Présentation du module Versement du projet SPAR
Système de Préservation et Archivage Réparti

Groupe PIN - Septembre 2008

Agenda

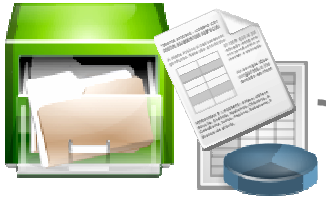
1. Contexte dans SPAR
2. Description générale du module
3. Description des données
4. Description des actions principales
5. Focus sur les développements
6. Conclusion

Le projet SPAR de la Bibliothèque nationale de France

- » PRESERVER le patrimoine numérique
- » ARCHIVER l'ensemble des données
- » REPARTIR l'accès aux données

Applications de production de données

Numérisation de conservation



Production administrative

Archivage WEB

...

»

SPAR

Applications de diffusion de données

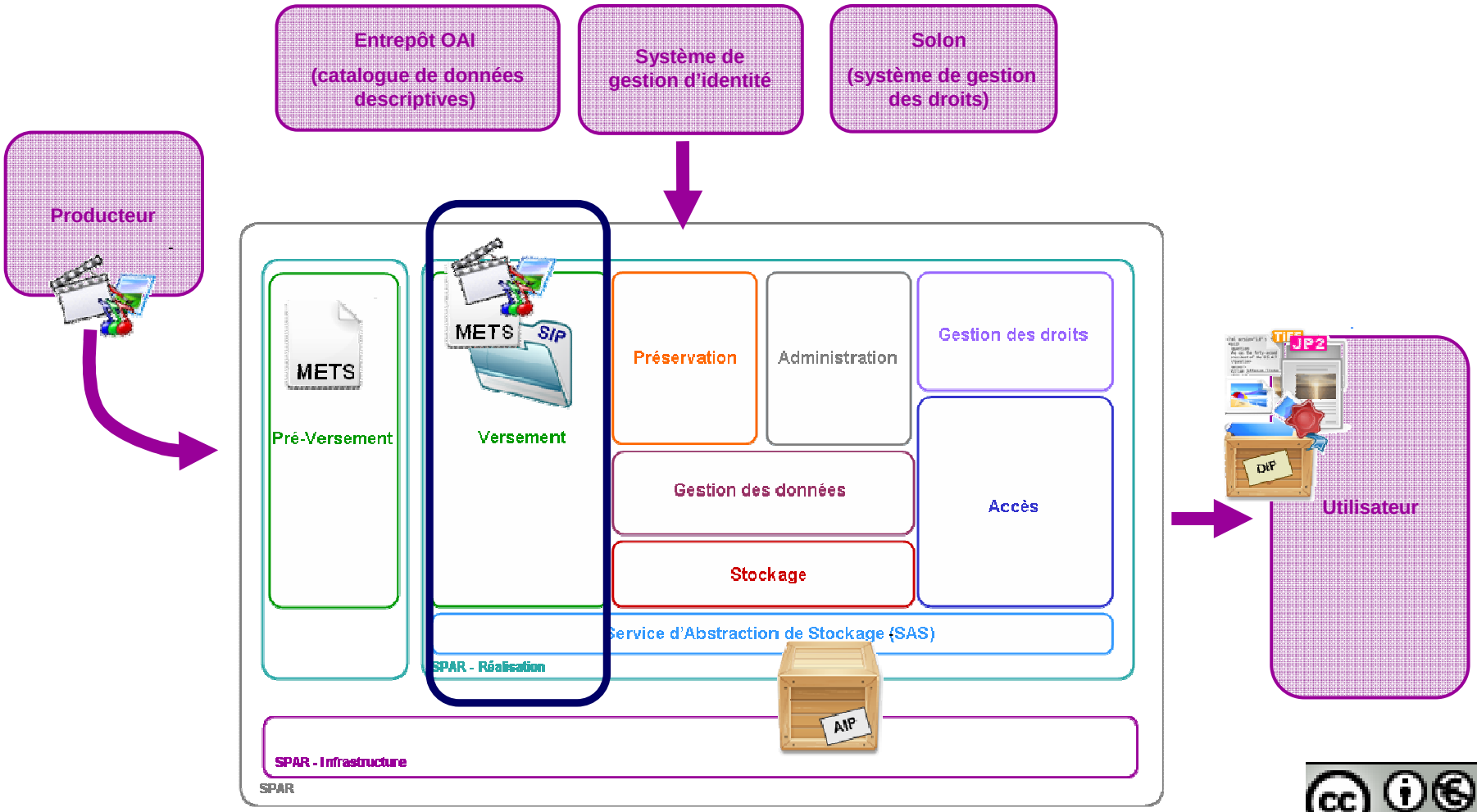
....

....

...

Contexte

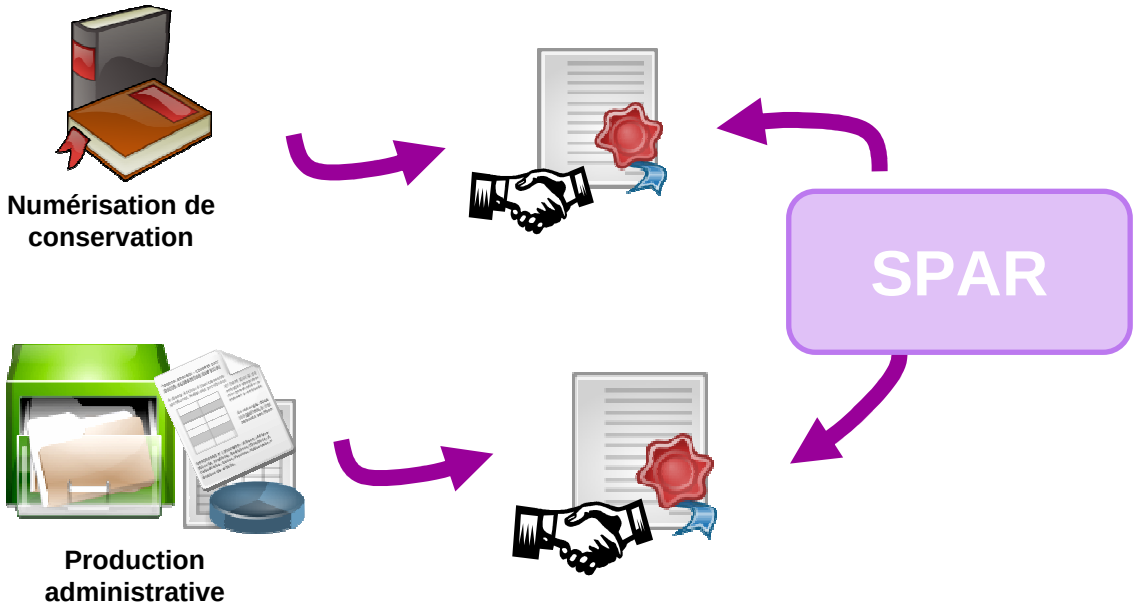
… Un système modulaire contenant des paquets d'information



Contexte

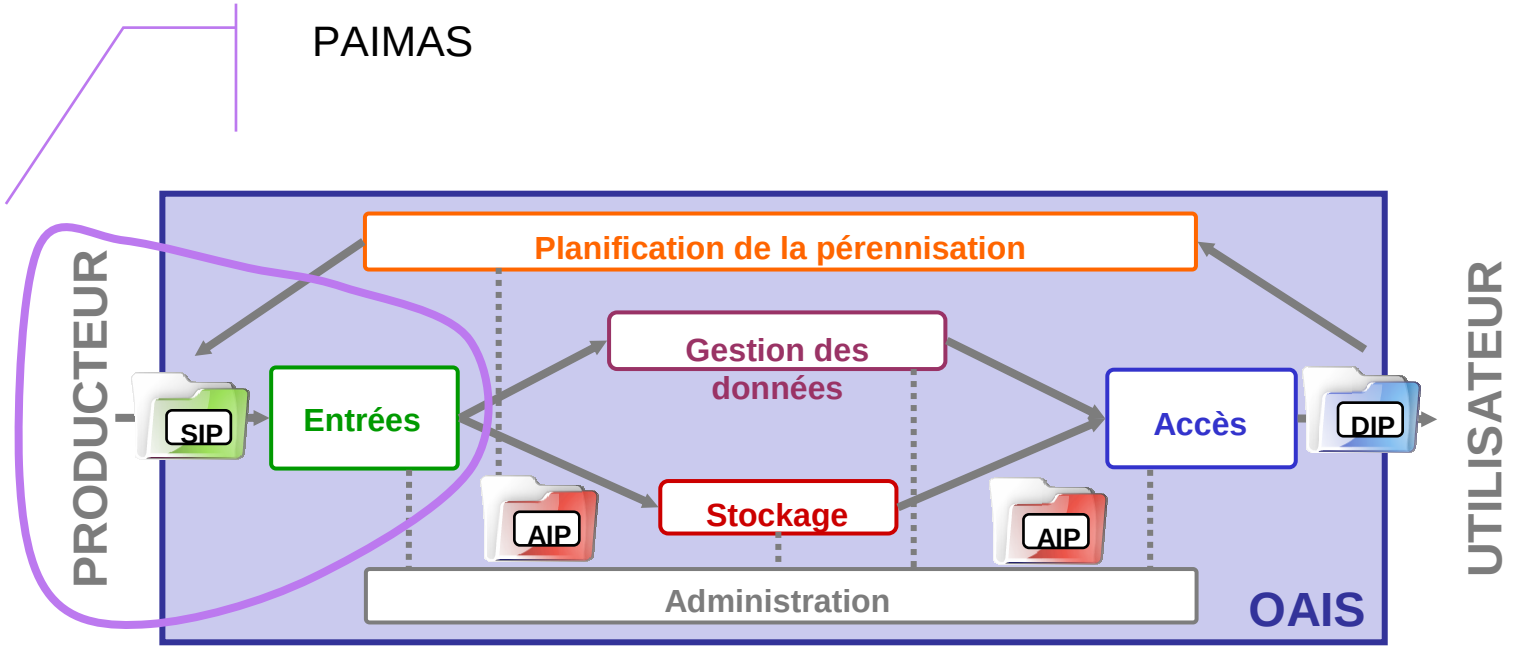
… La double fonction du module Versement

