Laboratoire ligérien
de linguistique

Bibliothèque nationale de France

UNIVERSITE D'ORLEANS

Retour d'expérience sur l'utilisation du format SIARD pour l'archivage des bases de données relationnelles

Michel Jacobson, le 6 novembre 2014, « Journée de sensibilisation à la sécurisation et à la pérennisation des données », rBDD

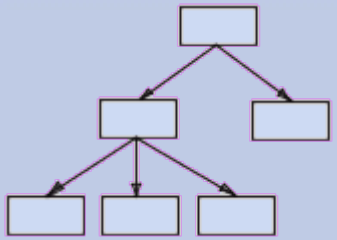
Plan

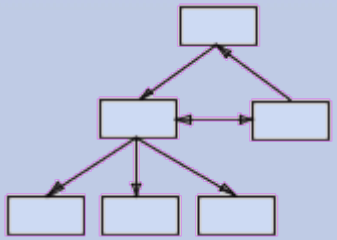
- Le contexte : les bases de données relationnelles
- Le format SIARD
- Exemple d'utilisation
  - La matrice cadastrale numérique

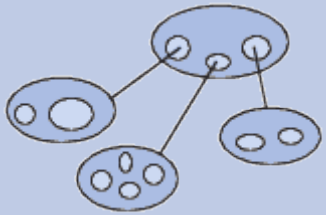
Les BDD relationnelles

BDD-R

- Les bases de données
  - Ensembles organisés d'informations exprimées dans un modèle formel et qui représentent des informations du monde réel ou des idées.
  - Ces informations peuvent être interrogées et mises à jour par une communauté d'utilisateurs à l'aide d'outils appelés des SGBD.
- Les bases de données relationnelles
  - Elles se caractérisent par leur Modèle de Description des Données (MDD) appelé le « modèle relationnel »
    - Par opposition à d'autres modèles : hiérarchique, réseau, objet..







SGBD-R

- Pas de définition d'un format de représentation partagé
- Définition d'un langage normalisé
  - SQL « Structured Query Language » ISO:1987
  - Définit un langage de description de données (LDD)
    - Le **domaine** d'une donnée est l'ensemble des valeurs qu'elle peut prendre : nombre, chaîne de caractères, date, booléen.
      - ➔ *INTEGER [entier de -2147483648 à 2147483647], DATE [calendrier grégorien]*
    - Des **contraintes** sur les valeurs peuvent être définies
      - ➔ *PRIMARY KEY [clé primaire], NULL [Autorise la valeur null])*
  - Définit un langage de manipulation de données (LMD)
    - Insertion (INSERT INTO...) ; Modification (UPDATE...) ; Suppression (DELETE...) ; Sélection (SELECT attributs FROM tables WHERE conditions)

Comment archiver des BDD relationnelles ?

Archivage des BDD relationnelles

- Objectif
  - Conserver les informations présentes dans des BDD en dehors du SGBD et au-delà de sa durée de vie
- La solution classique
  - Avoir une structure de représentation (format de fichier) décrivant les informations
  - Ce format de fichier doit avoir des qualités (spécifié, public, ouvert, normalisé, sans entrave légale, auto-porteur, outillé, utilisé, simple...) de manière à pouvoir s'assurer de la lisibilité à long terme des informations qu'il permet d'exprimer
- Solution alternative
  - Conserver l'environnement d'exploitation (émulation)

Archivage des BDD relationnelles

- Problème : il n'existe pas de format de représentation
- Mais des solutions d'export (SGBD en fonctionnement):
  - Les Dumps : solution pour une sauvegarde technique.
    - Cas d'utilisation : Rechargement en cas de corruption, changement de version de SGBD ou migration système.
    - Permet d'exprimer toutes les informations du SGBD (y compris droits, privilèges procédures, etc.), formats propriétaires.
  - L'export dans des fichiers plats
    - CSV, tabulaire, XML (ne permet d'exprimer que les données)
  - L'export en SQL
    - Problème de compatibilité entre SGBD dans leurs implémentations de SQL
  - L'export en RDF
    - Les recommandations du W3C « Direct Mapping » et « R2RML »

