Bases de données : sommaire

Ce Catalogue des formats de données d'archivage a atteint ses limites avec le sujet *bases de données*. Hormis pour de petits exemplaires créés par des programmes de bureautique très répandus, les bases de données ne peuvent être traitées qu'exceptionnellement sous forme de fichiers autonomes. La conversion de ces derniers dans un format d'archivage constitue un élément essentiel de l'archivage à long terme. Les bases de données font, au contraire, souvent partie d'un système plus global: applications techniques (dans l'administration), systèmes de gestion de contenu (pour sites internet, etc.), systèmes d'information géographique (GIS), logiciel d'archivage, etc. Souvent, les informations essentielles ne sont pas uniquement enregistrées dans la base de données, mais aussi en partie dans la logique de programmation du système ou dans l'interface avec l'utilisateur. Il n'est donc pas suffisant de ne convertir que la base de données dans un format d'archive. L'archivage de ces systèmes doit être effectué globalement et chaque fois individuellement.
Tout en gardant cette situation initiale à l'esprit, il peut être toutefois opportun d'archiver une base de données, plus exactement son contenu. Pour une évaluation future ou statistique du contenu, il ne suffit pas d'archiver les données brutes, il faut aussi tenir compte de leur structure et de leurs relations.
L'archivage actuel de bases de données, en général de bases de données relationnelles, non seulement le format cible est important, mais également la méthode d'archivage. Nous nous conformerons toutefois à la structure générale de ce catalogue de formats et diviserons les méthodes d'archivage malgré tout selon le format définitif des fichiers à archiver.

Formats examinés

- CSV
- SIARD
- SQLX
- SQL Script

Recommandation

- La prise en charge de tableurs de base de données dans un format CSV est répandue au sein des archives jusqu'à présent. Au vu des limites des CSV , notamment du manque de documentation et de méta-informations, ce format peut uniquement être considéré comme une solution d'urgence pour le moment. Il est recommandé d'utiliser un format plus global et spécifique à l'archivage de bases de données, tel que SIARD .

CSV

Informations générales

Titre	Comma-Separated Values; Colon-Separated Values; Character-Separated Values
Catégorie	Données structurées de tableurs et bases de données
Abréviation	CSV
Extension de fichier	.txt, .csv
Mime Type	text/CSV - text/comma-separated-values
Pronom PUID	x-fmt/18
Version	Il n'existe pas de norme générale du format de fichiers CSV. Un cahier des charges RFC du format d'un fichier CSV existe et sert en générale de référence: RFC 4180.

Description

Les données CSV sont des fichiers ASCII structurés sous forme de tableau. Les valeurs, les champs ou les colonnes individuels sont séparés par un délimiteur, par exemple une virgule ou un point-virgule. Les lignes des tableaux sont séparées par un saut de ligne. Lors du transfert d'une base de données relationnelle dans des fichiers CSV, chaque tableau de la base de données est copié dans un fichier CSV.
Les différentes variantes et quasi-normes du format de fichier CSV se distinguent par le masquage des délimiteurs et du retour automatique à la ligne dans les champs.
Deux variantes CSV sont particulièrement intéressantes dans le domaine des bases de données:

- Le *format CSV de Microsoft Excel* se distingue par des guillemets délimitant les champs et par un nombre variable de champs par ligne. La première ligne permet d'indiquer les noms des colonnes. Le format Excel CSV peut être lu par de nombreuses bases de données.
- Le *SQL-Loader File* d'ORACLE introduit un en-tête précédant le contenu du fichier CSV dans lequel la dénomination des champs, leur format, le jeu de caractères, etc. sont définis. De nombreuses bases de données peuvent importer ou créer des «SQL-Loader Files».

Evaluation

Ouverture du format : **4**
Il n'existe pas de norme générale du format de fichiers CSV. Par contre la spécification est contenue pour l'essentiel dans RFC 4180 et extrêmement simple.

Licence libre : **3**
Il n'existe pas de restrictions juridiques associées à une licence; cela est vraisemblablement aussi le cas pour *Excel CSV Format* et *SQL-Loader File*.

Diffusion : **4**
CSV est le format d'échange de données structurées le plus répandu, autrement dit entre bases de données et tableurs.

