Étude du format SIARD - Sofware Independant Archiving of Relational Databases

Sommaire

Introduction

La prise en charge pour archivage de données provenant de bases de données est une problématique complexe pour laquelle on ne dispose pas encore d'outils permettant une automatisation et une exploitation aisée du processus. L'initiative des archives fédérales suisses qui ont fait développer dès 2004 un format spécifique pour l'archivage des bases de données relationnelles (SIARD) et un programme *SiardSuite* associé permettant l'exploitation de ce format en est d'autant plus précieuse. Les archives de France ont par conséquent souhaité étudier ce format et tester le programme associé. Ce sont les résultats de ces études et de ces tests qui vous sont ici présentés.

L'analyse du format et du logiciel a été réalisée pour évaluer leur possible adaptation au contexte français. Il s'agit également de déterminer si leur emploi pourrait améliorer les procédures d'archivage des bases de données relationnelles actuellement utilisées par les services d'archives. En d'autres termes, il fallait identifier les avantages de SIARD par rapport à un simple export à plat des bases : récupération et exploitation non seulement du contenu de données, mais également de la structure de la base de données d'origine, des relations entre les tables et des informations de gestion. La présente étude distingue les observations faites sur le format, de celles faites lors de l'utilisation du programme *SiardSuite*, tant sur des jeux de test que sur des bases de production.

Le Format SIARD

Description générale

La structure du format SIARD est suffisamment décrite par les auteurs pour qu'il soit possible de consulter un fichier respectant ce format sans pour autant disposer de l'application *SiardSuite*.

Techniquement, un fichier SIARD est un conteneur ZIP64[1] non compressé. Ce conteneur contient deux dossiers : *header* et *content.* Dans le premier dossier (*header*), sont stockées les métadonnées de la base de données archivée. Le deuxième dossier (*content*) contient, lui le contenu des enregistrements de la base. Les métadonnées comme les enregistrements de la base sont écrits dans ces dossiers sous forme de fichiers au format XML[2] . Chaque fichier XML est lui même accompagné d'un autre fichier qui en définit de manière formelle la structure en utilisant la syntaxe XML-Schemas[3]. L'archiviste peut bien sûr éditer ces fichiers en utilisant un éditeur XML ou un simple éditeur de texte dans le but, par exemple, d'enrichir les métadonnées.

Conceptuellement, le format SIARD considère la base de données comme une entité unique (une base de données = un fichier SIARD) composée d'un contenu de données et de métadonnées qui comportent entre autre des informations de gestion.

Les atouts de SIARD

Contrairement à une exportation à plat d'une base de données, SIARD permet de coder la structure de la base (liste des tables, description pour chaque table de la liste de ses champs, description pour chaque champs de son type). Le format permet aussi de décrire les relations entre les tables ainsi que des informations de gestion.

La structuration explicite des métadonnées et des données confère à ce format une capacité d'automatisation et présente en cela un avantage de taille par rapport aux solutions techniques actuelles. En effet le format est suffisamment documenté pour permettre son déploiement (son contrôle, sa communication, etc.). Toutefois, il n'existe actuellement qu'un seul logiciel (*SiardSuite*) permettant sa manipulation. Il est cependant parfaitement envisageable de créer une application capable de manipuler le format SIARD, de le produire ou de le transformer.

Ainsi au cours de l'étude, nous avons mis au point un prototype en PHP[4] permettant d'exporter des bases de données MySql[5] au format SIARD, alors que ce SGBD[6] n'est pas supporté à ce jour par le logiciel *SiardSuite*. Bien qu'employant des outils et des technologies différents pour le créer, le fichier ainsi produit est parfaitement conforme aux spécifications du format SIARD et peut être utilisée dans le logiciel *SiardSuite.*

---

1 ZIP64 est un format relativement récent qui reprend en partie les dispositions du format PKZIP en permettant de contenir un plus grand nombre de données (plus de 4 Go).

2 Recommandation du W3C – Extensible Markup Language (version 1.0) 5ème édition de 2008

3 Recommandation du W3C – XML Schema (version 1.0) édition de 2001

4 PHP (Hypert PreProcessor), langage de script utilisé pour créer des pages WEB dynamiques.

5 MySQL, système de gestion de bases de données très employé orienté vers le service de données (accès en lecture de données très rapide).