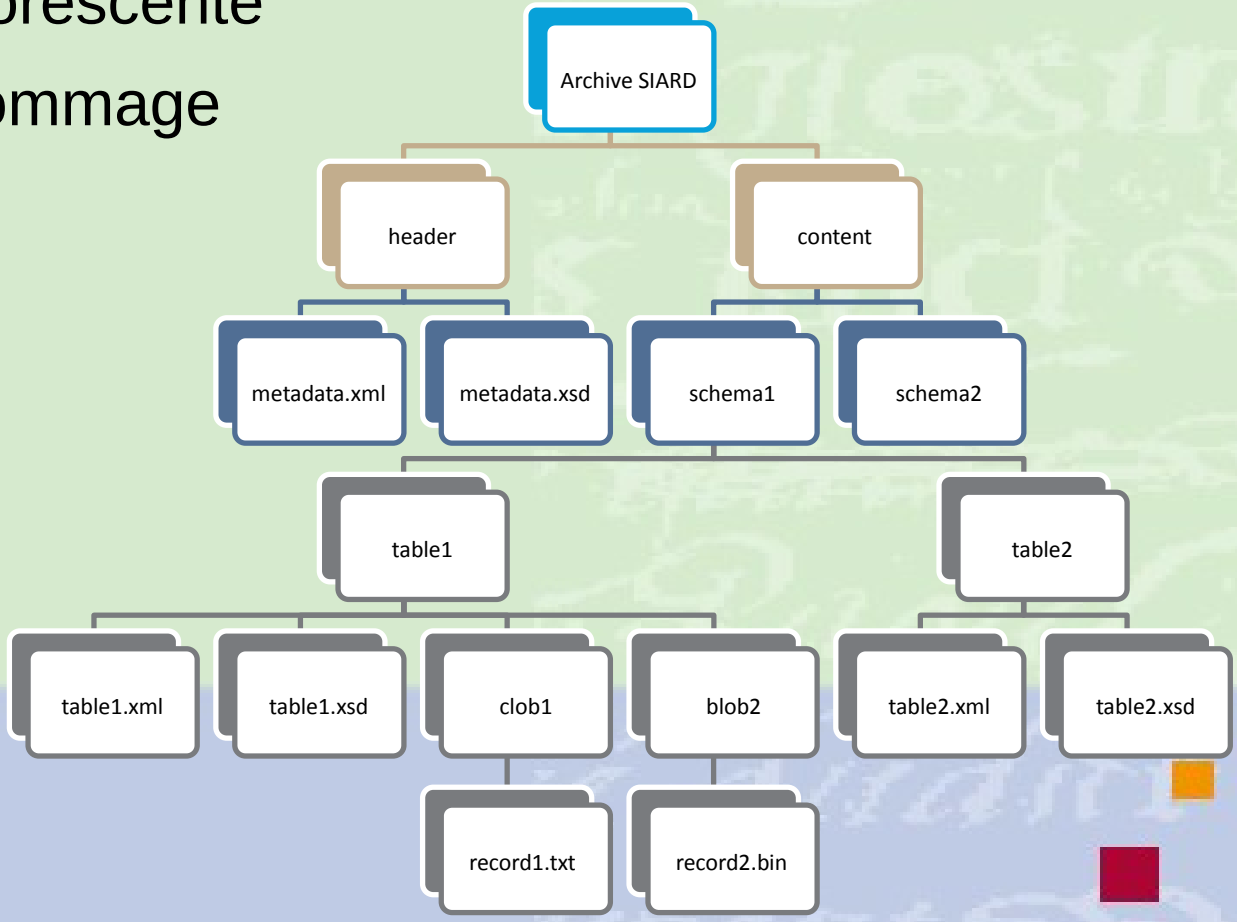
Le format SIARD

Le format SIARD

- Software-Independent Archiving of Relational Databases
- C'est un format de fichier
  - Le format a été définit par les archives fédérales Suisses (sept 2008). Les spécifications sont publiques et basées sur des formats bien connus :
    - XML pour les contenus
    - ZIP pour l'enveloppe (ZIP64)
    - SQL:99 pour les types d'informations
  - Le format permet de réunir dans un même conteneur, les données, leur structure et leur documentation.

