Standards Change Request

Extend Maximum File Name Length to 40 Characters **SCR3-1144.v2**

Provenance:

Date: 2008-09-16
Author(s): Susan Slavney (Geosciences)
Working Group: Susan Slavney (lead), Mitch Gordon, Ron Joyner, Elizabeth Rye

Date: 2008-09-09
Author(s): Susan Slavney (Geosciences)
Working Group: Susan Slavney (lead), Mitch Gordon, Ron Joyner, Elizabeth Rye
Title: Extend Maximum File Name Length to 40 Characters (SCR3-1144.v1)

Problem:

The MSL Project has requested that PDS increase the maximum length allowed for file names from the present limit of 27 characters plus a 3 character extension (27.3). The MSL file naming convention encodes a great deal of information in EDR file names, and there is not enough room in 27.3 characters.

Current Urgency:

MIPL is currently designing EDR product generation software for MSL. They need to make a decision soon on the file naming convention because it affects all EDR products. The next SIS draft deadline is Sept. 29, 2008.

Proposed Solution:

This SCR proposes to extend the maximum file name length to 44 characters, i.e. 40 characters plus a period plus a 3 character extension. The limit of 40 is chosen because that is the maximum length of the keyword PRODUCT_ID. Many data providers (for MSL and others) follow the common practice of using the file name as the product ID. Limiting the file name to 40 characters will allow it to be used as a valid product ID.

Impact Assessment:

The Geosciences Node has submitted this SCR in the belief that there will be minimal impact on PDS tools and documentation. The proposal was presented at the PDS Management Council teleconference on September 8, 2008, and the Node received permission to proceed from the MC.

PDS Standards Reference – Changes would be made to Standards Reference sections 8, 10, and 11. Specific changes are given below under Requested Changes.

Archive Preparation Guide – The following change would be made to the APG:

- Section 3.2.3 (p. 20), paragraph 4.
  Change:
  > PDS allows a maximum of 31 characters in directory and file names to ensure that all file names are unique.

  to:
  > PDS allows a maximum of 31 characters in directory names and a maximum length of 44 characters in file names to ensure that all file names are unique.

Proposer's Archive Guide – No change is required.

Data Product SIS template on PAG web page – No change is required.

Archive Volume SIS template on PAG web page – No change is required.

Planetary Science Data Dictionary – No change is required.

Software impact – The impact on the following PDS tools has been assessed by querying the PDS Nodes, including software developers at the Engineering Node.

Tool	**Impact**
Vtool	None
LTDTool	None
LVTool	None
NASAView	44-character filename not fully displayed in window banner on Macintosh. 1 week to resolve.
Tbtool	None
Line	None
PDS-D	None
Catalog database	None
Web interfaces	None
Basic Browser	Filename field in database needs to be increased to hold 44 characters. 1 week to resolve.
ATM Tools	None
GEO Tools	None
IMG Tools	None
NAIF Tools	SPICE Archiving Guide needs to be updated to specify

	40.3 file name length. A script that checks file name length needs to be updated. Impact is minor.
Rings Tools	None
SBN Tools	Estimate 1 man-week to fix problems recognized immediately and to conduct a thorough search of all node utilities, databases, and interfaces. Additional time to fix any problems thus identified.

Additional Information:

Three supporting documents are appended for additional information.

1. Proposal To Allow File Names Longer Than 27.3 In MSL Archives. This background information was submitted by the Geosciences Node to the PDS Management Council and discussed at the September 8, 2008, teleconference. It includes the rationale for this SCR and a further proposal to include this topic in plans for the PDS-4 standard.

2. MIPL Justification for Longer File Names. One concern that arose during Standards Working Group discussion was whether the longer file names would really be helpful to users. When the Geosciences Node reported this concern to the MIPL representatives, they submitted this justification for their request.

3. Informal Survey of User Preferences. The Standards Working Group assigned the Geosciences Node the action to conduct an informal survey of MER and Phoenix users to determine their opinions about longer file names and the practice of putting metadata into the file names. The survey results showed a definite user preference for longer file names.

Requested Changes:

The following text is from Sections 8, 10, and 11 of the PDS Standards Reference version 3.7. Proposed deletions are shown in strikethrough text and additions in red text.