Fonctionnalités : **1**
Les fichiers CSV sont ce que l'on nomme des *flat files*, autrement dit seules les informations d'un tableau peuvent être stockées dans un fichier. La plupart des bases de données ont toutefois recours à des structures hiérarchiques ou relationnelles pour stocker les données. Pour reproduire aussi cette structure dans une *flat file*, il est nécessaire de répéter des informations, d'où une redondance des données. En outre, les données stockées dans les fichiers CSV ne sont pas formatées. Il est impossible de reprendre des formats de champs, des structures de données ou des formules.

Implémentation : **4**
L'immense majorité des bases de données et des tableurs sont capables de créer ou de lire des fichiers CSV.

Densité de mémorisation : **3**
Pas de compression des données, les contenus des champs sont reproduits caractère par caractère. La représentation des structures entre les tableaux conduit inévitablement à la redondance des données (voir fonctionnalité ci-dessus).

Vérifiabilité : **2**
Seule l'extension du nom de fichier permet une reconnaissance ou une validation. Le manque de normalisation rend toutefois nécessaire la présence d'une documentation précise sur le masquage des séparateurs et des retours automatiques à la ligne, l'utilisation de guillemets et le problème de la variation du nombre de champs par ligne. Le codage des caractères et la représentation des types de données utilisées doivent aussi être fixés.

Bonnes pratiques : **2**
Pour des raisons historiques, les fichiers CSV sont relativement répandus au sein des archives.

Perspectives : **1**
CSV est souvent éliminé au profit de formats de tableaux basés sur XML, car celui-ci résout les problèmes du masquage des caractères de contrôle, du jeu de caractères et des types de données.

Classe de formats : **A**
CSV est un des plus anciens formats de l'informatique.

Conclusion

De grandes quantités de données sont déjà archivées en format CSV pour tableaux. Celui-ci gardera donc son importance. Toutefois, vu l'impossibilité de conserver des relations, des métadonnées et des informations structurelles dans ce format, seuls des ensembles de données en format CSV bien documentés garderont leur valeur. Un remplacement par XML (SIARD , SQLX , OOXML ou ODF) aura aussi lieu dans le domaine de l'archivage. N'oublions pas que XML a été élaboré dans une mesure non négligeable pour répondre au besoin d'un format d'échange de données structuré et pour remplacer les solutions entièrement basées sur le texte.

Références

Cahier des charges RFC du format des fichiers CSV
https://tools.ietf.org/html/rfc4180

Bibliographie

Wikipédia: CSV
https://fr.wikipedia.org/wiki/Comma-separated_values
Wikipédia: CSV (en anglais)|
https://en.wikipedia.org/wiki/Comma-separated_values
Creativyst Software, The Comma Separated Value (CSV) File Format
http://www.creativyst.com/Doc/Articles/CSV/CSV01.htm
Oracle SQL*Loader Tutorial
http://loader.datenbank-wissen.de/

Articles connexes

Le format CSV est aussi répandu comme format d’archivage et d’échange dans le domaine des tableurs .

SIARD

Informations générales

Titre	SIARD RDB DATA – Software Independent Archival of Relational Databases
Catégorie	Données structurées de bases de données
Abréviation	SIARD
Extension de fichier	.siard
Mime Type	-
Pronom PUID	fmt/161
Version	Version actuelle 2.0 (2016, eCH-0165 v2.0, a été développée en collaboration avec le projet européen d'archivage eARK) Versions antérieures: la version 1.0 a été publiée une première fois en 2008 par les Archives fédérales suisses. Une version au contenu identique a été publiée en 2013 en tant que norme eCH-0165.

Description

SIARD permet d'enregistrer dans un codage XML simple des structures (schémas, tableaux etc.) et le contenu de bases de données relationnelles. Les archives SIARD consistent en un fichier de contenu et un fichier de métadonnées comprenant des métadonnées de tous les niveaux. SIARD est basé sur des normes ISO (SQL:1999 [SIARD Version 1]/SQL:2008 [SIARD Version 2] et XML 1.0) et permet de conserver des bases de données relationnelles en provenance de différents systèmes, notamment MS Access, Oracle, MS SQL et MySQL. Il est également possible d'archiver des collections de fichiers CSV en format SIARD.

Evaluation

Ouverture du format : **4**
SIARD est une norme eCH.

