW4gYXV0cmUgZXhlcXBsZSAh</page2>`
 - `<page3 id="03" encoding="Base64">RXQgZW4gdm9pY2kgdW4gdHJvaXNp6G1llCE=</page3>`
 - ...
 - `<page10 id="10" encoding="Base64">RXQgbGUgZGVybmlldiAh</page10>`
 - `</dossier>`

Attributs de l 'objet et de l 'empreinte

- Média type (format du contenu de l 'objet):
test/xml
- Taille
- L 'élément empreinte de l 'objet est facultatif. Ses attributs sont d 'une part, l 'algorithme et d 'autre part, la MiseSousFormeCanonique (méthode de mise sous forme canonique retenue pour le calcul de l 'empreinte)

La description

- Elle porte à la fois sur le contenu, la gestion et le format.
- Elle doit refléter la structure logique de l 'objet, et indiquer, pour chacune de ses divisions, à quelle partie physique de l 'objet elle correspond
- Toute partie du contenu de l'élément <Objet> dont l'intelligibilité doit être maintenue à long terme doit être référencée dans l'élément <Description>. C'est notamment le cas des fichiers binaires encodés en Base 64.

La description et la DTD EAD

- Le contenu de l'élément <Description> doit être conforme à la DTD EAD dans sa version en vigueur (actuellement la version 2002) et au modèle décrit dans les paragraphes suivants.
- Ce modèle retient la structure, les éléments et les attributs qui ont semblé les plus pertinents pour les versements d'archives électroniques. Sur les 146 éléments présents dans la DTD, seuls 39 sont utilisés dans le présent standard.
- Il est important que la structure EAD retenue soit la plus simple et la plus précise possible. Une structure trop complexe ou trop mal typée compliquerait singulièrement la réalisation d'exports et d'imports.

La DTD EAD

- La DTD EAD permet une description à plusieurs niveaux : l'objet dans son ensemble puis chacune de ses parties et sous-parties, sans limitation.
- Toutefois, pour un objet (dossier ou groupe de dossiers en règle générale), cinq niveaux de hiérarchie, en incluant la description d'ensemble de l'objet, ont été considérés comme suffisants.

La DTD EAD

- Elle est générique par nature
- Ne restitue pas par conséquent toute la richesse métier
- Mais permet de réaliser un effort de synthèse

La notification

- Elle est émise en synchrone
- Elle permet la fourniture d'un identifiant unique du versement
- Si la notification ne parvient pas au service versant, celui-ci ré-émettra son versement avec le même identifiant service versant

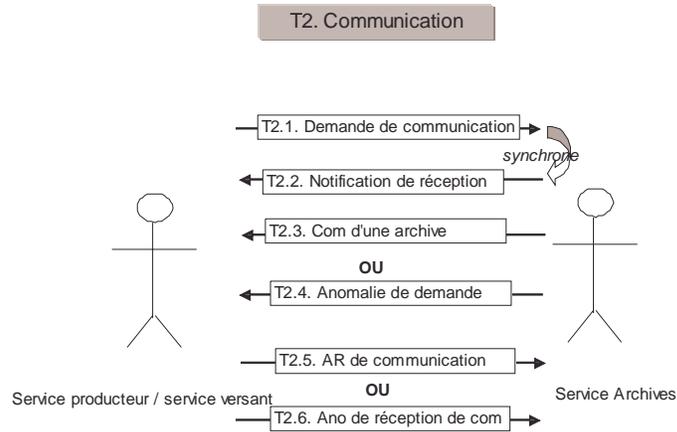
L 'accusé de réception

- Il est émis en asynchrone à la suite de résultats positifs des différents contrôles : empreinte, format, structure XML, EAD..
- La non réception au terme d 'un certain délai donnera lieu à un traitement hors ligne de l 'anomalie

L 'anomalie

- Permet de préciser la nature de l 'anomalie que ce soit pour le versement, la communication ou l 'élimination
- Exemples : échec de vérification de l 'empreinte, non conformité de la structure XML, non conformité du contenu, identifiant manquant ou non conforme

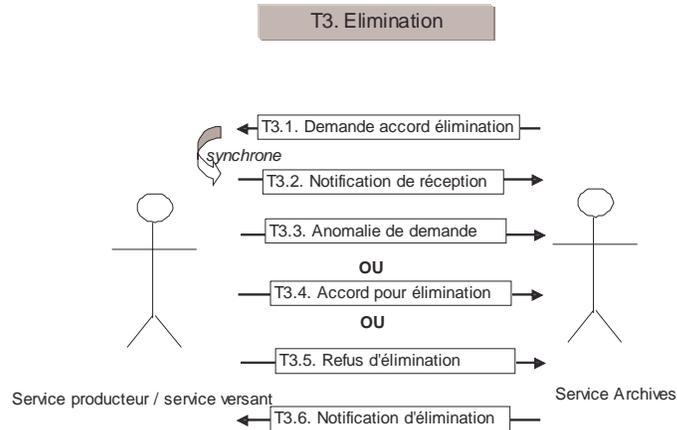
La communication



La communication

- Préalable : le service demandeur s'est fait enregistrer auprès du service d'archives
- Pourra être communiquée toute l'Archive mais également des sous objets
- Une notification de la part du service d'archives est prévue, en mode synchrone
- Lors de la communication (prévue en mode a-synchrone), est prévue la génération d'une empreinte, ainsi qu'un accusé de réception de la part de celui qui reçoit la copie (suite à un résultat positif des vérifications réalisées après communication)

L 'élimination



L 'élimination

- Est prévue une demande d 'accord au service versant, formulée par le service d 'archives
- Lors de la demande d 'accord, une copie de l 'Archive concernée est envoyée afin que le service versant puisse vérifier sur pièces de quoi il s 'agit

L 'élimination

- La demande ne peut porter que sur un versement entier (et non sur une partie de versement)
- Même principe de notification en synchrone de la demande d 'accord d 'élimination, qui ne donne pas lieu à l 'émission d 'un nouveau identifiant (celui fourni par le service d 'archives est suffisamment univoque)
- Une anomalie peut être déclenchée en mode a-synchrone par le service versant (document illisible par exemple)

L 'élimination

- Le service versant doit donner son accord
- Ou refuser l 'élimination
- Si l 'élimination a lieu, elle donne lieu à l 'émission d 'une notification au service versant.

Perspectives

- Vient de faire l'objet d'un appel à commentaires sur le site de l'ADAE
- D'ici le mois de janvier, 3 ateliers thématiques seront organisés