Establishing a Mechanism for Maintaining File Integrity within the Data Archive

Thomas C. Stein, Edward A. Guinness, Susan H. Slavney

Earth and Planetary Sciences, Washington University, St. Louis, MO, 63130
stein@wunder.wustl.edu, guinness@wunder.wustl.edu, slavney@wunder.wustl.edu

Abstract. An important but often overlooked part of the data archiving process is to ensure that archived data remain unchanged over time. Early data archives produced by the Planetary Data System (PDS) were written to write-only optical media (CD-ROM and CD-R, and later, DVD-ROM). These read-only media provided first-order assurance that data read from them were the equivalent to the original data. In addition, the distribution of several hundred copies of an archive (in the form of CD-ROMs) effectively provided backups of the data. Most current planetary data archives are stored in online using rewriteable media. As such, the data are vulnerable to accidental changes and deletions, as well as intentional changes by virus, Trojans, and the like. We are exploring mechanisms to maintain file integrity that could be integrated into standard PDS procedures. File integrity would begin with the data provider and continue through the life of the archive. In this paper we discuss the need for such a mechanism and the use of digital signatures as a means of determining file integrity. We survey standard cryptography algorithms and commonly available software tools that might be used to produce digital signatures. Finally, we discuss the role of digital signatures within the archive life cycle.

Introduction

Properly maintaining a long-term (> 100 years) archive includes ensuring that data formats are well-defined, that the data are accessible, and that the archive contents remain unchanged. NASA's Planetary Data System (PDS) Geosciences Node [Guinness et al., 1996] is charged with developing and maintaining long-term archives for the agency's past, present, and future orbital and landed missions to Mars, Venus, and the Moon. Given the growth of missions and data sets over the past 20 years coupled with missions planned through 2015 (Figure 1), a Data Management System is being developed to manage and maintain the Geosciences Node data holdings.

In this paper we will describe the rationale for the work and specify the high-level requirements for the Data Management System, and then focus on the use and selection of digital signatures as a mechanism for maintaining file integrity.

Rationale

There are a number of motivations for developing a system to maintain and track the file integrity within the PDS Geosciences Node data archive. First, the system would be useful for tracking and validating electronic deliveries from data suppliers. It would provide a tool for protecting the on-line repository at the Geosciences Node against file loss or corruption. Finally, information from the system, such as file size, checksum signatures, etc., could be provided to the data users who electronically download data from the Node to ensure that the transfer was complete and accurate.

The PDS Geosciences Node is responsible for archiving and distributing a large amount of planetary geosciences data. Our online repository currently holds 143 data sets derived from 15 different missions. These data sets comprise approximately 8 TB of data (Figure 2). The amount of data in the Geosciences Node repository will more than double in the near future as the Mars Reconnaissance Orbiter (MRO) and Lunar Reconnaissance Orbiter (LRO) missions begin to deliver data to the PDS. The large amount of data and number of files present new challenges to maintaining the archive and ensuring its integrity that were not encountered when the archives were small enough to be mass-produced on CD-ROMs and shipped to hundreds of science users. Managing the planetary geosciences data archives is even more demanding considering that the data sets from several on-going missions continue to grow. Active missions typically revise and redeliver parts of or entire data sets during their lifetime, making configuration control important to ascertain that the proper versions of products are housed within the archive and made available to users. Therefore, a system that can track the inventory of data products and ancillary files and test for possible corruption of data files is essential for maintaining the reliability the Geosciences Node archives.

Requirements

This section describes the high-level requirements levied on the data management system.

Maintain inventory and state information. An integral part of the system is the data management database which must contain an inventory of the archive and associated state information for each data set. The inventory is a listing of all components of a data set—date products, product labels, and ancillary files—along with product type, format, file location, size, and digital signature. The state information includes backup status, and, in a future release of the system, data access permissions.

Check in new or updated data. Instrument teams are responsible for providing data products and ancillary files to the PDS on a mission-defined schedule. In some cases, multiple deliveries are made during the life of the mission. The DMS must be able to check in new or updated data so that the inventory and state information are up to date.

This process includes verifying that all files are present and that the contents of the files match what was sent by the data provider.

Maintain integrity of the online repository. Released data are placed in the online repository for access by planetary scientists and the general public. It is incumbent on the PDS that the integrity of the archive be checked regularly. This check includes automated checks of the manifest (filename, location, and size), content (via digital signature), and access permissions. Manual integrity checks of portions of a data set should be available. A report of the results should be produced automatically.

Track backups of the data. Archive data and ancillary files must be backed up regularly, with a copy available onsite and offsite. In addition, simulated restores should be carried out, including integrity checks on the sample restorations.

