

Sample 3/4/5 Data Files

2009-09-29

Goal.....	1
The “ownership” Database	1
Folder Structure	2
Individual File Structure	2
Element <root>	2
Required Attribute @count of <root>.....	2
Repeating Optional Element <Submission> of <root>	2
Required Attribute @sAdsh of <Submission>	2
Required Attribute @sAcceptanceDatetime of <Submission>	2
Repeating Required Element <Attachment> of <Submission>.....	2
Required Attribute @sFormType of <Attachment>	2
Required Attribute @iSequence of <Attachment>	2
Required Attribute @sFilename of <Attachment>	3
Required Element <ownershipDocument>	3

Goal

This document explains the contents of the ten files comprising the “3/4/5” sample data and how it relates to the ‘ownership’ database that accompanies them.

The “ownership” Database

The accompanying file ownership-2009-09-27.mdb is an Access database that contains only the “header” information of each attachment. That header information consists of the following data:

Column	Type	Description
dCIK	Double	One of the CIK’s to which the Attachment applies. Forms 3, 4, and 5 report transactions that will typically involve the CIK’s of buyers, sellers, and the company whose securities are changing hands.
sCompanyName	String (255)	The Conformed Company Name of the CIK
sSubType	String (10)	Submission Type (regex= (((3) (4) (5))(/A)?)
dtFilingDate	DateTime	The filing date, which may be after the submission date but never before.
sAdsh	String (20)	20-character accession number, with dashes
sDocType	String (10)	Same type as sDocType, usually same value
iDocSeq	Integer	Order in which the attachment appeared in the submission.
iDocSize	Integer	Number of bytes in the attachment

The table has no primary key; it is essentially a join table between dCIK and sAdsh.

Folder Structure

There is a single folder, data. It contains ten xml files named, appropriately, 0.xml, 2.xml, 3.xml, etc. Each file is approximately 32MB uncompressed.

Each file contains the contents of all of the submissions whose accession number ended with 0, 1, 2, respectively, from 1 June 2009 to 16 September 2009.

Because the accession numbers are simply unique ID's generated in sequence, the 10 files form a pseudorandom partitioning done so as to ease generation and testing of code and transport of the data; their union is the dataset of interest.

Individual File Structure

Element <root>

The root element of the ten files is "root" and serves only as a container for the Submission elements.

Required Attribute @count of <root>

The number of child Submission elements. There is no limit to the number of Submissions, but the sample data has about 5000 in each file.

Repeating Optional Element <Submission> of <root>

This element is a container for data extracted from the original .txt submission file.

Required Attribute @sAdsh of <Submission>

The value is a 20-character string corresponding to the sAdsh field of the ownership database.

Required Attribute @sAcceptanceDatetime of <Submission>

This is a 16-digit string that provides a complete timestamp for the entire submission down to the second granularity. The first ten digits of this string SHOULD be the same date as in the database's dtFilingDate except where the submission was made after 5:30pm EST. In general it is the sAcceptanceDatetime that is more useful for analysis than the filing date.

Repeating Required Element <Attachment> of <Submission>

Usually there is only one Attachment but there could be more. An attachment is the container for the actual XML submission.

Required Attribute @sFormType of <Attachment>

This is a string that corresponds to sSubType in the ownership database.

Required Attribute @iSequence of <Attachment>

A positive integer indicating the original sequence in which the file was attached to the submission. Almost always 1.

Required Attribute @sFilename of <Attachment>

A string indicating where the original XML file can be found on EDGAR. For example if the original submission was at

<http://www.sec.gov/Archives/edgar/data/99999/1111111111011111/1111111111-09-111111.txt>

then if @sFilename is edgar.xml it can be found at

<http://www.sec.gov/Archives/edgar/data/99999/1111111111011111/edgar.xml>

However, the SEC site would not welcome 50,000 separate downloads of these files from the EDGAR site, which is why they have been provided to you in this form.

Required Element <ownershipDocument>

This is the actual XML data. The content of this element is documented at:

<http://www.sec.gov/info/edgar/ownershipxmlspec-v1-r1.doc>