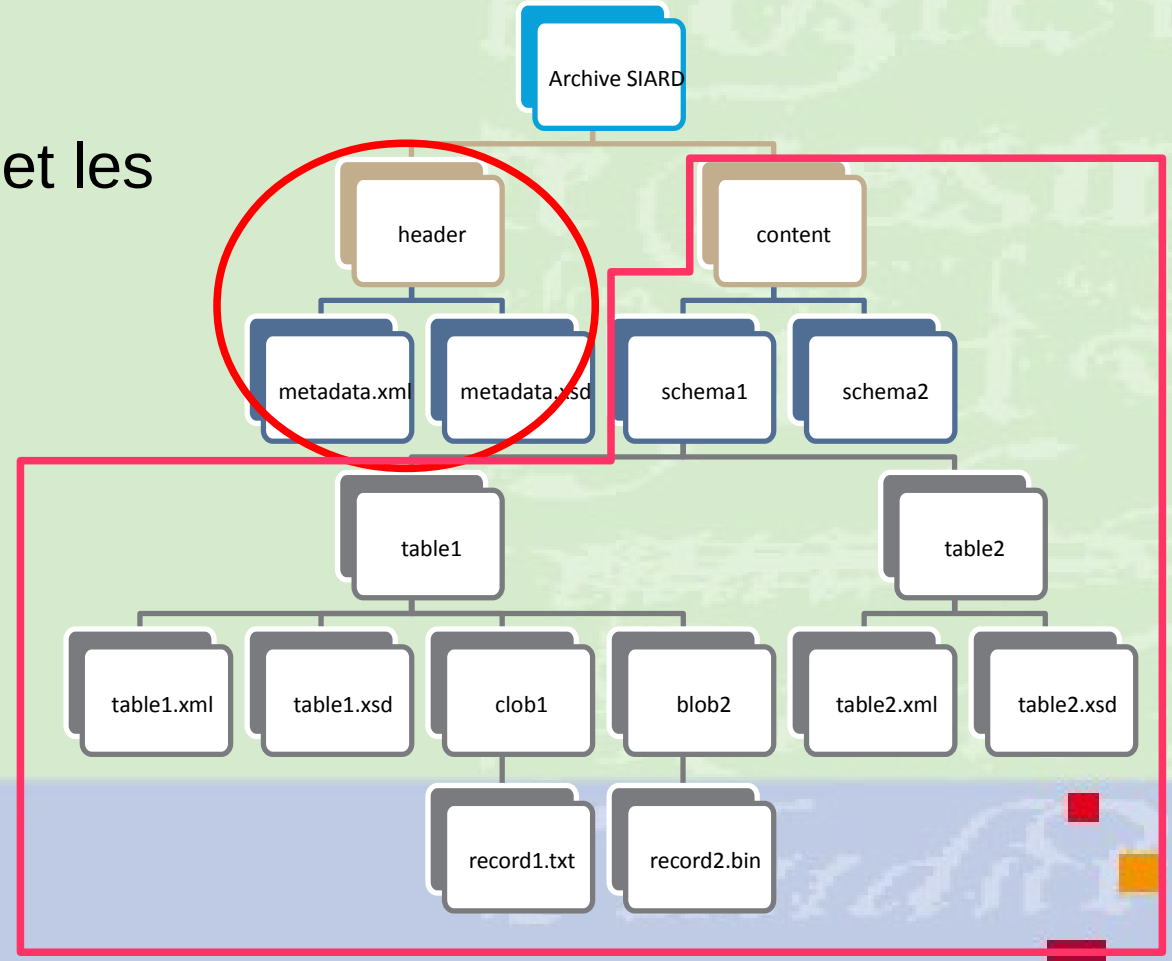
Structure d'un fichier SIARD

- Les informations sont organisées
  - Dans une structure arborescente
  - Respecte un plan de nommage