- Un lien entre le Production et le système ou l'Archive
  - Vérifier que les paquets fournis par le producteur sont conformes
  - S'assurer que le système peut garantir la qualité de service promise
- Comment faire? En formalisant dans un **contrat** les engagements et les souhaits du producteur et de l'Archive => Accord sur la qualité de service = SLA en anglais : service level agreement.



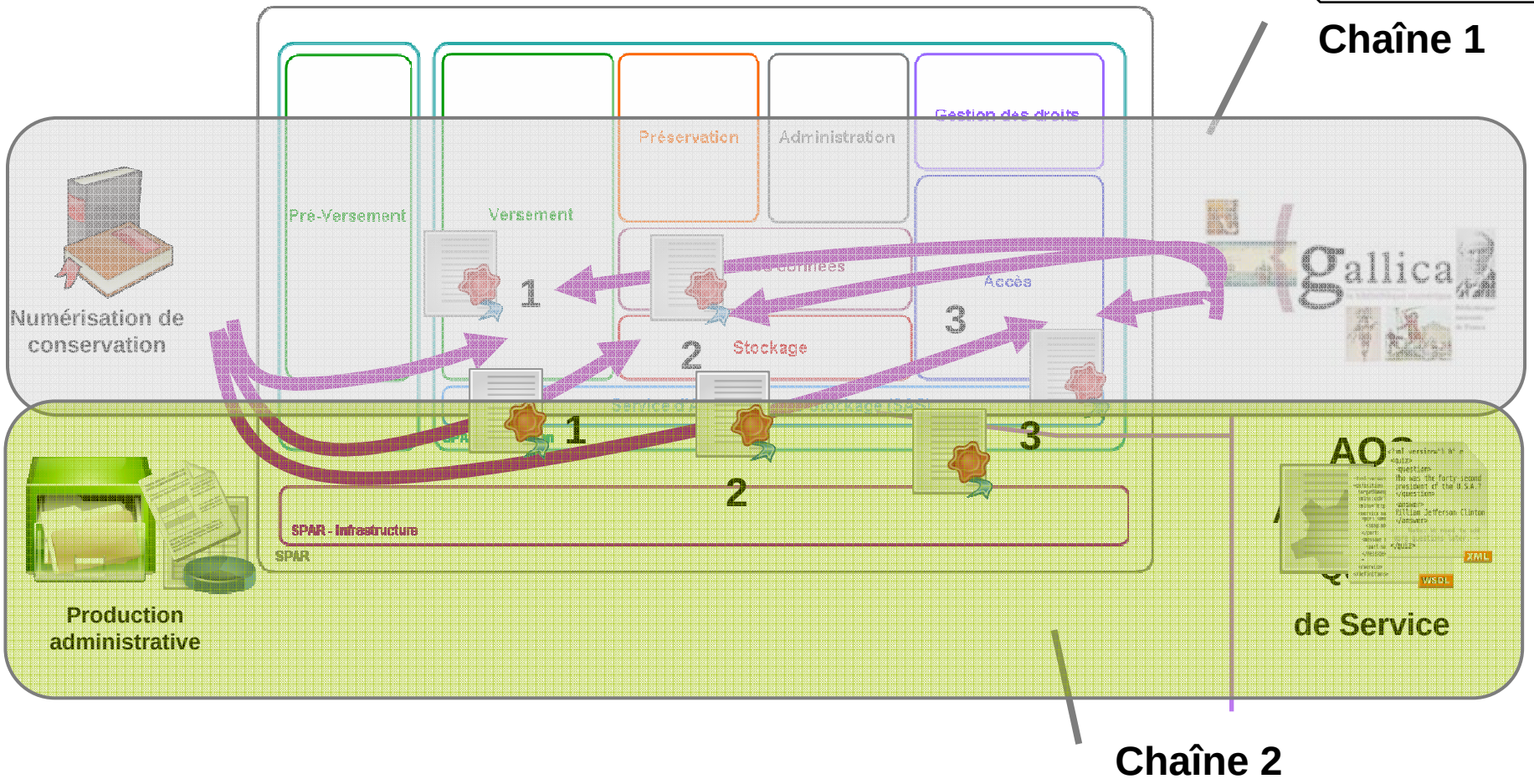
Contexte

… Les accords sur la qualité de service : les normes existantes



Contexte

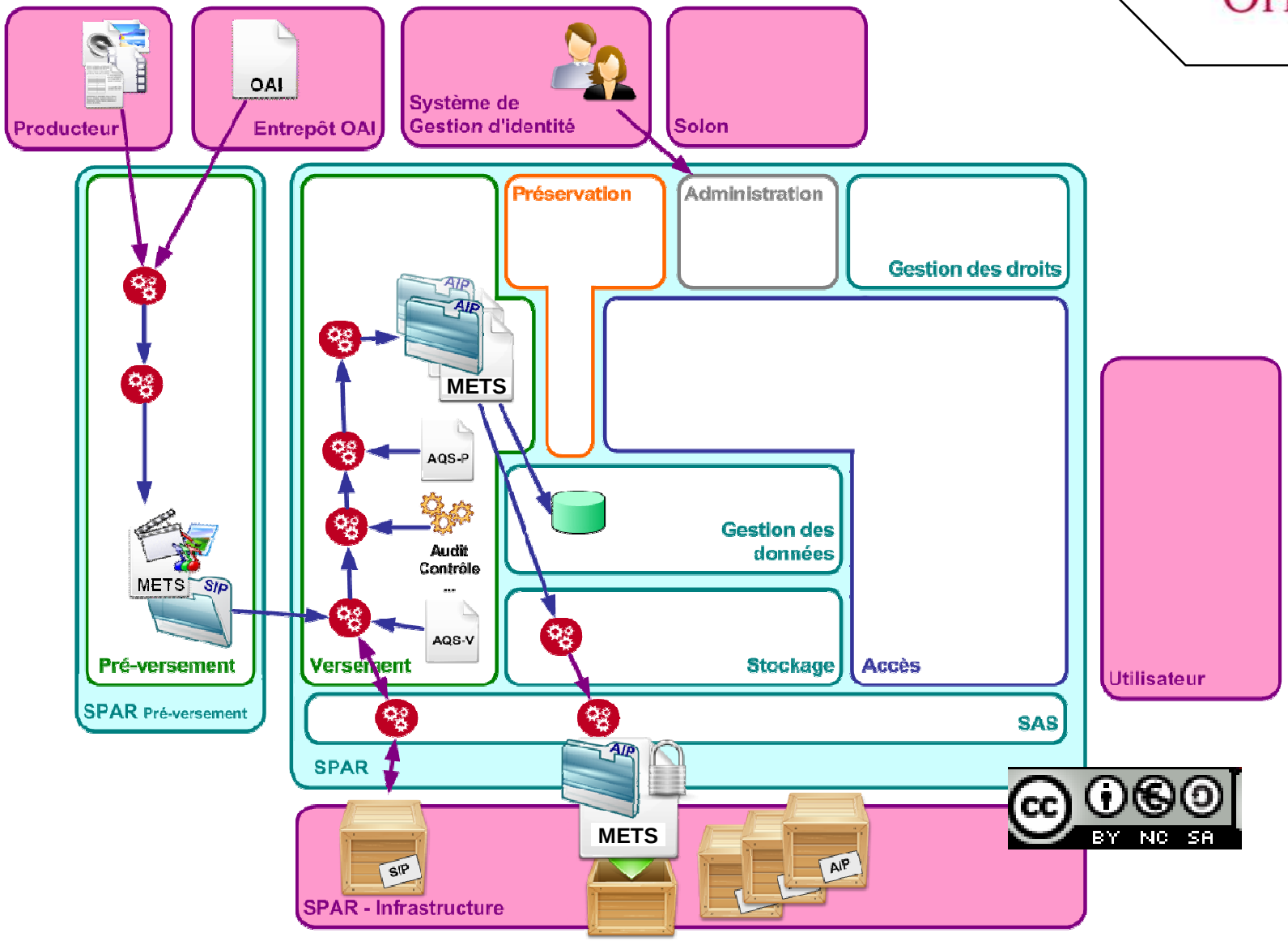
… Les accords sur la qualité de service : périmètre appliqué à SPAR



Etc.