6 Système de Gestion des Bases de Données

Le logiciel SiardSuite

Description générale

L'application SiardSuite comportent deux fonctions principales que sont a) la création d'une archive au format SIARD à partir de l'exportation des informations d'une base de données et b) la ré-importation des informations d'une archive au format SIARD dans un SGBD. Cette ré-importation n'a pas vocation à être exhaustive et ne doit pas être considérée comme un processus d'interopérabilité entre différents SGBD. Elle a été développée pour permettre aux utilisateurs d'utiliser les capacités de recherche des SGBD (interfaces graphiques ou requêtes en langage SQL).

L'application *SiardSuite* met en œuvre la norme SQL-3[7] pour extraire les informations des SGBD. Cette application ne supporte actuellement que trois SGBD: Oracle, Microsoft Sql Server et Access. Cette sélection semble couvrir les besoins suisses[8]. Le programme est écrit en Java et se connecte aux SGBD et aux bases qu'ils contiennent à l'aide de pilotes fournis par les éditeurs des SGBD. Pour MS Access, il s'agit d'un pilote de type ODBC[9] qui doit se configurer directement dans le système d'exploitation. Pour les deux autres SGBD il s'agit de pilotes de type JDBC[10] livrés dans un dossier du programme *SiardSuite*[11].

S'il est évident que le respect d'une norme comme SQL-3 constitue une démarche prudente dans une volonté de pérennisation et d'interopérabilité, un certain nombre de difficultés pratiques se rencontrent toutefois. En effet, les bases de données relationnelles sont développées dans des programmes informatiques complexes, appelés SGBD-R[12] qui ne respectent pas nécessairement ni complètement ni uniquement cette norme. En particulier, les fonctionnalités proposées par ces programmes qui ne respectent pas cette norme auront donc beaucoup de mal à être archivées.

Mise en marche et déploiement

La connexion aux SGBD réclame de l'opérateur responsable de l'archivage une certaine connaissance de son installation de base de données (compte administrateur, nom du serveur de base et port d'écoute notamment). Une fois la connexion au SGBD établie, le processus de création de l'archive SIARD peut se dérouler normalement. Ce processus est constitué d'une suite indivisible d'opérations atomiques. Lorsque ce processus échoue dans l'une de ses opérations, l'application *SiardSuite* indique l'erreur rencontrée qui peut être complexe a interpréter car la documentation de l'application ne traite pas des erreurs d'exécution[13]. En cas d'erreurs, le processus se bloque de telle sorte qu'il est en théorie impossible de récupérer une archive SIARD incomplète.

---

7 Norme ISO/CEI 9075 : 1999 - Structured Query Language (SQL-3)

8 Quelques repérages effectués au sein d'administrations françaises permettent également de confirmer le nombre important de SI basés sur ces produits. On peut notamment citer l'étude menée en 2009 au conseil général de la Haute-Saône (étude menée sous le contrôle des archives départementales).

9 ODBC : Open DataBase Connectivity, format édité par Microsoft pour permettre à des clients de base de données de communiquer avec les SGBD.

10 JDBC : Java DataBase Connectivity, API fournie avec Java permettant de se connecter à des SGBD.

11 Il en existe donc deux différents, un pour Sql Server et un pour Oracle.

12 SGBD(R) : Système de Gestion des Bases de Données, le R correspondant à Relationnelle.

13 En annexe 3 ont été reportées les erreurs relevées ainsi que des explications permettant de les comprendre.

Les dépendances logicielles

Les versions des SGBD ont beaucoup évolué depuis leurs création et continuent à évoluer. Les pilotes font également l'objet de mises à jour, en particulier pour tenir compte de ces évolutions. L'application *SiardSuite* se connecte et interroge, elle, de manière similaire tous les SGBD sans se préoccuper ni de leur version ni des versions des pilotes utilisés.

Cette dépendance de l'application *SiardSuite* vis-à-vis des SGBD ainsi que des pilotes a été observée lors de l'étude où l'on a pu, par exemple, mettre en évidence des résultats différents pour une même base de données chargée dans diverses versions d'un même SGBD.