Provide the location and digital signature of data on demand. The DMS should provide several methods for acquiring archived files and digital signatures. The URL and local file system location of a file, as well as its digital signature, should be made available programmatically. Multiple files and signatures should be packaged on the fly into standard formats such as tar and zip when requested.

Approach

Now we focus in on the use of digital signatures within the system as a method of tracking file integrity.

We propose the use of digital signatures throughout the archive cycle as a mechanism for maintaining file integrity (figure 2). Digital signatures of individual data products and associated files can be generated using standard cryptography algorithms in order to detect bit-level differences between multiple instances of the data. Changes to one or more bits of a file would be reflected in a corresponding change to that file's digital signature.

For this model, in which data flows from the mission instrument team to the PDS, the digital signature is generated by the data producer using a freely-available hashing algorithm. One signature is generated for each file, whether a data product or ancillary file (detached PDS label, document, or software code). A manifest of files and digital signatures would be sent to the PDS archiving facility as part of the standard delivery.

The PDS archiving facility checks the digital signatures against the set of files received. When a digital signature does not match the original file, both the file and signature are retransmitted by the data provider. Once validated, archive files are placed into online storage, and the digital signatures are loaded into a database for long-term archive maintenance. In addition, the archive files are backed up to long-term storage media, such as a backup tape set or secondary hard disk media. Two copies of digital signatures are backed up in separate locations.

Maintenance checks are an integral part of the archive process. In this model, online archive files are checked against digital signatures stored in the database. In the event of a mismatch, the signature is checked against the backup copy of the file and

the online copy is replaced with the backup upon successful matching. If the digital signature in the database does not match either the online or backup copy of the archive file, then the three copies of the digital signature are compared, and the database version is replaced if necessary.

With proper scheduling, an instance of three different digital signatures should not occur for a single archive file. Also, it is important to regularly refresh backup sets to minimize loss due to unreadable media.

Hashing Algorithms

Several hashing algorithms are commonly used in the computer industry as a means of secure communications. These algorithms use one-way encryption techniques to authenticate digital signatures and other content. Although a number of these algorithms have been shown to have security flaws, they remain a viable method for detecting file changes, and therefore, are a key tool for maintaining the integrity of data archives.

Hashing algorithms apply a key against a data source to produce a nearly-unique signature, or hash. The successful use of cryptography algorithms for file change detection lies in the fact that a small change in the source will result in a notable change in the signature (Table 1).

Table 1. Example MD5 hash for similar input strings.

Phrase	Hash
"Long-term Data Preservation"	8D4CBED173FC2FB6B02AE668990D728A
"Long term Data Preservation"	9DB0C2E108E0EE7F930D390E5BAA5182

There are a number of commonly-used algorithms from which to choose, including MD5, SHA, and RIPEMD. Multiple versions of SHA and RIPEMD are available, with bit lengths being the delineating factor. A number of factors were considered in selecting an algorithm for this application:

Availability. MD5, SHA, and RIPEMD algorithms are commonly used for generating digital signatures. As a result source code and executables (binaries) are available for many platforms at little or no cost.

Speed. Both MD5 [Rivest, 1992] and SHA-1 [NIST, 1995] algorithms were tested by generating signatures for the Clementine Long-Wave Infrared data set which consists of 213,129 files in 4,527 directories. The test server was running under normal operating conditions. MD5 signatures for the data set were created in 1 hour and 37 minutes. SHA-1 signatures were created in slightly more than an hour. In both cases signatures were written to an output file and were not read into a database or compared against an existing set of signatures.

Hash size. MD5 hashes are 32 bytes in length. SHA-1, SHA-256, and SHA-512 hashes are 40, 64, and 128 bytes respectively. The RIPEMD-160, RIPEMD-256, and RIPEMD-320 hashes are 40, 64, and 80 bytes respectively. Longer hash sizes require more storage space within the data management database.

Ease of use. Coding both the MD5 and SHA-1 algorithms is straightforward. In addition, binaries are easily downloaded or compiled from source code on most operating system platforms. This is an attractive feature for the distributed nature of the data provider/PDS system.

Security. The MD5, SHA-1, and RIPEMD algorithms have been "broken," i.e., collisions have been shown such that a file may be modified without a corresponding change in the digital signature [Wang et al., 2004, 2005]. In addition, the National Institute of Standards and Technology has not approved MD5 as a secure hashing algorithm and MD5 is starting to be phased out [Kaminsky, 2004].