Description générale du module

… Quelques fonctionnalités clés d'un système et organisation pérennes



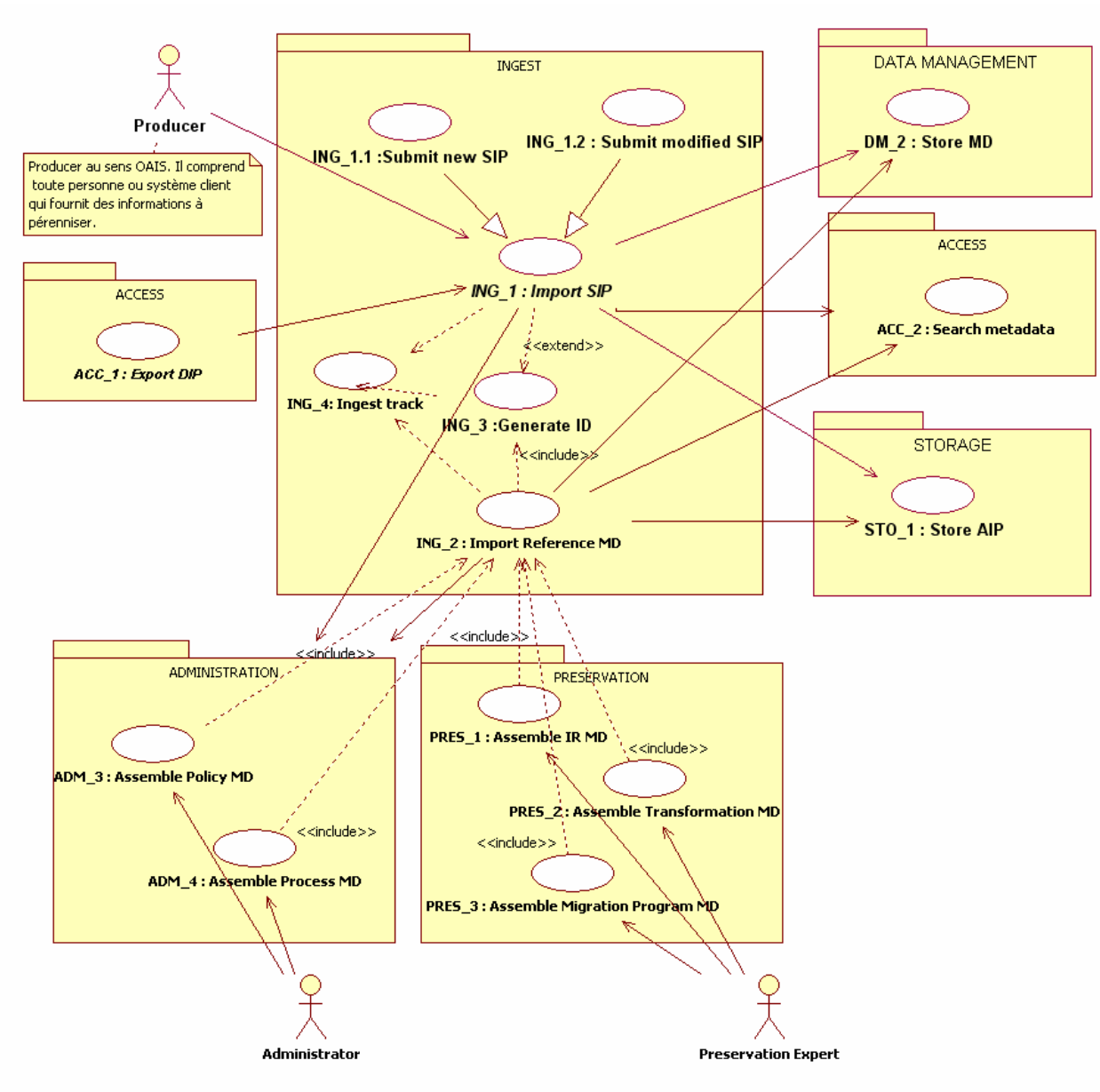
Description générale du module

… Méthodologie

- Cas d'utilisation
  - ING_1 : Import de SIP (avec ou sans mise à jour)
  - ING_2 : Import de description de référence
  - ING_3 : Génération de l'identifiant unique
  - ING_4 : Gestion de la base de suivi

- Actions
  - Un cas d'utilisation est composé de plusieurs actions
  - Une action est une unité de traitement sur le projet dans la description fonctionnelle et dans les développements
    - ACT_01 : Réception de la requête de versement
    - ACT_06 : Audit et caractérisation des fichiers contenus dans un SIP
    - ACT_09 : Finalisation du versement
    - ACT_10 : Mise à jour d'un paquet