8.2 Formation of Directory Names

…

2. Directory names must comply with the ISO 9660 Level 2 standard and not exceed 31 characters in length. Users are encouraged to keep directory names as brief as practical in the interests of providing succinct file paths and easy to read directory listings. The total length of the directory path and file name must not exceed 255 characters.

…

8.3 Path Formation Standard

The PDS standard for path names is based on Level 2 of the ISO 9660 international standard. A pathname may consist of up to eight directory levels. Each directory name is limited to 31 characters; the forward-slash character ("/") is used as the separator in path names. The total length of the directory path and file name must not exceed 255 characters. Path names typically appear on PDS volumes as data in index tables for locating specific files on an archive volume. They may also appear as values in a limited number of keywords (e.g., FILE_SPECIFICATION_NAME, PATH_NAME, and LOGICAL_VOLUME_PATH_NAME).

…

Chapter 10. File Specification and Naming

The *File Specification and Naming* standard defines the PDS conventions for forming file specifications and names. This chapter is based on levels 1 and 2 of the international standard ISO 9660, "Information Processing - Volume and File Structure of CD-ROM for Information Interchange."

ISO 9660 Level 1 versus ISO 9660 Level 2

PDS recommends that archive products delivered on physical media adhere to the ISO 9660 Level 1 specification. Specifically, CD-ROM volumes that are expected to be widely distributed should use file identifiers consisting of a maximum of eight characters in the base name and three characters in the extension (i.e., "8.3" file names), as described in Section 10.1.1. When there are compelling reasons to relax the 8.3 file name standard, the ISO 9660 Level 2 specification with respect to file names may be used, subject to the restrictions listed in Section 10.1.2.

Electronic Transfer and Storage of Archives

The ISO 9660 Level 1 and Level 2 standards are specifically for CD-ROM media. Data providers may negotiate with their PDS Node to deliver archives electronically rather than on CD-ROM. As electronic delivery of archives is becoming a more common practice, PDS is pursuing a more comprehensive standard. Pending the results of this investigation, PDS requires that electronically delivered data meet the restrictions listed in Section 10.1.3.

10.1.1 ISO 9660 Level 1 Specification

…

10.1.2 ISO 9660 Level 2 Specification

…

10.1.3. Specification for Files Delivered Electronically

Electronically delivered files must adhere to the ISO 9660 Level 2 specification with the exception that the base name of the file may be up to 40 characters long. Thus the total file name

length, including a period and 3-character extension, is 44 characters. This format may be referred to as the "40.3" format. The limit of 40 characters is chosen because it is the maximum length of the value of the PRODUCT_ID keyword. As it is a common practice to use the file name as the unique product identifier, this limit will ensure that the file name is not too long to be a PRODUCT_ID.

Notes:

1. The 40.3 format is a looser restriction than the ISO 9660:1988 Level 2 standard.
2. Directory names for electronic delivery must still follow the ISO 1990:1988 Level 2 standard, i.e. they are restricted to 31 characters as described in Section 8.2.
3. The total length of directory path and file name must not exceed 255 characters.
4. The 40.3 rule is specifically for electronic delivery. Files delivered on CD or DVD media must conform to the 27.3 rule.
5. Delivery of files on a "data brick", that is, a computer hard drive that can be mounted directly onto a computer or network, is considered an electronic delivery.

Chapter 11. Media Formats for Data Submission and Archive

This standard identifies the physical media formats to be used for data submission or delivery to the PDS or its science nodes. The PDS expects flight projects to deliver all archive products on magnetic or optical media or by electronic delivery, as ~~. Electronic delivery of modest volumes of special science data products may be~~ negotiated with the science nodes.

Archive Planning - During archive planning, the data producer and PDS will determine the medium (or media) to use for data submission and archiving. This standard lists the media that are most commonly used for submitting data to and subsequently archiving data with the PDS. Delivery of data on media other than those listed here may be negotiated with the PDS on a case-by-case basis.

Physical Media for Archive - For archival products only media that conform to the appropriate International Standards Organization (ISO) standard for physical and logical recording formats may be used.

1. The preferred data delivery medium is the Compact Disk (CD-ROM or CD-Recordable) produced in ISO 9660 format, using Interchange Level 1, subject to the restrictions listed in Section 10.1.1.