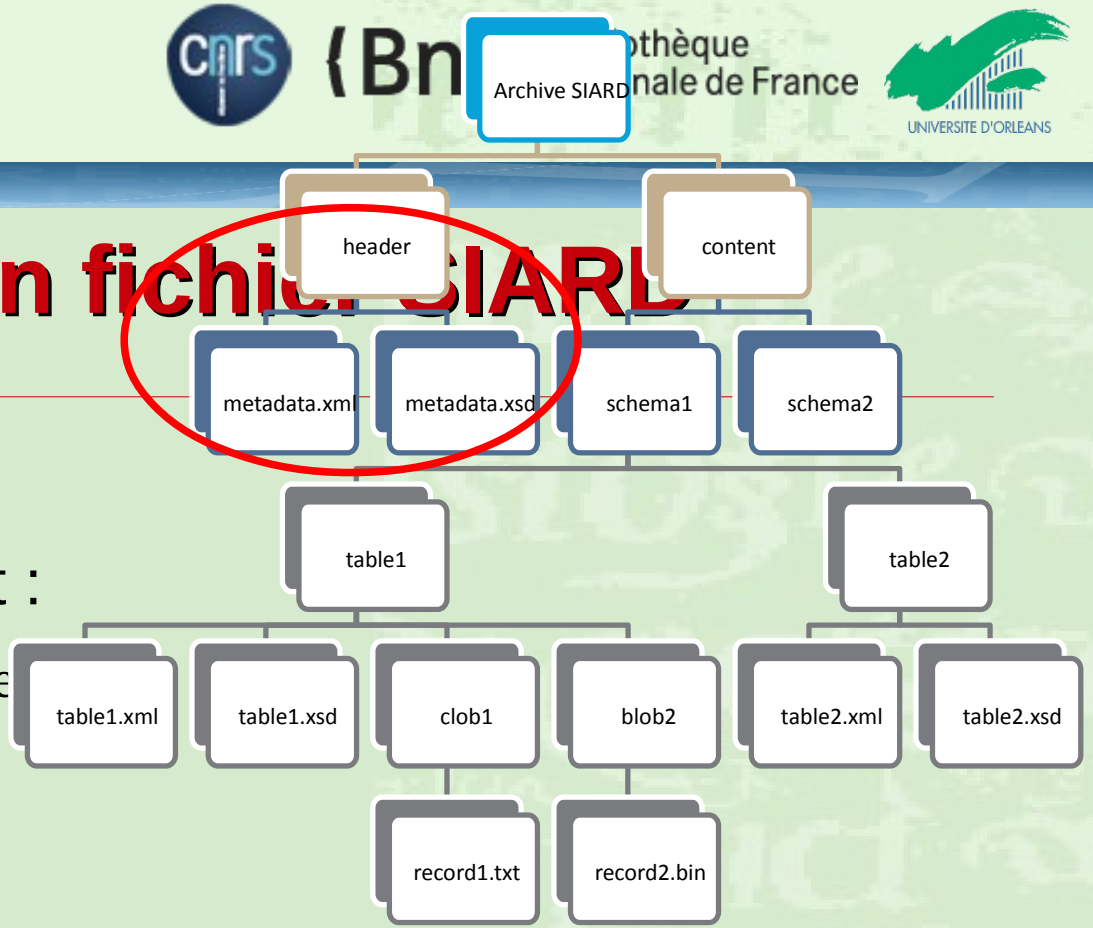


Structure d'un fichier SIARD

- 2 dossiers obligatoires:
  - « header » contient les informations de structure et les métadonnées ;
  - « content » contient les données de la base



Structure d'un fichier SIARD



- « header » comporte :
  - Le fichier metadata.xml décrit :
    - La répartition des fichiers du dossier « content » (schémas/tables/...)
    - La structure de chaque table
      - ➔ *Colonnes, types, contraintes...*
  - Le fichier metadata.xsd décrit la syntaxe du fichier précédent
  - Un fichier metadata.xsl pour faciliter la visualisation