Description générale du module … Cas d'utilisation



Description générale du module

… Cas d'utilisation ING_1

Processus du module Versement : enrichissement et validation du paquet

★ **Validation du METS**

★ Validation de chaque fichier du paquet

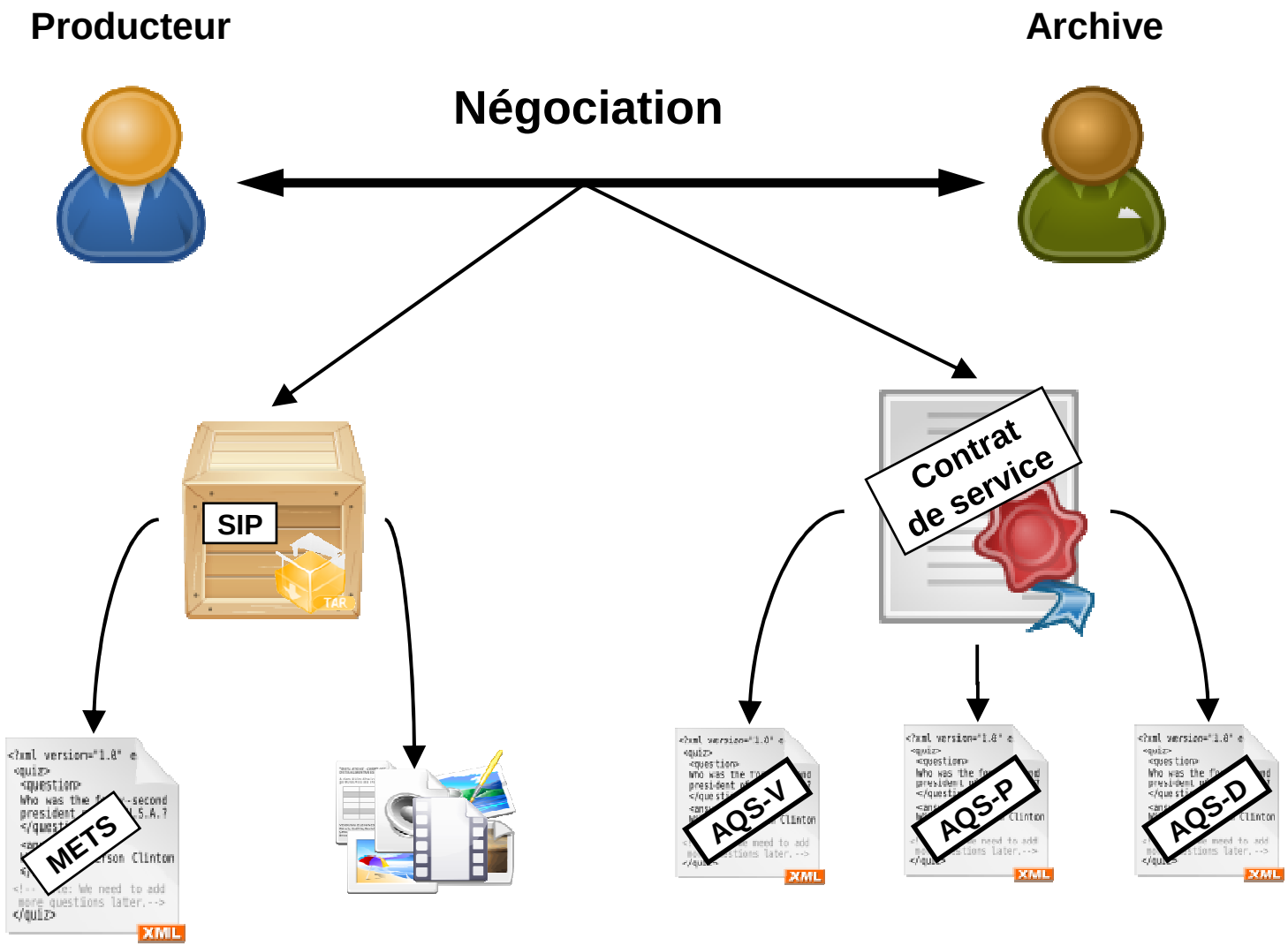
★ Recherche d'un paquet dans SPAR

★ Gestion de la mise à jour

AQS

Modèles de données

… Les différentes données manipulées par le module Versement sont le résultat de la négociation entre le producteur et l'Archive

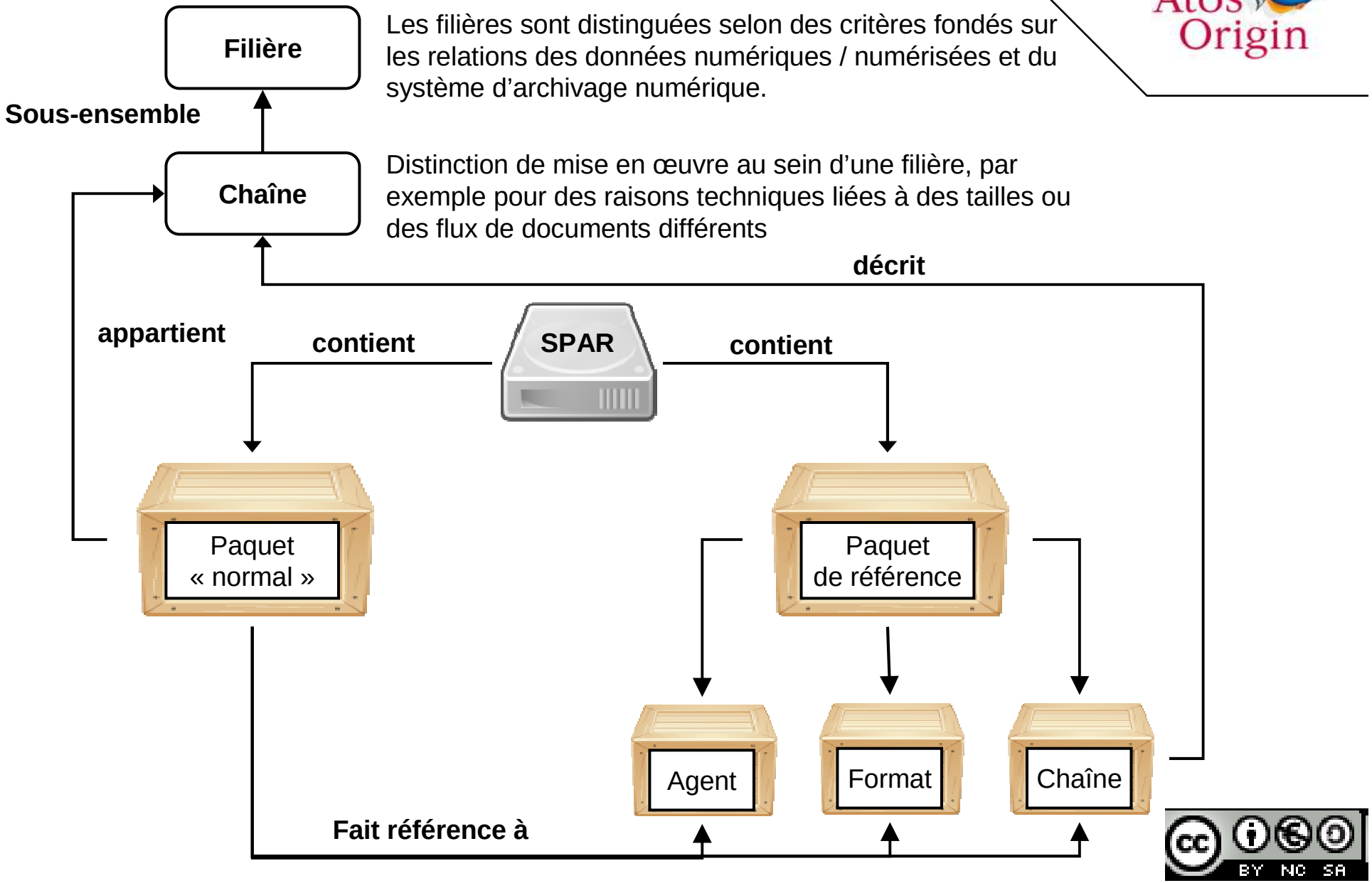


Modèles de données

… Les cadres de la négociation

- Notions générales de SPAR
- Granularité des objets numériques dans SPAR
- Les types de formats de fichiers
- Les critères de l'AQS-V
- Le METS Profile de la BnF
- La politique des identifiants de la BnF

Modèles de données … Les notions générales



Modèles de données

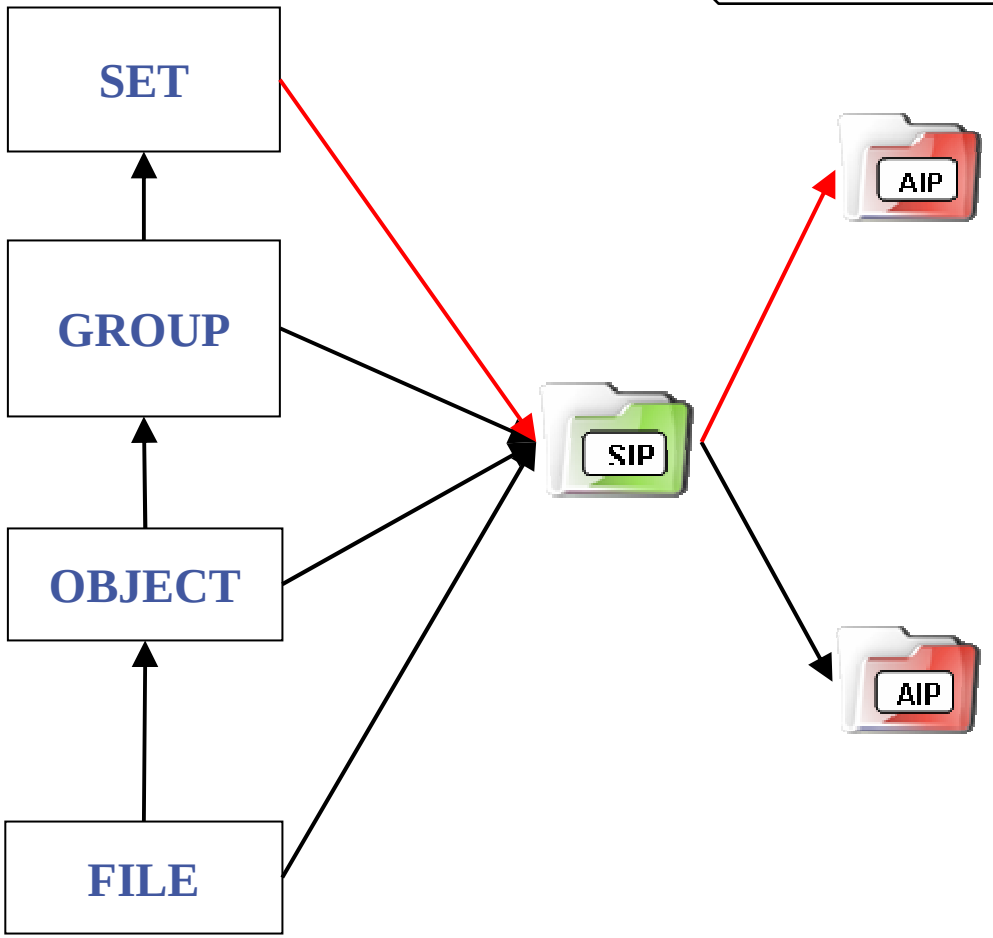
… La granularité des objets dans SPAR

Périodique
Série Monogr. à Plusieurs Vol. (MPV)
Document Multimédia multi-support (MMS)

Volume
Fascicule
Cassette, disque, disque vinyle, planche de diapos, disquette

Page
Fichier html
Face d'un disque

Vue de détail (fichier image)
Fichier texte
Fichier vidéo
Fichier audio
Image (d'une planche de diapo, d'un fichier html)



Un modèle générique pour tous les objets numériques conservés

Modèles de données … Les types de formats de fichiers

La BnF a mis au point une typologie des formats de fichiers. A chaque type sont associés une politique de préservation et des comportements spécifiques de l'archive.