Périmètre des données et des métadonnées archivées

Le format SIARD permet de décrire tant les fonctions de gestion spécifiques aux SGBDR que les données primaires des bases qu'ils contiennent. En revanche, les historiques de création et de fonctionnement des bases qui en tant qu'informations de contexte devraient idéalement être connues et ajoutées à l'archive, ne sont pas prévues dans le format. La liste qui suit[14] donne une typologie des informations que le format SIARD permet d'exprimer :

- Contenus de données
- Type de données (par exemple numérique, chaîne de caractères, binaires…)
- Contraintes sur les données
- Contraintes d'intégrité
- Contraintes d'unicité
- Contraintes référentielles
- Informations de gestion
- La vie d'une base (procédures, clauses de vérification, déclencheurs)
- L'interface d'utilisation (vues ou formulaires)
- Les comptes et les droits d'accès

Les données primaires, les contraintes essentielles[15] et la structure des bases[16] sont systématiquement traitées par le processus d'archivage de *SiardSuite*. Pour les autres informations, même si le format d'archivage SIARD prévoit leur description (cf. la liste ci-dessus), le programme *SiardSuite* opère des choix et ne les archive pas toutes.

Alors que dans les pratiques d'archivage de bases de données par les services d'archives, les déclencheurs, les vues, les procédures stockées et les droits d'accès n'ont jusqu'ici jamais fait l'objet d'un traitement archivistique particulier, l'application *SiardSuite* propose de conserver ces informations et n'écarte dans son traitement que les clauses de vérification et les déclencheurs[17]. Le processus évacue les informations de gestion les moins pertinentes pour les archivistes. Ainsi par exemple, pour l'archivage des procédures stockées, leurs « corps originaux » (morceaux de codes) ne sont généralement pas récupérés parce qu'ils sont écrits dans des langages propriétaires et qu'ils ne présentent pas un grand intérêt dans une finalité historique. A l'inverse, les types de données définis pour chaque colonne d'une table sont

14 En annexe 2 sont reportées une introduction au modèle relationnel ainsi que la définition des termes techniques relatifs à l'usage d'un SGBDR.

15 Il s'agit des contraintes d'intégrité (clés primaires), des contraintes d'unicité (clés secondaires) et des contraintes référentielles (clés étrangères), autrement dit de la définition des identifiants et des relations entre les tables.

16 La structure d'une base inclut la liste des tables, la description de leurs champs et des types de données.

17 Les clauses de vérification et les déclencheurs sont des bouts de code SQL ou non, permettant de lancer des actions de gestion sur une base en production, avant ou après insertion ou mise à jour.

systématiquement traités par le programme d'archivage car ils renseignent sur l'état de la base archivée et permettent l'import des données primaires dans des applications de consultation.

Le protocole de test

Afin de comparer les résultats sur les trois SGBD supportés par *SiardSuite,* une même base de données a été définie et a été utilisée pour l'ensemble de nos tests[18]. Sa structure ne se compose que de neuf tables au maximum selon les fonctionnalités que l'on décide d'archiver. On a veillé à appliquer dans cette base l'ensemble des contraintes et méthodes de gestion[19] que le format SIARD est censé archiver. Les tests se sont déroulés en plusieurs étapes distinctes. Dans la première, le fichier de métadonnées a été particulièrement expérimenté. Dans un second temps, c'est le fichier SIARD qui a fait l'objet de tests. L'encodage des caractères et celui des *lob*[20] y ont été par exemple examinés. Enfin, les aspects de re-import des archives SIARD dans les SGBD ont été vus. Les pertes d'information étant relativement importantes au cours de cette étape, c'est plutôt l'intégrité des données primaires qui a été étudiée.

Après cette phase exploratoire à l'aide d'une base fictive, nous avons testé l'archivage de bases de production, possédant peut-être moins de méthodes de gestion, mais représentant mieux la réalité, notamment en ce qui concerne le nombre d'enregistrements. Les éléments posant problème sont exclusivement des informations de gestion secondaires. Les contraintes des bases les plus pertinentes que sont les clés primaires, secondaires et étrangères[21] sont bien archivées. Les résultats ont été synthétisés sous la forme d'un tableau reproduit en annexe.