2. Compact Disks may be produced in ISO 9660 format using Interchange Level 2, subject to the restrictions listed in Section 10.1.2.

3. Digital Versatile Disk (DVD-ROM or DVD-R) should be produced in UDF-Bridge format (Universal Disc Format) with ISO 9660 Level 1 or Level 2 compatibility.

Because of hardware compatibility and long-term stability issues, the use of 12-inch Write Once Read Many (WORM) disk, 8-mm Exabyte tape, 4-mm DAT tape, Bernoulli Disks, Zip disks,

Syquest disks and Jaz disks is not recommended for archival use. WORM disk formats are proprietary to the specific vender hardware. Helical scan tape (8-mm or 4-mm) is prone to catastrophic read errors. Bernoulli, Zip, Jaz, Syquest and other vendor-specific storage media are prone to obsolescence.

Electronic Transfer and Storage of Archive – The ISO and UDF-Bridge standards mentioned above are specifically for CD-ROM, CD-R, DVD-ROM, and DVD-R media as noted. PDS recognizes that electronic delivery of archives is becoming as common a practice as delivery on physical media, and therefore is pursuing a more comprehensive standard. Pending the results of this investigation, the following restrictions apply to archival products delivered and stored electronically.

6. File names are restricted to 44 characters total length, as described in Section 10.1.3. This is a looser restriction than the ISO 9660:1988 standard.
7. Directory names are restricted to 31 characters total length, as described in Section 8.2. This is the same restriction as the ISO 9660:1988 standard.
8. The total length of directory path and file name must not exceed 255 characters.

…

Appendix 1. Proposal To Allow File Names Longer Than 27.3 In MSL Archives

Summary

The MSL Project has requested that PDS increase the maximum length allowed for file names from the present limit of 27 characters plus a 3 character extension. The Geosciences Node believes the request should be honored as long as there is no impact on PDS tools. The Node proposes a two-step solution: first, choose an intermediate limit to meet the present needs of MSL, and second, upgrade the file name standard for PDS-4 to be compatible with the current ISO standard. If there is general approval from the Management Council, the Geosciences Node will submit an SCR to implement the first step.

Details

Most or all of the MSL EDR products will be generated by MIPL. As was done for the MER and Phoenix missions, MIPL will use a standard file naming convention across all EDR data sets. The naming convention allows a great deal of information to be encoded in the file name for ease of programming and selection of files, such as spacecraft clock, sequence ID, rover motion counter, and various instrument-specific fields. With the increase of complexity of MSL instruments and the expected long duration of the mission, they have run out of room within the 27.3 limit. Justin Maki has sent a formal request to the Geosciences Node manager, Ray Arvidson, to allow longer names in MSL archives.

The Geosciences Node believes PDS should honor the request if it can be done without impact to PDS software tools in this time of transition to PDS-4. We believe that extending the file name maximum length to 40 characters will help the MSL Project and will not impact PDS software. The limit of 40 is chosen because that is the maximum length of the keyword PRODUCT_ID. If MIPL or other MSL producers follow common practice and choose to use the file name as the value for PRODUCT_ID, a limit of 40 characters makes sense.

For the long term, we recommend that the working group(s) developing standards for PDS-4 include a file name standard that is compatible with the latest ISO standard, ISO 9660:1999. The original reasons for adopting the 27.3 standard were (a) to be compatible with the then-current ISO standard, ISO 9660:1998, and (b) to work with certain commercial software for writing CD and DVD volumes. Since that time the ISO standard has been revised to allow longer names, and now that PDS data is seldom if ever distributed or archived on optical media, the software rationale is also obsolete.

With the go-ahead from Management Council, the Geosciences Node will submit an SCR to extend the maximum file name length to 40 characters. The SCR would include a survey of the full technical impact on the nodes, which the Engineering Node has agreed to help coordinate.

Appendix 2. MIPL Justification For Longer File Names

Justification of MSL Request to Increase Maximum Length of File Name
Justin Maki, Doug Alexander
9/13/08

A. Purpose of the MSL Product Filename

1) Uniqueness - We need to be able to assign unique filenames to products for which we want to protect content from being overwritten, at our discretion. This also supports the practice during Operations to temporarily copy/link files away from the mission filesystem to flat "work space" directories in support of ad hoc (off-nominal) processing. In this regard, filenames need to be unique independent of a filesystem's directory path.