	Stocké	Identifié	Connu	Maîtrisé
Explications	Format dont on ignore les caractéristiques techniques (non identifié) et pour lequel on n'assure que la conservation du train de bits.	Format dont on connaît les caractéristiques techniques (identifiées grâce à un répertoire de formats) mais pour lequel aucun suivi et aucune trajectoire de migration / émulation ne sont prévus. Un format identifié devient maîtrisé ou connu si on met en œuvre une telle trajectoire.	Format non maîtrisable pour lequel la BnF possède au moins un outil de référence, connaît les usages qui en sont faits, sur l'évolution duquel elle assure un suivi et une veille, et pour lequel elle a défini une trajectoire en vue de sa transformation en format maîtrisé ou de son émulation.	Format maîtrisable pour lequel la BnF possède la documentation publiée et au moins un outil de référence, sur l'évolution duquel elle assure un suivi et une veille, et pour lequel elle a défini des contraintes d'application vis-à-vis des producteurs.
Répertoire de formats de SPAR	Non	Non	Oui	Oui
Filière de numérisation de conservation A	Interdit	Interdit	Interdit	» TIFF compressé G4 » TiFF BnF nb » TiFF BnF couleur » JPEG » XML Alto BnF » XML TdMNum » XML TEI

Modèles de données

… La description d'un format dans SPAR

Les formats connus ou maîtrisés sont décrits dans un fichier XML versé dans SPAR au sein d'un paquet de référence et indexé dans le module Gestion de données.

Les différentes informations du fichier de description d'un format :

- **Les propriétés d'identification (extension, type mime, PUID, MagicName…) ;**
- **Le schéma de caractérisation ;**
- **Les propriétés de caractérisation spécifiques à chaque format ;**
- **Les outils d'identification et de caractérisation et leur exploitation.**

Exemple : Le format maîtrisé XML Alto BnF

- **PUID : info:pronom/fmt/101 ;**
- **Type mime : text/xml ou application/xml ;**
- **Schéma de caractérisation : TextMD ;**
- **Version de XML : 1.0 ;**
- **URI de l'espace de nom : http://bibnum.bnf.fr/ns/alto_prod ;**
- **Charset : ISO_8859-1:1987 ou UTF-8 ;**
- **Outil de caractérisation : Jhove + module « XML-hul ».**

Modèles de données

… Les critères de l'AQS_V

Le contrat de service qui lie le producteur et l'archive se concrétise par trois fichiers XML rassemblant respectivement, sous une forme exploitable par un programme, les critères liés au versement, à la préservation et à l'accès.

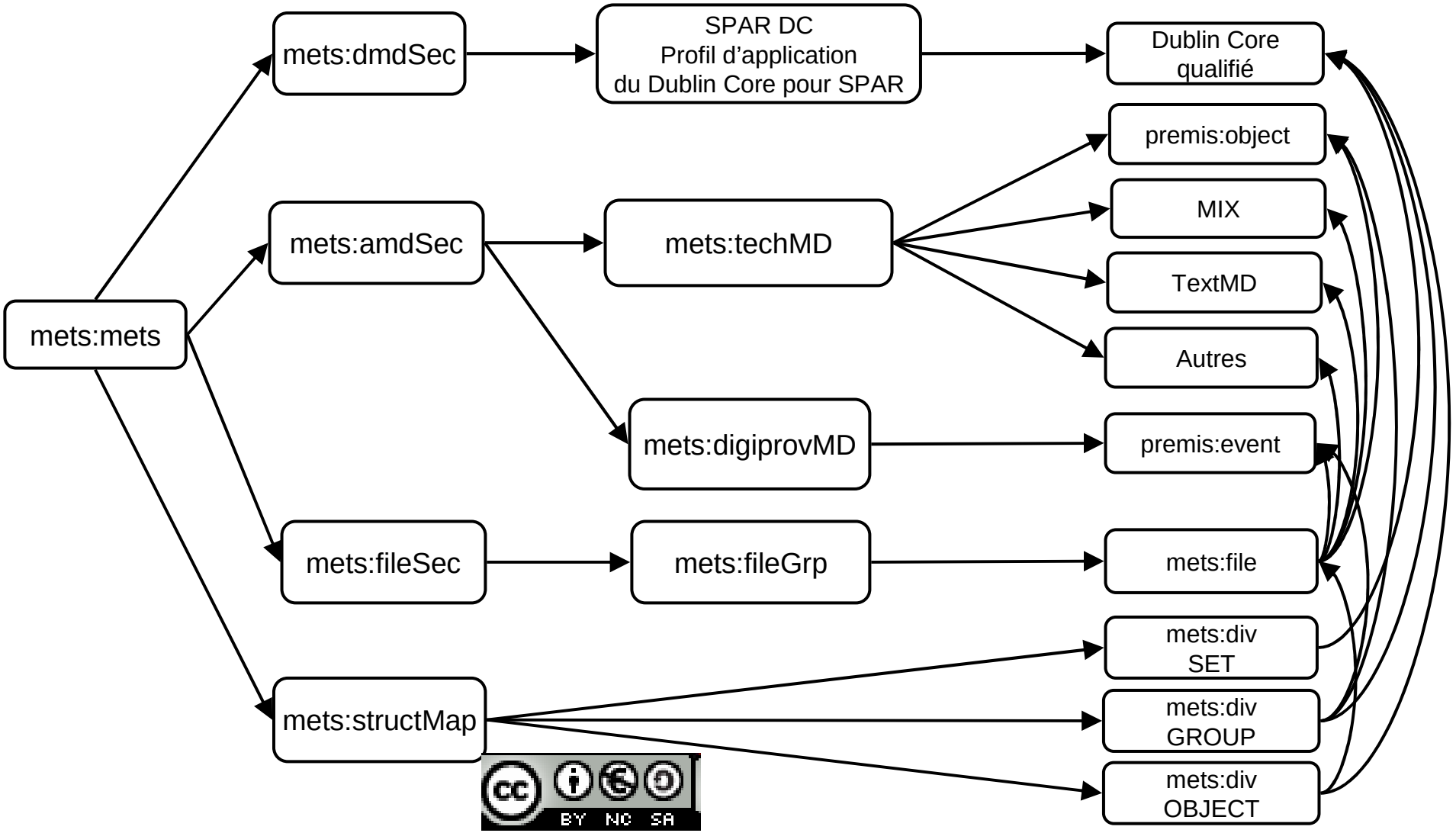
Un AQS-V spécifique à chaque chaîne détermine les critères suivants :

Exigences sur la chaîne	» Horaire d'ouverture et de fermeture » Durée maximum d'indisponibilité
Exigences sur le paquet	» Poids minimum et maximum du paquet » Nombre minimum et maximum de fichiers à l'intérieur du paquet » Les formats refusés et autorisés pour la chaîne et leurs niveaux de préservation
Exigence sur le stockage	» Les paramètres des capsules de stockage pour la copie sécurisée du SIP
Exigences sur le processus de versement en lui-même	» Le nombre minimum et maximum de versement sur une ou plusieurs périodes de temps données ; » Le temps minimum et maximum de prise en compte de la livraison ; » Les utilisateurs autorisés à verser.

Modèles de données

… Le METS Profile de la BnF

METS est le format d'empaquetage choisi par la BnF. Pour encadrer leur utilisation de METS, la BnF a mis au point un METS profile.



Modèles de données

… La politique des identifiants dans SPAR

Trois types principaux d'identifiants sont utilisés dans SPAR pour identifier les différents types de ressources.