Licence libre : **4**
Les Archives fédérales suisses sont détentrices du copyright du processus SIARD. En recourant à la normalisation par eCH, les Archives fédérales suisses renoncent à la perception de droits de licence.

Diffusion : **2**
La diffusion de SIARD se limite pour l'essentiel au monde des archives. Le format est utilisé dans différents archives en Suisse et à l'étranger.

Fonctionnalités : **3**
L’enregistrement de SIARD permet de reconstituer intégralement un schéma de base de données avec tous ses objets, en les rouvrant dans un SGBD relationnel. On ne peut distinguer logiquement la base de données de l’original.

Implémentation : **4**
Il existent divers outils qui permettent de créer et lire les fichiers SIARD parmi lesquels figurent également des produits open source.

Densité de mémorisation : **3**
L’utilisation de fichiers XML pour le stockage des données primaires entraîne un volume de mémorisation relativement grand. La compression « deflate » améliore cependant la densité de mémorisation.

Vérifiabilité : **3**
Il existe un validateur pour les fichiers SIARD (KOST-Val).

Bonnes pratiques : **2**
SIARD est utilisé dans l'archivage de bases de données en Suisse et à l'étranger et constitue en outre un format d'archivage officiel de différents projets européens (PLANETS, E-ARK).

Perspectives : **3**
SIARD est reconnu dans le monde des archives; on peut s'attendre à ce qu'il continue à s'imposer.

Classe de formats : **B**
SIARD a été conçu spécialement pour l’archivage des bases de données les plus couramment utilisées (bases de données relationnelles). SIARD est utilisé dans de nombreuses archives et sa diffusion ne cesse de croître, pas uniquement en Suisse.

Conclusion

SIARD offre une possibilité pour l'archivage de bases de données relationnelles. Il implémente un codage XML simple pour la conservation à long terme de données d'archives.

Références

eCH
eCH-0165: Spécification de format SIARD, Version 2.0
2016
https://www.ech.ch/vechweb/page?p=dossier&documentNumber=eCH-0165&documentVersion=2.0

Bibliographie

Kaufmann, Roger und Voss, Andreas
Save your databases using SIARD!
2014
https://web.stanford.edu/group/dlss/pasig/PASIG_September2014/20140917_Presentations/20140917_15_Introduction_to_SIARD_Roger_Kaufm
Ohnesorge, Krystyna

SIARD – The Swiss Solution for Archiving Relational Databases
2016
http://www.eark-project.com/resources/conference-presentations/finconfpres/82-day-2-5-the-use-of-siard-in-e-ark/file

Articles connexes

Il existe une relation fonctionnelle avec SQLX et SQL Script .

SQL script

Informations générales

Titre	*Scripting Database*
Catégorie	Données structurées de bases de données
Abréviation	SQL
Extension de fichier	.sql
Mime Type	-
Pronom PUID	-
Version	Il ne s'agit pas d'un format de données mais d'une méthode permettant de créer une base de données relationnelle à partir d'un fichier de scriptage (fichier texte) et réciproquement. Le fichier texte contient un ensemble de commandes SQL. Ces commandes SQL doivent se conformer à une version SQL déterminée (par exemple SQL-92). SQL est standardisé par ANSI et ISO.

Description

La méthode consistant à construire une base de données contenant tous les objets de la base de données au moyen d'un ensemble de scripts SQL est utilisée en général dans le domaine du développement de bases de données, lors de la création de bases de données au moyen d'outils d'assistance CASE, pour la documentation et lors de la sauvegarde de données.
Un ensemble de commandes DDL (Data Definition Language) permet de définir la structure ou les objets de la base de données. Les commandes DML (Data Manipulation Language) permettent finalement de placer les données dans les objets de la base de données/tableaux. Toutes ces commandes, placées séquentiellement dans un fichier, sont exécutées par l'interpréteur SQL et créent un schéma de base de données avant de remplir les tableaux de données. Il existe plusieurs outils permettant de générer des scripts. Certaines bases de données comportent déjà ces outils (SQL-Server: «Generate SQL Script wizard», PostgreSQL: «SQL Manager», etc.).
Cette méthode, semblable d'ailleurs en cela à SQLX ou SIARD , n'archive que la structure et le contenu de la base de données. La logique d'application, qui peut constituer une partie nécessaire à la compréhension d'une application technique, n'est pas traitée.