2) Human Interpretability - Aside from "uniqueness", the 2nd (of 3) objective of the filename design is to strike a balance between what is humanly intuitive and what is interpretable/parsable by only software.

The Mars landed surface missions (MPF, MPL, MER, PHX, MSL) operate within compressed ("tactical") planning timelines that can ill afford to be without a human-in-the-loop presence. The ad hoc nature of Operations requires a very reactive system over which human intuition presides.

As an example, manually-intensive methods (such as new algorithms) are sometimes developed on-the-fly to resolve anomalies in short order. The more humanly intuitive the system is, including filenames that have a high degree of interpretability, the lower the risk of the human engineer failing to error/quality check (QC) the new methods in the ad hoc environment.

3) Metadata - Aside from "uniqueness" and "human interpretability", the 3rd (of 3) objective of the filename is to strike a balance between metadata stored only in the label and metadata that is more readily available as fields in the filename.

Including some metadata in the filename has been shown during previous Mars landed surface missions to improve the efficiency in ground processing of the file. For the human consumer, the information characterizes the file intuitively. For event-driven software tools that key off of the metadata, the extraction of the information is less CPU-intensive than having to perform label processing to extract keyword/values.

For MER and MSL, there is a core set of metadata information that purports to address the fundamental questions about the file's content:

a) ***When*** was the instrument data acquired? (a spacecraft timestamp such as SCLK)
b) ***Where*** was the instrument data acquired? (rover mobility indices or AZ/EL coordinates)
c) ***Why/How*** was the instrument data acquired? (Command Sequence ID and/or Activity ID and/or Campaign ID, etc)
d) ***What*** was the instrument data acquired? (Instrument ID)

In addition to the above information, other identifiers traditional to the Mars landed surface mission filenames include ***Product Type, Spacecraft ID,*** and ***Producer ID***.

Presence of the above information in the filename facilitates improved methods of searching and sorting the files in both the mission filesystem's directory hierarchy and flat "work space" directories.

4) Ops/Archive Duality - We need the filename designed for Operations use to be the same for Archive use, so that software tools built for Operations will also have success in processing the Archive dataset years from now. In general, Ops needs have always been different, and more demanding, than Archive needs. As a result, the filename may appear needlessly complex to the PDS archivist, but rich in critical information to the Ops user.

B. Why a Longer Filename than 27.3 for MSL ?

MSL is very MER-like in its requirements, and baselines a large amount of MER heritage in many areas of design. This includes the product filename convention, since the MER filename has proven to work well. All fields in the MER filename have been shown to be valuable when one applies the aforementioned rationale that the filename should be "unique", "humanly intuitive", and "metadata serviceable".

As it is, the MER filename just barely fits in 27.3. The MSL filename must expand beyond MER's 27.3 conventions, and here is why:

1) MSL is a much larger mission with larger requirements. This leads to expansion of many of the filename fields:

- SCLK (9 positions in MER filename) needs subsecond resolution to accommodate video frames, which MER did not have.

- Site counter needs 3 characters instead of 2 (for MSL, more sites are anticipated).

- Position counter needs 4 characters instead of 2 (in several instances, MER overflowed the maximum range limits enabled by a very cryptic base-36 encoding of 2 chars)

- Sequence ID has expanded from 5 to 8 characters.

2) MER lessons learned show the need for another character (Special Processing flag) to ensure unique filenames for products produced in off-nominal ways.

3) In order to accommodate the larger MSL requirements within 27.3, severe compression of the existing fields would be necessary.

   For example, nearly all numeric fields (SCLK, rover mobility indices, etc.) would have to be **base-36 encoded**. The Product Type would have to compressed from 3 characters to 2, compressing the alpha portion of Sequence ID from 3 to 2 chars (breaking the mapping to the Sequence nomenclature familiar to Ops personnel), and removing the Venue field important in discerning Flight from Testbed data sources.

4) Any cryptic encoding and/or compressions make it exceptionally difficult for humans to interpret the filename, a detriment to risk minimization during times of ad hoc anomaly recovery involving new development (previously described above).