Identifiant de production

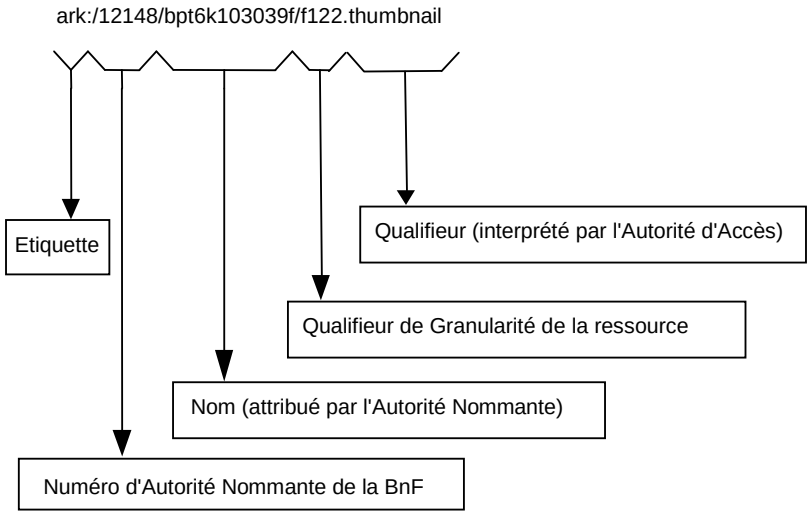
Laissé à l'appréciation du producteur sur le modèle :

Identifiant-du-producteur_identifiant-donné-par-le-producteur

Ressource concernée :

- **SIP**

Identifiant des objets numériques dans SPAR



Ressources concernées :

- **AIP (y compris les paquets de référence)**
- **DIP**

Identifiant des autres types de ressources

info:bnf/spar/ressource

Ressources concernées :

- **Ressources qui ne sont pas des objets numériques, exemple :**
- **Evénements**
- **Prédicat pour les triples RDF**

Description des actions principales

… Les interfaces entre le producteur et l'Archive

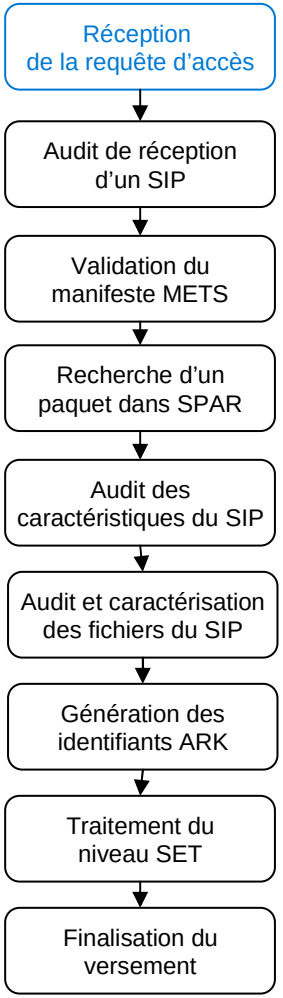
L'ensemble des interfaces entre le producteur et l'archive sont développés selon l'architecture **ReSTFul**, c'est-à-dire une requête HTTP et une réponse en XML.

Ce choix présente plusieurs avantages :

- Respecte strictement la norme et les méthodes HTTP (GET, POST, PUT, DELETE)
- Simple à implémenter et à maintenir
- Développement orienté ressources convient mieux à une archive OAIS que le développement orienté service de SOAP.

Deux ressources définies dans le module Versement

- **Fourniture d'un paquet au module versement**
  URL : http://ingest.spar.bnf.fr/packages/new
  Méthode HTTP : POST
  Arguments : identifier, profile, empreinte, type d'empreinte
  Réponse en XML : identifiant d'un jeton
- **Suivi d'un versement**
  URL : http://ingest.spar.bnf.fr/processes/<idToken>
  Méthode HTTP : GET
  Argument : mode (verbose ou no verbose)
  Réponse en XML : le statut du processus et le détail de chaque action



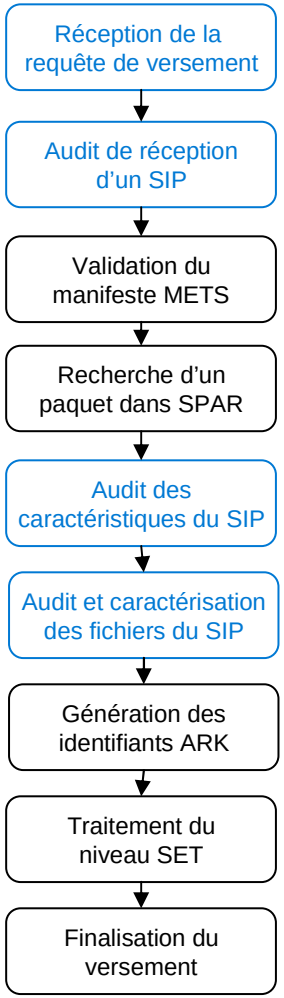
Description des actions principales … La vérification des critères de l'AQS_V

L'ensemble des critères de l'AQS-V sont vérifiés au cours du versement dans quatre actions spécifiques :

- Exigences de stockage (Réception de la requête de versement)
- Exigences sur la chaîne (Audit de réception d'un SIP)
- Exigences sur le processus de versement (Audit de réception d'un SIP)
- Exigences sur le paquet (Audit des caractéristiques du SIP et Audit et caractérisation des fichiers du SIP)

Les fichiers XML d'AQS sont indexés au sein d'une base de données RDF (Triple Store) après une conversion en RDF/XML. Le triple store est interrogé avec le langage de requête SPARQL, recommandation du W3C.

Sujet	Prédicat	Objet
ark:/12148/b123456789/r1	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	info:bnf/spar/context#ingestSLA
ark:/12148/b123456789/r1	info:bnf/spar/context#applicableFor	info:bnf/spar/context/fil_num_cons_A
ark:/12148/b123456789/r1	info:bnf/spar/context#definesRequirement	blank node r1221519205r3953r4
blank node r1221519205r3953r4	http://www.w3.org/1999/02/22-rdf-syntax-ns#type	info:bnf/spar/context#packageRequirement
blank node r1221519205r3953r4	info:bnf/spar/context#name	maxNumberOfFiles
blank node r1221519205r3953r4	http://www.w3.org/1999/02/22-rdf-syntax-ns#value	32



Description des actions principales … La vérification des critères de l'AQS_V

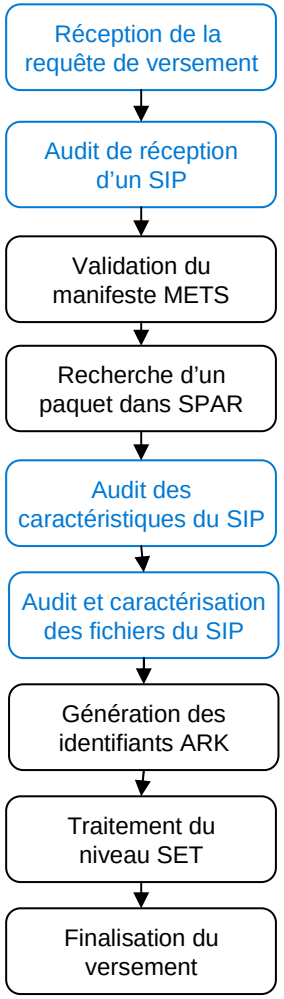
L'ensemble des critères de l'AQS-V sont vérifiés au cours du versement dans quatre actions spécifiques :

- Exigences de stockage (Réception de la requête de versement)
- Exigences sur la chaîne (Audit de réception d'un SIP)
- Exigences sur le processus de versement (Audit de réception d'un SIP)
- Exigences sur le paquet (Audit des caractéristiques du SIP et Audit et caractérisation des fichiers du SIP)

Les fichiers XML d'AQS sont indexés au sein d'une base de données RDF (Triple Store) après une conversion en RDF/XML. Le triple store est interrogé avec le langage de requête SPARQL, recommandation du W3C.

Requête : je recherche l'exigence du nombre maximum de fichiers dans l'AQS-V de la chaîne filière de numérisation de conservation A (fil_num_cons_A)

```
SELECT ?maxnumberoffiles
WHERE
{
 ?ingestSLA  a <info:bnf/spar/context#ingestSLA>;
             <info:bnf/spar/context#applicableFor> <info:bnf/spar/context/fil_num_cons_A>;
             <info:bnf/spar/context#definesRequirement>?requirement.
?requirement <info:bnf/spar/context#name> 'maxNumberOfFiles';
              <http://www.w3.org/1999/02/22-rdf-syntax-ns#value> ?maxnumberoffiles.
 }
```



Description des actions principales

… La validation du manifeste METS

Deux système de validation sont utilisés pour valider le METS du SIP :

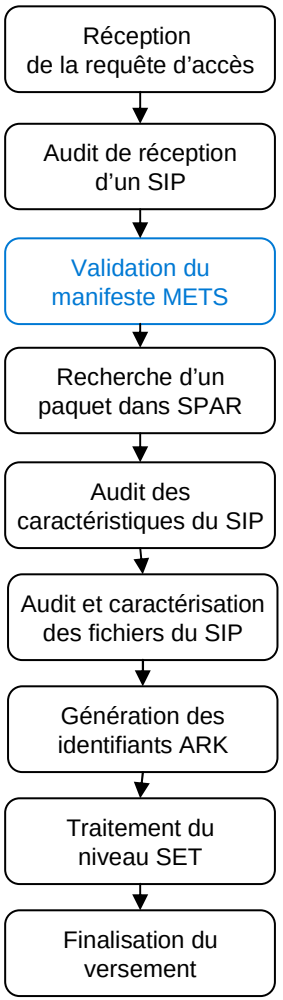
- Validation par les schémas XML utilisés par le manifeste METS (Mets, spar_dc, Dublin Core elements, Dublin Core terms, Premis V2)
- Schematron, norme ISO pour valider plus précisément un fichier XML

Pourquoi Schematron ?