Conclusion

Les résultats de la présente étude ont été transmis aux créateurs du SIARD et de l'application *SiardSuite* et ont donné lieu à de nombreux échanges. Il en ressort qu'il existe un certain décalage entre ce que prévoit la documentation et le traitement effectif du processus d'archivage. Pour autant, les créateurs semblent avoir conscience de certains manques de la documentation et considèrent toujours leur logiciel comme devant être amélioré. Les tests ont été menés au début de l'étude, sur la version 1.19 de la *SiardSuite* et ont été poursuivis sur la version 1.20 sortie entre temps. Des améliorations ont pu être relevées entre les deux moutures, preuve que le développement se poursuit et va dans le bon sens. Les concepteurs ont indiqué qu'un certain nombre de modifications issus des résultats de cette étude seraient appliquées à la prochaine version de la *SiardSuite*[22]. En l'état actuel, il parait peu probable que le format SIARD lui-même fasse l'objet d'une mise à jour. Le sentiment qui résulte de l'étude est que ce dernier est fonctionnel, assez bien documenté, qu'il a fait l'objet d'une phase de réflexion préalable intéressante qui lui permet à présent de répondre le plus souvent de manière convaincante aux contraintes qu'impose l'archivage de bases de données relationnelles.

Il parait en revanche primordial de poursuivre le développement du logiciel d'archivage *SiardSuite* dont les sources sont détenues par les archives fédérales suisses, afin de régler

18 Sa description technique n'est pas reportée dans ce document mais le sera en détail dans les annexes.

19 Voir le paragraphe précédent et l'annexe technique.

20 LOB : Large Object Binary (BLOB ou CLOB), type de données permettant de stocker des données binaires (de très grande taille, type images, sons...).

21 Les contraintes référentielles sont l'essence même du schéma des bases de données relationnelles.

22 Une version 1.21 a pu être étudiée mais elle n'est pas encore disponible au téléchargement.

certains problèmes. En particulier, il serait bon de permettre une meilleure opérabilité ou compatibilité du programme avec les différentes versions des SGBD qu'il supporte. Le support d'autres SGBD rendrait bien sûr l'outil plus utile et adapté à la diversité des réalités du terrain. Enfin il est également indispensable que la documentation soit mise à jour car en l'état actuel, les différences relevées sont très importantes.

Ce programme démontre toutefois qu'il est possible d'utiliser le format SIARD de manière convaincante et ouvre des perspectives intéressantes dans son maniement notamment sur la question de la consultation des données.

Contact, Téléchargement

Le format SIARD et le programme d'archivage *SiardSuite* sont librement et gratuitement téléchargeables sur le site WEB des archives fédérales suisses à l'adresse suivante :
http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=fr
Il suffit de s'inscrire pour disposer d'un login et d'un mot de passe valides. L'utilisation de *SiardSuite* nécessite d'avoir une machine virtuelle Java 1.5 installée sur son poste informatique. Pour se connecter aux SGBD, les procédures diffèrent selon les produits mais la documentation les expose bien.

Pour communiquer avec les responsables du programme de développement du SIARD, les coordonnées des personnes suivantes peuvent être intéressantes.

Aux archives fédérales suisses :
Amir Berstein : Amir.Bernstein@bar.admin.ch
Urs Meyer : Urs.Meyer@bar.admin.ch

Société qui a créée le format et le programme :
Hartwig Thomas : hartwig.thomas@enterag.ch

Remerciements

Le service interministériel des Archives de France tient à remercier pour leur précieux concours les interlocuteurs qui ont collaboré et répondu à notre demande , à savoir les conseils généraux de l'Aube[23], de la Haute-Saône[24] d'une part. D'autre part, les conseils généraux de la Loire[25], des Pyrénées-Orientales[26] ainsi que Brest métropole Océane[27] pour avoir fournis des exemples de bases de données en production. Les archives de France adressent également leurs remerciements à Fred Grevin[28] pour ses conseils avisés.

23 Service d'archives départementales, Céline Guyon.
24 Service d'archives départementales, Georges Rech.
25 Service d'archives départementales, Agnès Goudail et service informatique Jean-Marc Faure.
26 Service d'archives départementales, Christine Langé et service informatique, Xavier Dibusi.
27 Service d'archives, Coline Vialle et service informatique, Jean-Luc Breton.
28 Service en charge du « records management », des archives et de la bibliothèque documentaire de la ville de New York.