Evaluation

Ouverture du format : **3**
La compréhension d'un script est grandement facilitée quand la syntaxe est conforme à une version SQL déterminée. Toutefois, de nombreux outils génèrent des scripts pour une base de données particulière et utilisent alors, principalement dans le domaine DDL, des commandes spécifiques à un type particulier de base de données qui ne sont pas conformes à la norme SQL.

Licence libre : **3**
Il n'existe pas de restrictions juridiques associées à une licence pour les scripts SQL.

Diffusion : **2**
La diffusion principale est dans les domaines du développement de bases de données, de la création de bases de données au moyen d'outils d'assistance et de la documentation de bases de données. Cette voie n'est pas empruntée très souvent pour l'archivage de bases de données.

Fonctionnalités : **4**
Un fichier SQL-Script activé permet de reconstituer intégralement un schéma de base de données avec tous ses objets. La base de données ne peut se distinguer logiquement de l'original.

Implémentation : **4**
Il existe aussi un grand nombre d'outils de scriptage indépendants des producteurs de bases de données. Un outil de scriptage peut être en outre décrit ou implémenté complètement en SQL.

Densité de mémorisation : **1**
Les commandes SQL des scripts provoquent une énorme augmentation de la quantité de données. C'est pourquoi cette voie est rarement empruntée dans des buts d'archivage (c'est pourquoi SQL-Loader crée bien des commandes DDL dans l'en-tête mais écrit finalement les données sous forme de fichier CSV).

Vérifiabilité : **2**
Seule une reconnaissance de format rudimentaire au moyen de l'extension de nom de fichier est possible.

Bonnes pratiques : **1**
Cette méthode n'a actuellement presque aucune signification pour l'archivage de bases de données. Elle constitue toutefois la base d'approches plus prometteuses ou plus répandues pour l'archivage comme SIARD et SQLX .

Perspectives : **1**
Etant donné la redondance élevée des données (toujours les mêmes commandes DML répétées), il ne faut pas s'attendre à ce que cette méthode s'impose pour l'archivage de données structurées provenant de bases de données.

Classe de formats : Ø

Conclusion

Cette méthode est théoriquement intéressante car toutes les étapes sont soumises à la norme SQL. Le fichier de scriptage peut être complètement créé par un script SQL et contient de nouveau un script SQL. SQL (Structured Query Language) est un langage qui a joui d'un développement stable et de longue durée. Il sert à interroger et à manipuler des données dans les bases de données relationnelles et possède donc en théorie de ce point de vue une aptitude élevée à l'archivage. Toutefois, étant donné que de nombreux systèmes de bases de données, en particulier dans le domaine DDL, ne respectent pas entièrement la norme SQL, un accès simple et durable à l'information n'est possible que dans une mesure limitée. Puisqu'il n'y a pas de validateurs SQL, la conformité à la norme ne peut pas simplement être vérifiée et prouvée. En outre, le fait que les données archivées ne puissent être exploitées utilement que sous forme comprimée ne joue pas en sa faveur. Pour des raisons pratiques, il n'est donc pas conseillé d'utiliser le script SQL pour l'archivage, surtout si les archives ne préparent pas elles-mêmes les données.

Bibliographie

Wikipédia: SQL
https://fr.wikipedia.org/wiki/SQL
Microsoft.com (ed.), Documenting and Scripting Databases
http://msdn2.microsoft.com/en-us/library/ms191299.aspx
SQLScripter
http://www.sqlscripter.com/
PostgreSQL, EMS SQL Manager 2005 for PostgreSQL ver.3.6 released!
https://www.postgresql.org/about/news.570

Articles connexes

Il existe une relation fonctionnelle avec CSV et SIARD .

Contact
A propos
Impressum
Événements
Newsletter
RSS

SQLX

Informations générales

Titre	SQL/XML
Catégorie	Données structurées de tableurs et bases de données
Abréviation	SQLX
Extension de fichier	.xml, .sqlx
Mime Type	-
Pronom PUID	-
Version	Version actuelle: 5 (L'échange entre la présentation des données XML et les bases de données relationnelles est spécifié dans le chapitre 14 de SQL:2016 «XML-Related Specifications (SQL/XML)».)