Structure d'un fichier SIARD

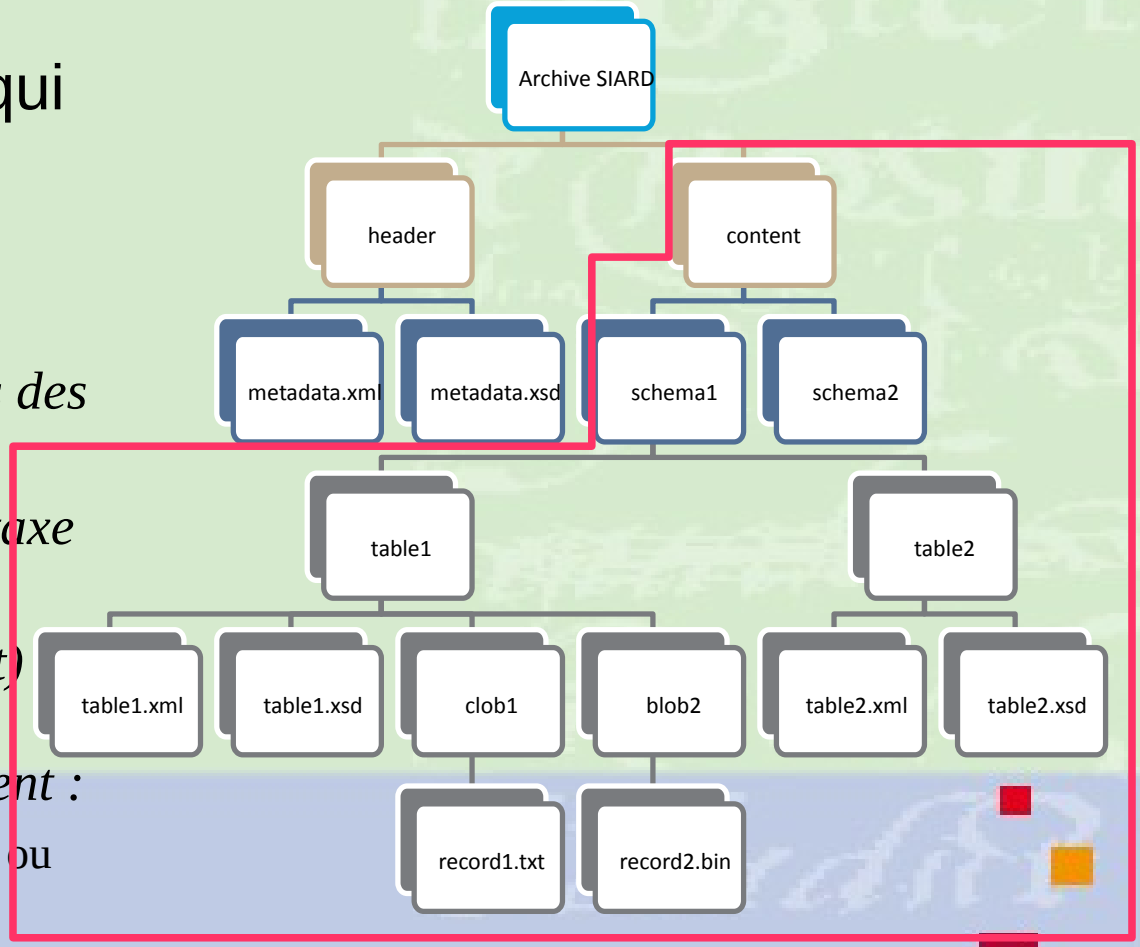
- La structure arborescente du XML reproduit la structure arborescente des fichiers du dossier « content » (schémas/tables)
  - Pour chaque fichier (=table), sont décrits :
    - Les colonnes
      - ➔ *nom*
      - ➔ *type [domaine]*
      - ➔ *Contraintes*

```
<schema>
   <name>mon_schema</name>
   <folder>schema0</folder>
   <tables>
      <table>
         <name>Fournisseurs</name>
         <folder>table1</folder>
         <columns>
            <column>
               <name>identifiantFournisseur</name>
               <type>CHARACTER(36)</type>
               <typeOriginal>varchar(36)</typeOriginal>
               <nullable>false</nullable>
               <description>identifiant du fournisseur</description>
            </column>
```

  - Des champs description peuvent être ajoutés sur tous les niveaux

Structure d'un fichier SIARD

- « content » contient :
  - Des dossiers « schemas » qui contiennent :
    - Des dossiers « table » qui contiennent :
      - ➔ *Un fichier xml avec les valeurs des données de la table*
      - ➔ *un fichier xsd décrivant la syntaxe du fichier de données*
      - ➔ *des dossier LOB (Large Object) BLOB (Binary LOB) ou CLOB (Character LOB) qui contiennent :*
        - Des fichiers de données binaires ou texte



Structure d'un fichier SIARD

- Exemple de fichier xml et de son schéma

```
<row>
    <c1>id_2</c1>
    <c2>2007-04-22</c2>
    <c3>0</c3>
    …
    <c17>0001000</c17>
</row>
<row>
    <c1>id_3</c1>
```

```
<xs:complexType name="rowType">
   <xs:sequence>
      <xs:element minOccurs="0" name="c1" type="xs:string"/>
      <xs:element minOccurs="0" name="c2" type="xs:date"/>
      <xs:element minOccurs="0" name="c3" type="xs:integer"/>
```

L'outil SIARD-Suite

- Siard-Suite est un logiciel Java développé pour les Archives fédérales Suisses
- Fonctionnalités :
  - Exportation d'une base d'un SGBD dans une archive SIARD
  - Importation d'une archive SIARD dans un SGBD
    - pour permettre aux utilisateurs d'utiliser les capacités de recherche des SGBD (interfaces graphiques ou requêtes en langage SQL).
  - Navigation dans les données et métadonnées d'une archive SIARD
  - Edition des métadonnées
- SGBD couverts :
  - Oracle, SQL-server, MS-Access, MySQL, DB/2