- METS est un schéma XML générique dont les profils d'applications varient selon l'usage, la validation avec les schémas est insuffisant ;
- METS profile n'est pas « machine actionable », il faut donc trouver un mécanisme pour contrôler son respect
- Le grain de validation peut être très fin, par exemple, la valeur précise d'un argument (le type de division, entre autres)
- Schematron est un fichier XML ce qui en facilite la préservation
- Schematron est une norme ISO

Principes de Schematron :

- Définition de règle (ou « Pattern ») indiqué sous la forme de requête Xpath que le fichier doit suivre
- Définition de test pour chaque règle et d'un comportement associé



Description des actions principales

… Le traitement des fichiers

Identification d'un fichier (Type mime ou magic number)

Choix d'un outil de caractérisation en fonction de l'identification

Interrogation du module Gestion de données

Caractérisation du fichier

Recherche des propriétés de caractérisation spécifiques

Interrogation du module Gestion de données

Extraction des valeurs des propriétés du fichier issu de la caractérisation

Recherche du format dans le répertoire de format

Interrogation du module Gestion de données

Vérification dans l'AQS-V si le format est autorisé

Interrogation du module Gestion de données

Ajout des métadonnées dans le manifeste METS

Réception de la requête d'accès

Audit de réception d'un SIP

Validation du manifeste METS

Recherche d'un paquet dans SPAR

Audit des caractéristiques du SIP

Audit et caractérisation des fichiers du SIP

Génération des identifiants ARK

Traitement du niveau SET

Finalisation du versement

Description des actions principales

… Le traitement des fichiers

text/xml

Jhove +
module XML Hul

Un fichier XML
au format textMD issu
du fichier XML de Jhove

Interrogation du
module Gestion de
données

Version de XML
Espace de nom
Charset

1.0
http://bibnum.bnf.fr/ns/alto_prod
UTF-8

ark:/12148/fXML_ALTOs

Interrogation du
module Gestion de
données

Interrogation du
module Gestion de
données

XML-ALTO est un format
maîtrisé autorisé par l'AQS
de fil_num_cons_A

2 techMD
1 digiprovMD
Lien avec le <file>

Interrogation du
module Gestion de
données

Réception
de la requête d'accès

Audit de réception
d'un SIP

Validation du
manifeste METS

Recherche d'un
paquet dans SPAR

Audit des
caractéristiques du SIP

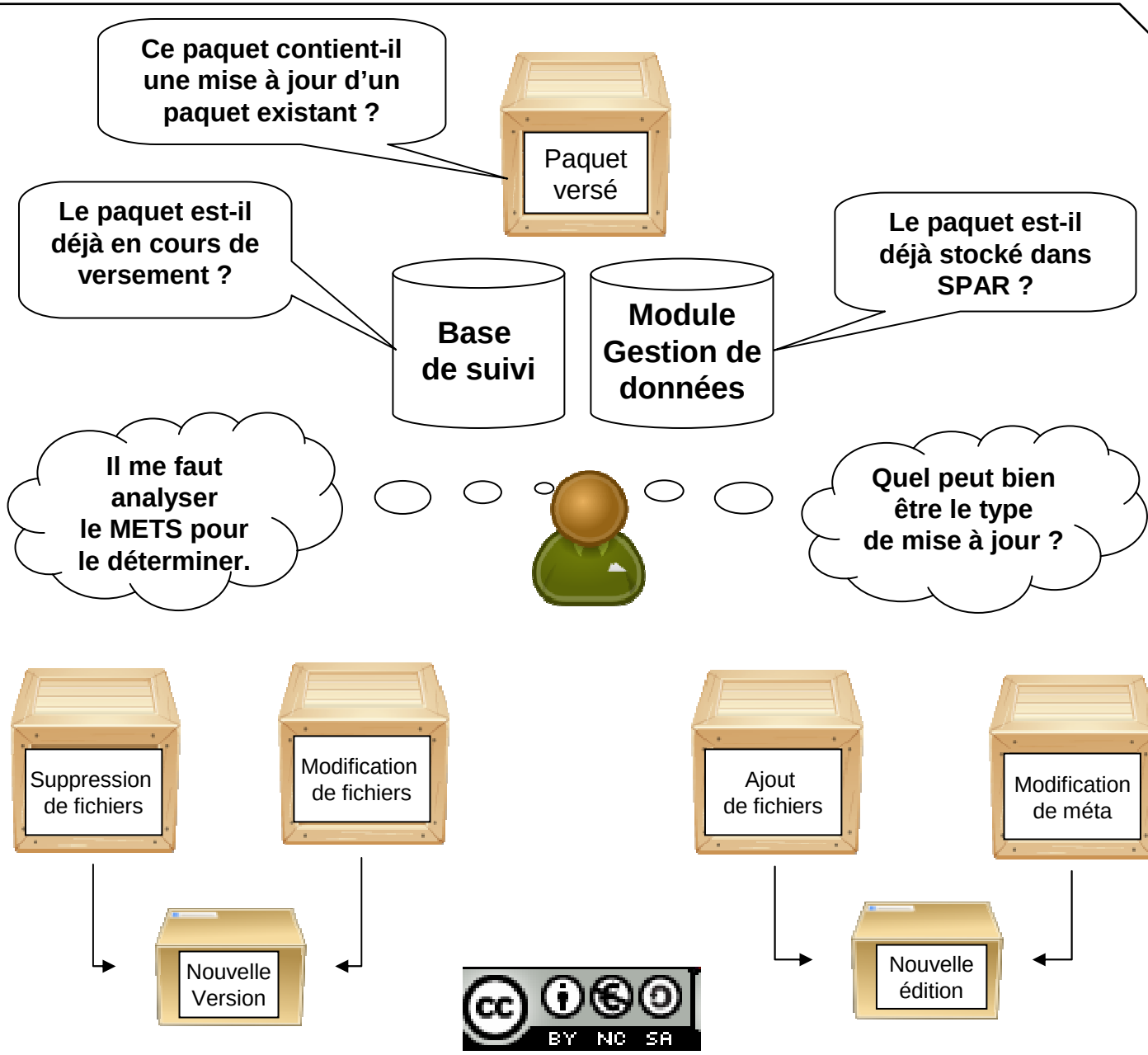
Audit et caractérisation
des fichiers du SIP