Description

SQLX n'est pas un format de fichier au sens strict (le format du fichier est toujours XML dans ce cas). Il s'agit plus précisément d'un ensemble de fonctions basées sur SQL pour l'exportation («publish») de tableaux à partir de fichiers XML et pour l'importation («extract/store») de ces mêmes fichiers XML dans une base de données relationnelle. L'exportation de tableaux individuels est triviale et est déjà maîtrisée par la majorité des systèmes de bases de données et des tableurs. L'application de l'intégralité d'une base de données relationnelle en représentation XML des données est plus délicate voire résolue de façon insatisfaisante. Le problème fondamental est posé par l'application du modèle relationnel des données au modèle hiérarchique XML.

Evaluation

Ouverture du format : **2**
Les spécifications sont devenues entre-temps plus étendues mais sont plutôt exprimées sous forme d'une proposition de normalisation. Il faut encore s'attendre à des modifications.

Licence libre : **4**
La proposition de normalisation d'une partie de SQL:2016 n'est pas soumise à des restrictions juridiques associées à une licence. Les implémentations de la fonctionnalité SQLX dans chaque système de base de données seront bien entendu propriétaires.

Diffusion : **3**
SQLX jouit déjà d'une diffusion significative comme possibilité d'exportation et pour l'échange de données de tableaux individuels. SQLX remplace CSV et corrige ses faiblesses connues. La résolution de l'exportation intégrale des schémas de la base de données en représentation XML des données semble encore présenter des aspects non résolus.

Fonctionnalités : **3**
SQLX permet de reprendre la dénomination et le format des champs d'un tableau dans un fichier XML. L'intégration des données primaires et des métadonnées ainsi que la normalisation sont bien mieux résolues que dans le cas de CSV . L'application des structures hiérarchiques des données d'une base de données peut être effectuée sans formation de redondances. L'application d'une base de données relationnelle dans un fichier XML n'est pas encore possible de façon générale.

Implémentation : **3**
La fonctionnalité SQL disponible permet d'implémenter les fonctions SQLX dans presque chaque base de données et tableur. Un grand nombre de fournisseurs (Oracle, Microsoft, etc.) ont déjà implémenté les fonctions SQLX, la conformité à la norme de l'implémentation varie cependant énormément.

Densité de mémorisation : **2**
Il ne se forme aucune redondance tant que les seules relations appliquées sont d'ordre hiérarchique (voir la fonctionnalité ci-dessus). Le stockage des balises (tags) XML dans le fichier ne provoque aucun «gonflement» indu de celui-ci. En règle générale, cela est accompli par une compression des données (ZIP) lors de l'enregistrement.

Vérifiabilité : **2**
Seule une reconnaissance de format rudimentaire est possible.

Bonnes pratiques : **1**
Bien que SQLX puisse remplacer CSV sans problèmes, il n'est pas à l'ordre du jour des archives.

Perspectives : **1**
SQL/XML est une approche de solution intéressante pour l'échange de données et l'archivage à long terme. Cette approche est déjà bien répandue et va continuer à s'étendre. Elle peut difficilement s'établir dans le domaine des archives.

Classe de formats : **D**
Le cahier des charges du format n'est pas encore irréprochable mais peut intéresser l'archivage.

Conclusion

SQLX est une approche assez prometteuse de solution pour l'archivage de données structurées. Il existe des solutions bien au point pour la conversion de tableaux plats en fichiers XML. L'exportation de bases de données relationnelles dans leur intégralité n'est en revanche pas résolue dans tous les cas. Il faut aussi s'attendre à une poursuite du développement dans le domaine de la standardisation. Mais comme les spécifications de SQLX concernent la fonctionnalité de «publish» et de «extract/store» et non pas la forme de la représentation XML des données elles-mêmes, la poursuite de ce développement ne constitue pas un handicap pour l'utilisation.

Références

ISO/IEC 9075-14:2016 «Information technology — Database languages — SQL — Part 14: XML-Related Specifications (SQL/XML)»
https://www.iso.org/standard/63566.html
[payant]

Bibliographie

Wikipédia: SQL/XML (anglais)
https://en.wikipedia.org/wiki/SQL/XML
OracleBase, SQL/XML
http://www.oracle-base.com/articles/9i/SQLXML9i.php

Articles connexes

SQLX peut aussi être utilisé dans le domaine des tableurs.
Il existe une relation fonctionnelle avec CSV .