Retours d'expériences

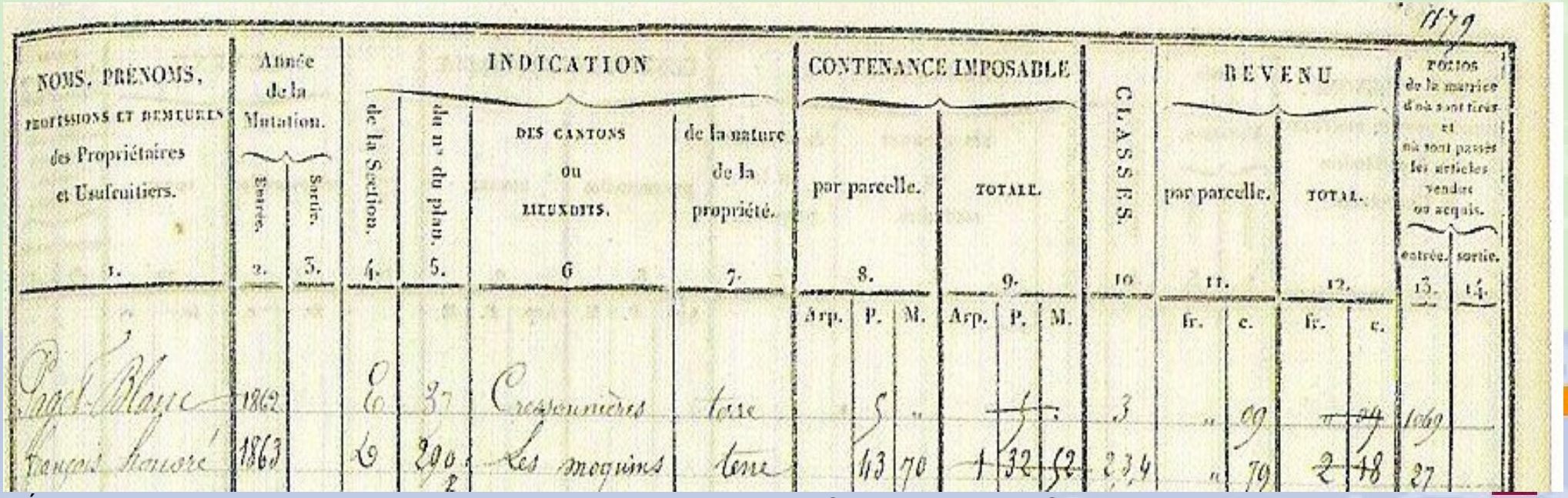
Expérimentations

- Les archives de la ville de Paris
  - Données sur les « aides sociales facultatives »
    - Issues d'une base de données relationnelle Oracle...mais les relations n'étaient pas codées dans le SGBD
- La matrice cadastrale
  - Données sur les propriétés et les propriétaires
    - Issues de fichiers dans un format « ASCII » à champs fixes

La matrice cadastrale

Historique de la matrice

- La matrice cadastrale est un document qui indique, pour chaque propriétaire, la liste des propriétés foncières en précisant leur consistance et leur revenu
- Sert au calcul de l'assiette de l'impôt



1179

NOMS, PRÉNOMS, PROFESSIONS ET DEMEURES des Propriétaires et Usufruitiers.	Année de la Mutation. Entrée.	Sortie.	INDICATION de la Section.	du n° du plan.	DES CANTONS ou LIEUXDITS.	de la nature de la propriété.	CONTENANCE IMPOSABLE par parcelle. Arp.	P.	M.	TOTALE. Arp.	P.	M.	CLASSES.	REVENU par parcelle. fr.	c.	TOTAL. fr.	c.	FOLIOS de la matrice d'où sont tirés et où sont passés les articles vendus ou acquis. entrée.	sortie.
1.	2.	3.	4.	5.	6	7.	8.			9.			10	11.		12.		13.	14.
Paget Blanc	1862		E	37	Cressonnières	terre		5	"		5		3	"	09	"	09	1069	
François Honoré	1863		D	290	Les moquins	terre		43	70	1	32	52	2,3,4	"	79	2	18	27	

État des propriétés de François Louis Honoré Paget Blanc, domicilié aux Cressonnnières, en Suisse [vers 1866].
Arch. Dép. Jura, 3 P 1633, f° 1178 (fonds du Cadastre).

La matrice

- Organisation des fichiers
  - Une production annuelle par la Direction Générale des Finances Publiques (DGFiP)
  - Pour chaque département :
    - 4 fichiers principaux :
      - ➔ *Propriétaires d'immeubles (FP) ;*
      - ➔ *Propriétés bâties (FPB) ;*
      - ➔ *Propriétés non bâties(FBNP) ;*
      - ➔ *Annuaire topographique initialisé réduit (FANTOIR).*
    - 2 fichiers complémentaires :
      - ➔ *Propriétés divisées en lots ;*
      - ➔ *Lots – locaux.*
    - Volume moyen par département : 1,2 Go (format ASCII d'origine non compressé)

La matrice

- Le format d'origine
  - « ASCII à champs fixes » accompagné d'une documentation au format .doc
  - Chaque ligne correspond à un article
  - Les articles peuvent suivre différents modèles (ou dessins d'enregistrement)