Génération des
identifiants ARK

Traitement du
niveau SET

Finalisation du
versement

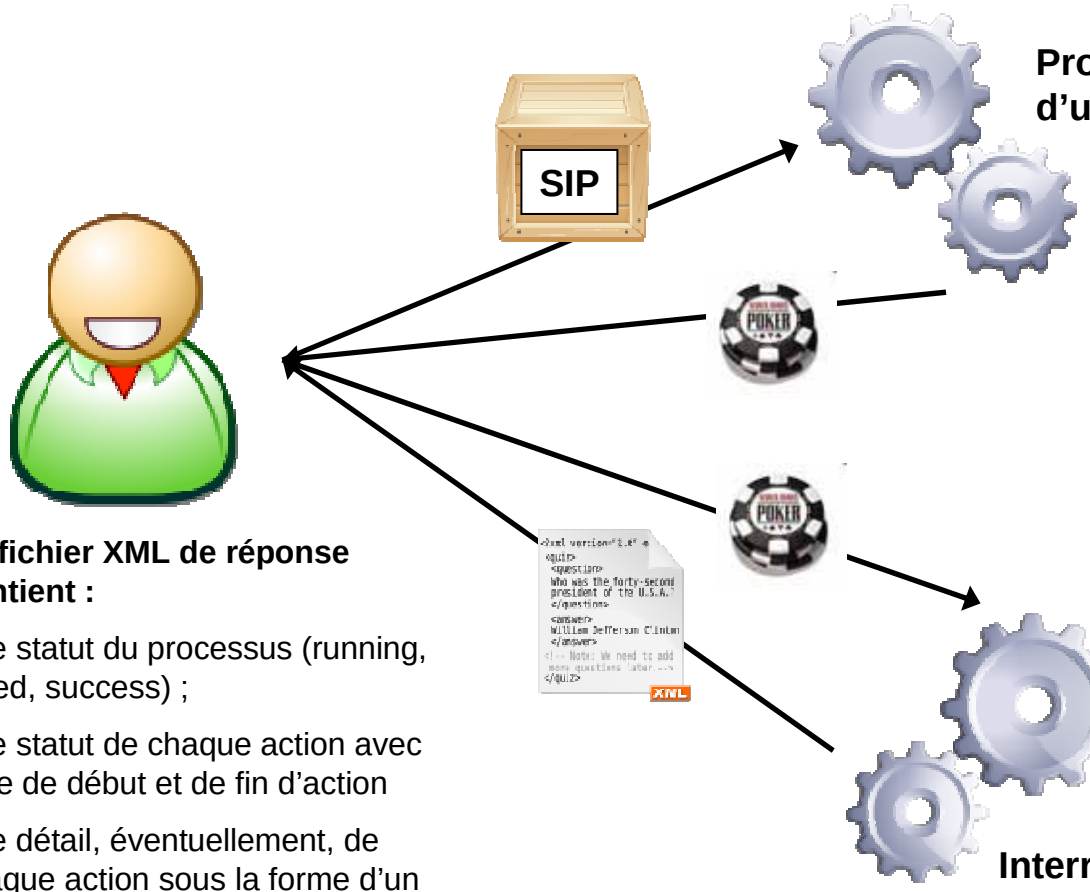
Description des actions principales … La mise à jour des fichiers



- **Réception de la requête d'accès**
- Audit de réception d'un SIP
- Validation du manifeste METS
- Recherche d'un paquet dans SPAR
- Audit des caractéristiques du SIP
- Audit et caractérisation des fichiers du SIP
- Génération des identifiants ARK
- Traitement du niveau SET
- Finalisation du versement

Description des actions principales… Le système de suivi

Pour des raisons de performance, le traitement d'un SIP est asynchrone, c'est-à-dire que le producteur n'est pas mis en attente le temps de l'accomplissement du versement. Il est donc nécessaire de mettre en place un système de suivi avec un mécanisme de jeton.

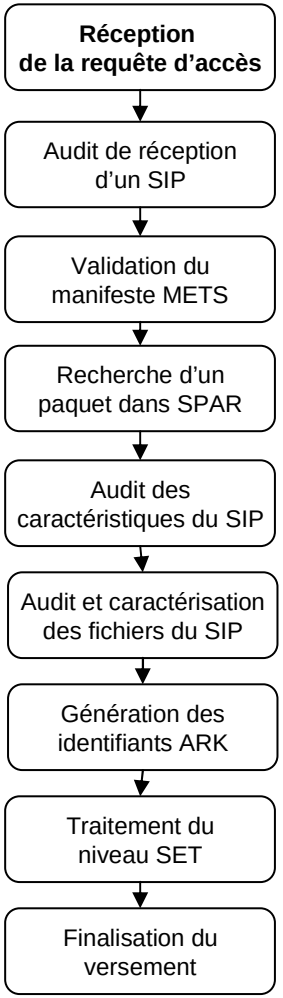


Processus de versement d'un paquet

Interrogation du suivi d'un processus de versement

Le fichier XML de réponse contient :

- Le statut du processus (running, failed, success) ;
- Le statut de chaque action avec date de début et de fin d'action
- Le détail, éventuellement, de chaque action sous la forme d'un premis:event, y compris le détail et un code d'erreur, dans le cas d'une erreur



L'avancement

- Le workflow de traitement du module Versement a été découpé en 5 cas d'utilisation et plus de 10 actions différentes.
- Ce découpage permet de réaliser une application en composants.
- A ce jour, la majorité des actions à été en partie ou complètement développée.
- A chaque action est associé un algorithme ou un composant technique spécifique. La majeure partie des composants techniques sont issus de bibliothèques Open Source.
- Deux point de visibilité destinés au client à ce jour :
  - Le dernier : le 9 septembre 2008 : Présentation du module dans un environnement d'intégration.

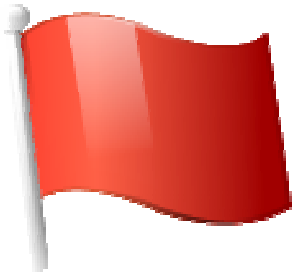
Les difficultés

» L'exploitation d'un grand nombre de bibliothèques Open Source implique une grande rigueur quant à la gestion des dépendances au sein du code source (langage Java).

- Un temps non négligeable est nécessaire pour valider la compatibilité des composants les uns vis-à-vis des autres.

» Les bibliothèques retenues initialement n'ont pas obligatoirement répondu aux contraintes de respect des standards.

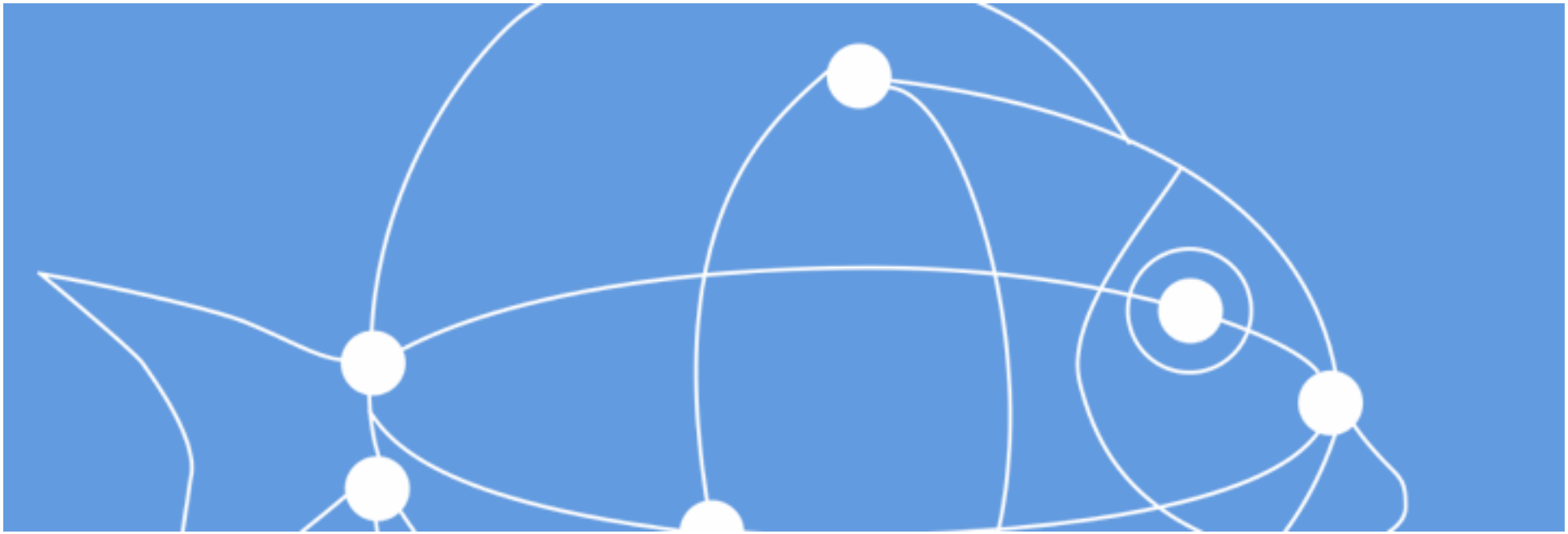
- Exemple : Le framework de gestion du protocole Service Web selon le modèle REST a du être changé pour assurer la prise en compte de la JSR 311.



Les points à retenir

- Le module Versement est un élément clé du système SPAR. Il convient de spécifier avec exhaustivité et rigueur l'aspect fonctionnel et l'architecture des données. De ces spécifications découle une grande partie de la structure applicative de la solution logicielle.
- De nombreux composants Open Sources permettent de mettre en œuvre tout ou partie des actions constituant le module Versement.
- Une méthodologie de type « Agile » permet d'assurer la construction de SPAR. Cette méthode permet d'associer plus facilement le client aux développements de la solution et de rapprocher les experts fonctionnels, l'architecte des données et les développeurs. Ainsi, l'effet tunnel est évité et le résultat est au plus proche du besoin exprimé.

Questions ?



Vos contacts :

Gautier Poupeau **+33 (0)6 62 03 86 35**

EIM Consultant gautier.poupeau@atosorigin.com

Charlotte Fabre **+33 (0)6 28 42 00 99**

EIM Consultant charlotte.fabre@atosorigin.com