Nevertheless, we consider this risk negligible for the purpose of verifying file contents.[1] It is possible to use a longer hash such as SHA-512 or RIPEMD-320. Collisions are less likely in this case, but the computational cost of the longer hash outweighs the benefit of greater security.

Based on these factors and given the requirements for this system, we have selected the SHA-1 hashing algorithm for producing digital signatures.

4. Future Work and Conclusion

Our next step is to carry out a file integrity case study on several existing data sets. The data sets are the Clementine Long-Wave Infrared data set, Mars Odyssey Gamma Ray Spectroscopy data set, and the Mars Exploration Rovers Panoramic Camera science data set. We intend to populate a database with digital signatures and then perform a verification test to compare the stored signatures with those generated on the fly. The target metrics include database size, time to populate the database, time to perform the verification, and any mismatches in the signatures.

The use of digital signatures for file integrity is a part of our overall plan to create a Data Management System to support archive operations at the Planetary Data System Geosciences Node. Initial tests indicate that the low computational and temporal costs make digital signatures a viable part of a long-term archive solution.

5. References and Acknowledgements

—: Specifications for Secure Hash Standard. Federal Information Processing Standards Publication 180-1, National Institute of Standards and Technology (1995)

[1] Wang, Yin, and Yu (2005) shows that 2^{69} hash computations are required to find a collision in the SHA-1 algorithm, less than the theoretical 2^{80} computation upper limit.

Guinness, E.A., Arvidson, R.E., Slavney, S.: The Planetary Data System Geosciences Node. In: Planetary and Space Science, Vol. 44, No. 1, Elsevier, London (1996) 13-22
Kaminsky, D.: MD5 to Be Considered Harmful Someday. Avaya White Paper (2004)
Rivest, R.L.: The MD5 Message-Digest Algorithm, Network Working Group Request for Comments: 1321. MIT Laboratory for Computer Science and RSA Data Security, Inc. (1992)
Wang, X., Yin, Y., Yu, H.: Finding Collisions in the Full SHA-1. The 25th Annual International Cryptology Conference (2005)
Wang X., Feng, D., Lai, X., Yu, H. Collisions for Hash Functions MD4, MD5, HAVAL-128 and RIPEMD (2004)

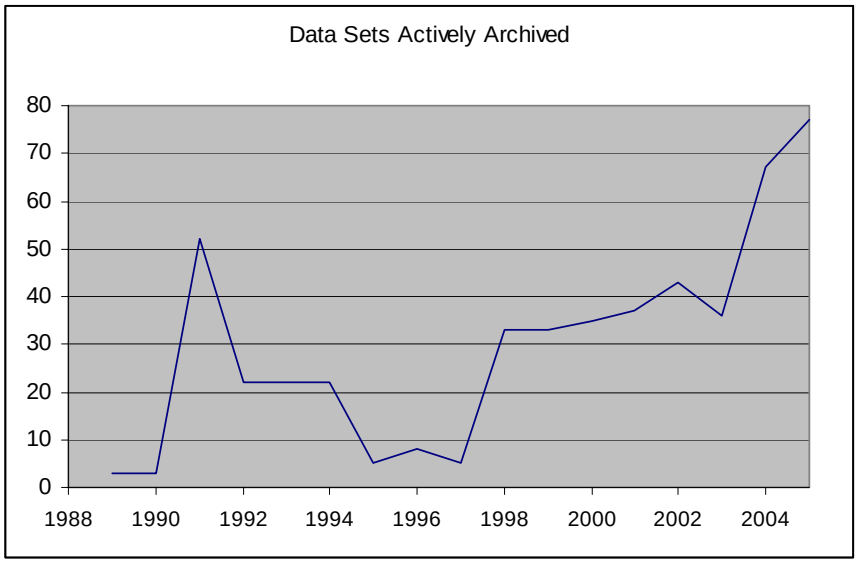

Figure 1. Number of data sets with data being actively archived for each year from 1988 to present. Only data sets being archived are included. Thus, data from the active missions Messenger and Mars Reconnaissance Orbiter are not included