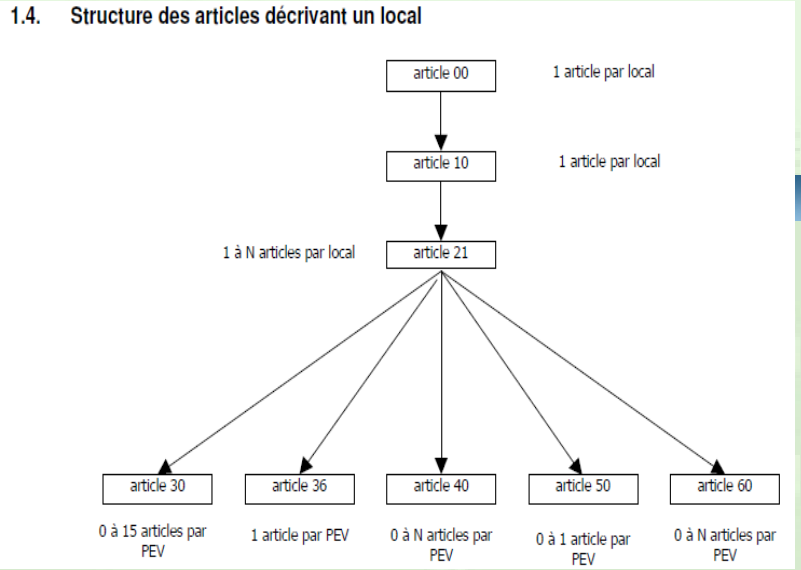
```
07.................................ARDECHE                        D421042010  201
0700010010000001                 00       A0323  A010001001B039 000385001 0551VILLEBRION...
0700010010000001                 10   0AS0002420031997      C 1P 000 MA000000518 0000023...
0700010010000001              00121   HA5M12000{000185000000518000002335
0700010010000001              00136   000000518 000002335 000001168 000000518 000002335 ...
```

Article « Tête Direction » : ..code et nom du département, l'année couverte...

Article de type « 00 » :l'indicatif cadastral et l'adresse du local.

Article de type « 10 » : contient les données générales du local. Un local peut être constitué de une à N parties d'évaluation ou PEV.

La matrice



Documentation élaborée par la DGFiP en 2012

- Les modèles
  - Entretiennent des relations entre eux
  - Ils sont décrits champ par champ, en indiquant :
    - le libellé, la position, la taille, le type, éventuellement la liste des valeurs autorisées, leur caractère obligatoire ou facultatif, une description intellectuelle.

```
07.................................ARDECHE                          D421042010  201
0700010010000001              00       A0323  A010001001B039 000385001 0551VILLEBRION...
0700010010000001              10   0AS0002420031997      C 1P 000 MA000000518 0000023...
0700010010000001           00121   HA5M12000{000185000000518000002335
0700010010000001           00136   000000518 000002335 000001168 000000518 000002335 ...
```

Article 10, champs « gpdl » (caractère 36) :
La valeur '1' signale que le local appartient à une propriété divisée en lots (PDL), même si le local n'est pas lui-même lié à un lot.

Article 10, champs « dsrpar » (caractère 37) :
Toujours servie. C'est celle de la parcelle d'assise ou de la parcelle de référence. En cas d'anomalie, elle est initialisée à 'A'.

Article 10, champs « dnupro » (caractère 38) :
Les modalités de reconstitution sont inchangées par rapport à celles existant sous Majic-Inter. Un local, dont le propriétaire n'est pas trouve, est attribue au compte fictif Y99999, avec une date de l'acte a blanc.
…

L'archivage des informations de la matrice

La matrice

- Coté service producteur (DGFiP)
  - Travail de re-documentation
    - Re-écriture de la documentation a posteriori (dans les quatre dernières années) par le service métier en documentant également les différences entre les différentes versions du format (différences sur les champs, leur appellation, leur contenu...)
- Coté service d'archives (SIAF pour les AD)
  - Travail de modélisation
    - Reprise des informations de la documentation au format .doc dans un modèle formel (en XML). Explicitation de la structure des données et de l'évolution dans le temps de cette structure

La matrice

- Mise au point d'un modèle formel
  - Déclaration des dessins d'enregistrements
  - Déclarations des champs
  - Déclaration des évolutions du format années après années
  - Déclaration des spécificités locales
- Objectifs
  - Contrôler la qualité des données
    - Écriture d'un programme de contrôle permettant de repérer automatiquement les écarts entre la documentation et les données produites
  - Formater les données
    - Conversion vers un format d'archivage pérenne intégrant données et documentation