A new proposed MSL filename length extension would allow these fields to expand to retain their uniqueness, human readability, searchability and sortability.

We can work with 36.3 or greater.

Appendix 3. Informal Survey of User Preferences

The following email was sent to 14 users of MER and Phoenix data known to us at the Geosciences Node. As Phoenix data has not been publicly released yet, the users we polled are necessarily people involved in the mission, mostly students. MER users were both students and scientists inside and outside the mission whom we know to have experience using the data.

Survey Question:

> Hi,
>
> You've been selected as a user of MER and/or Phoenix data to provide your input on a highly unscientific and informal survey. The PDS is currently considering whether to extend the maximum length of file names permitted in an archive. The present limit is 27 characters plus a 3-character extension. The MSL Mission would like to have more characters available for their file names. Like the MER and Phoenix data providers, MSL plans to encode a lot of information in the file name such as instrument ID, spacecraft clock time, sol number, product type, and other things. Having more characters would allow more information to be included in the file name, and perhaps some separator characters to make the name easier to read.
>
> PDS is trying to assess the impact of extending the length to 40 characters. We're trying to move quickly to accommodate the MSL software developers, and therefore we don't have time for a more structured survey, but we would like some input from MER and Phoenix users on how they like the file names used in those archives. Specifically, do you use the information encoded in the file name? Would more characters be an improvement or not, in your opinion?
>
> Please send me your thoughts by Tuesday, Sept. 16. And feel free to forward this to other interested users.
>
> Thanks,
>
> Susie Slavney
>
> ---
> Ms. Susan Slavney (susan.slavney@wustl.edu)
> Washington University
> One Brookings Drive, Campus Box 1169
> St. Louis, Missouri 63130

314-935-9295 Fax: 314-935-4998

Geosciences Node, NASA Planetary Data System
http://pds-geosciences.wustl.edu

Survey Results:

Of the 14 people polled, 9 replied as of this writing. All were in favor of more characters in the file name. Below are excerpts of their replies. Complete responses are available from the Geosciences Node.

“I'm all for additional information in file names, so 40 characters would work for me. That is, as long as we don't break existing software (or software that can't be easily modified).”

“Speaking from my MER experience, I did use the information in the filename, and think an expansion to 40 chars (or at least more than 27.3) would be a useful thing.”

“40.3 is a whole lot better than 27.3. 80.3 is even better, and this is what I use in my databases.”

“Completely agree with [the previous comment]. I would say the number of characters in the file name should be limited so that you can print out the filename in one line. 80.3 is too long there, e.g. 60.3 would be better (for the old DOS 80 columns). The infos you mentioned (instrument ID, spacecraft clock time, sol number, product type) will be very useful for sieving through the lists without opening the files.”

“Since I primarily use windows, the only thing that I would like to retain is the 3 character extension. The 80 characters that [another user] suggested seems a bit much, but I don't have a problem with changing from 27 to 40.”

“Everything in the filenames for Phoenix data is essential to me: the sol number, 4-character ops token, the sclk, the three-character product type code, the version type, etc. I really think that the MSL mission filenames should have at least the information in Phoenix. I also think it is worth trying out more separator characters. Whether or not the filename should be extended further than that would really depend on what information the extension would give.”

“Having additional separator characters would be a huge benefit, so if extending the length means being able to include more of these, I would definitely appreciate it. On the other hand, having a longer block of characters without a significant number of

separators would be difficult to work with, so the number of characters for additional information would still need to be somewhat limited for adequate spacing."

"The more info in the names, the better!"

"With MER, many of my processing scripts make use of information encoded in filenames. In fact, the type of information that you mention (sol, local time, etc.) is what we use when we reformat the PDS-format files for Mini-TES (and Pancam)...and rename the files to something more practical. It is a simple set of commands that allow one to split a filename apart and construct the relevant arguments (say for a 'retrieval' command or from a calculation of radiative flux from a particular Mini-TES temperature profile) without needing to actually grab information out of the header file. In addition, from the point of view of file organization and use, longer filenames that allowed something more than the cryptic scheme used for MER might be sufficiently beneficial that I would be tempted to not store the files in another format with a more useful name string. So, basically, the answer to your two questions are 'yes' and 'yes'."