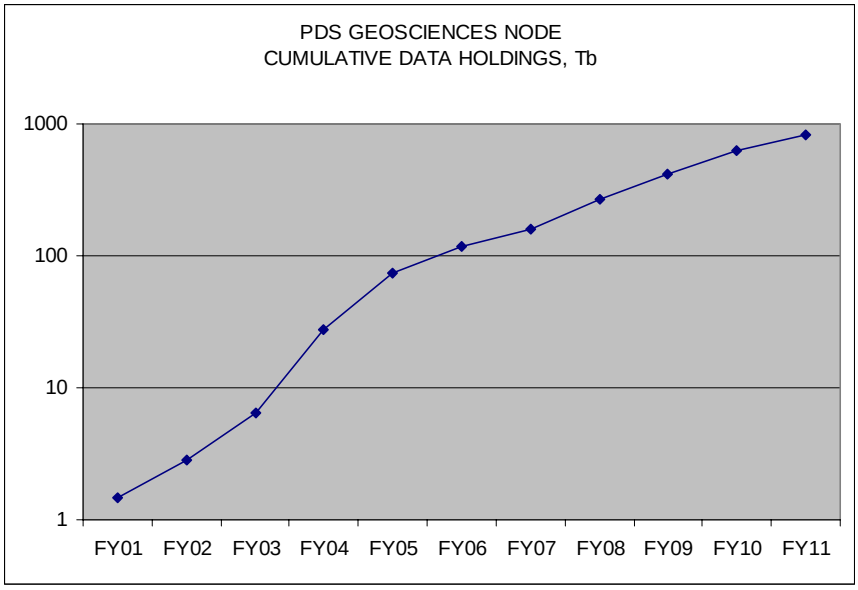

Figure 2. Growth of cumulative data holdings in Tb for the PDS Geosciences Node. The Mars Science Laboratory Mission (2009-2015) is not included. Fiscal Years (FY) are from 2001 through 2011

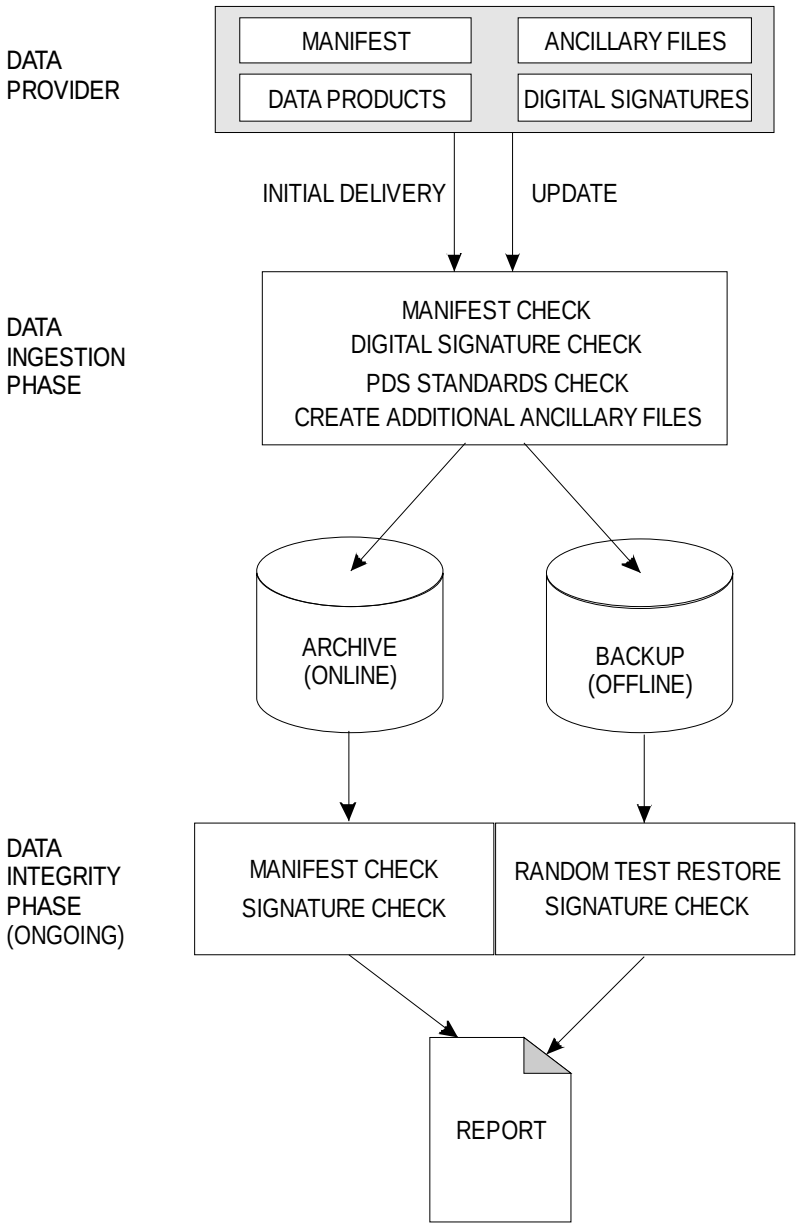

Figure 3. Data integrity checks play a vital role in the data archive life cycle