La matrice

- Itérations modélisation / documentation
  - Jeux de tests
    - Erreurs de saisie (30 février, valeur non définie dans les dictionnaire de données, absence de champs obligatoires...)
      - ➔ *Relâchement de contraintes dans le modèle*
    - Erreurs d'interprétation (identifiants répartis, occurrences multiples, jargon métier opaque...)
      - ➔ *Affinement de la documentation et ajustement du modèle*
  - Le programme de contrôle a été utilisé pour tester toutes des données du stock existant (années 2004- 2011)
  - Des propositions ont été faite à la DGFiP dans l'objectif d'améliorer plus en amont la qualité des données en basculant du formalisme peu explicite et peu outillé actuel vers un formalisme d'export en XML directement contrôlable.

La matrice

- Pourquoi le format d'origine ne nous satisfait pas ?
  - La validation du format ASCII est pauvre. Il faut développer notre propre validateur ($$);
  - Le format ASCII ne permet pas de faire d'autres usages que ceux d'un éditeur de texte. Il faut développer notre propre outil de recherche et de consultation ($$$) ;
  - La documentation est exprimée dans un format (.doc) qui ne permet pas une exploitation automatisée ; Les traitements seront à la charge du lecteur ou médiatisés par un archiviste (ressources humaines, risque d'image) ;
  - La documentation et les données sont séparées. Il faudra dans la description du fonds expliciter cette relation (facile) ;

La matrice

- Justification du choix du format SIARD
  - C'est un format définit par et pour un usage « archives »
  - Les spécifications du format sont publiques et basées sur des formats bien connus (XML, ZIP, SQL:99).
  - Les données et la documentation peuvent être rassemblées dans le même conteneur et leurs relations explicitées.
  - Les fonctionnalités de recherche/édition des SGBD peuvent être utilisées pour développer rapidement des interfaces de consultation.

La matrice

- Justification du choix du format SIARD (suite)
  - Les éléments de structures décrits dans la documentation sont facilement *mappable* avec les concepts du modèle relationnel SQL-1999
    - Dessins d'enregistrement → tables
    - Les liens entre les dessins → contraintes relationnelles
    - Les noms des champs, leurs types (date, nombre, caractères...) → idem
    - Les contraintes d'unicité, nullable → idem
    - Les vocabulaires contrôlés → tables liées aux champs qui les utilisent
    - Les valeurs des champs → idem;

La matrice

- Inconvénient 1: pas d'outil de validation du format SIARD
  - Justification : Pour les Archives fédérales Suisses la confiance repose sur l'outil de création (SIARD-Suite) et la remontée des données dans un SGBD existant
  - Solution: Développement d'un outil de validation en collaboration avec le CINES (siard_validator https://adullact.net/projects/ff-toolbox/)
    - Validité du format de l'enveloppe ZIP64 non compressé
    - Vérification de l'arborescence et du plan de nommage
    - Validité du fichier metadata.xml
    - Validité de chaque fichier tableX.xml
    - Vérification du nombre d'enregistrements de chaque table
    - Vérification de l'unicité des clés primaires
    - ...

La matrice

- Inconvénient 2: Pas d'outil de consultation
  - Solution :
    - Remontée des données dans un SGBD (logiciel Siard-Suite)
    - Développement d'une interface de consultation pour un usage « archives » (différent de l'usage métier : calcul de l'assiette pour l'impôt)

Conclusions

Conclusions

- Avantages de SIARD :
  - C'est un format de fichier dont les spécifications sont publiques...
  - Le format a été suffisamment expressif pour nous permettre d'exprimer toutes les informations
- Inconvénients de SIARD
  - Peu d'outils existants
    - Mais possibilité d'utiliser les SGBD-R (Attention les métadonnées ne sont pas chargées dans la base)
    - Facilité de développer de nouveaux outils

Conclusions

- Ce que nous a appris cette expérimentation
  - Au-delà du choix d'un format de représentation, archiver une BDD conduit à se poser des questions sur :
    - Le changement de finalité : Pour quels usages ?
    - La qualité de la documentation : Problème de jargon du métier du producteur. Documenter est une opération très coûteuse et chronophage.
    - La qualité des données : Est-ce que les données sont bien comme on croit qu'elles sont ?
    - Le tri des informations:
      - ➔ *Attention avec l'utilisation de Siard-Suite, le tri doit être fait en amont (création de bases intermédiaires, attention à l'intégrité de la base)*
  - Finalement on est passé d'un projet où on cherchait un format d'archivage et on a aboutit à la définition d'un format pour la consultation.