EMC²
where information lives®

NDMP Localization/Internationalization Support for NetWorker –A Detailed Overview

EMC Proven Professional Knowledge Sharing 2009

Implementation Engineer, Backup and Recovery Specialist Version 4.0 (EMCIE)

Jyothi A.Deranna
Senior Software Quality Engineer
NetWorker QA, SSG India

EMC² PROVEN PROFESSIONAL

Table of Contents

Disclaimer: The views, processes or methodologies published in this compilation are those of the authors. They do not necessarily reflect EMC Corporation's views, processes, or methodologies

Disclaimer: The views, processes or methodologies published in this compilation are those of the authors. They do not necessarily reflect EMC Corporation's views, processes, or methodologies

Introduction

Computer internationalization and localization are important due to the numerous differences that exist among countries, regions and cultures with respect to language (not only distinct languages but also dialects and other differences within a single language). Also, there are differences in weights and measures, currency, date and time formats, names and titles, citizen identification numbering systems, telephone numbers, addresses and postal codes, religious, cultural and political sensitivities, profanity and legal systems.

NetWorker® 7.4 offers internationalization (I18N) enhancements to support multilocale data zones and improved usability of an I18N NetWorker and/or localized NetWorker. NetWorker 7.4 enables the support of remote operations involving non-ASCII data across machines running in different locales (for example, scheduled backup of non-ASCII savesets). In addition, the legacy client GUIs are updated to work in a multilocale data zone.

The Network Data Management Protocol (NDMP) is a TCP/IP-based protocol that specifies how network components communicate with one another for the purpose of moving data across the network for backup and recovery.

EMC® NetWorker's NDMP client connection feature provides NAS vendors fast, flexible backup and restore of mission-critical data residing on filers. NetWorker with NDMP client connections provide backup and recovery support for more than eight NAS hardware providers including EMC, NetApp, Auspex, Netforce, BlueArc, and others.

The NDMP localization support partially existed with NetWorker since the 6.1.3 release to support the NDMP vendor's requirement. With the NetWorker 7.4 release, during which NetWorker was fully internationalized, there were some major changes in NDMP localization support as well.

Different NAS vendors have their own mechanism to store the non-ASCII characters in their specific filers. NDMP is a single interface that will help NetWorker understand the way each vendor is handling the non-ASCII data. Using NDMP, NetWorker is able to backup filer's data and store it without data corruption.

This article offers insight on how NetWorker manages to backup and recover non-ASCII data residing in the NAS filers using the NDMP client connection feature. It focuses primarily on EMC Celerra® and Network Appliance filers, although there is no difference in the way NetWorker handles NON-ASCII data using NDMP for all the vendors it supports.

Since each NAS filer has its own way of handling non-ASCII characters, NetWorker administrators must configure NAS filers that differ from vendor to vendor before backing up the non-ASCII file systems. This article provides best practices and troubleshooting tips to avoid data corruption and increase performance.

Audience

The intended audience for this article is administrators dealing with non-ASCII data who are using EMC NetWorker NDMP to back up and recover data to and from their NetApp and/or EMC Celerra file servers. Support and sales staff may also find this paper helpful as a reference.

Terminology

Data server: An NDMP server that transfers data between primary storage and the data connection.

Tape server: An NDMP server that transfers data between secondary storage and the data connection, and allows the Data Management Agent (DMA) to manipulate and access secondary storage.

dump: On EMC Celerra, this is a backup format in PAX that traverses a file tree in mixed width first and depth-first order. On NetApp, this is an inode-based backup that traverses a file tree in directory first and file-based order.

tar: A backup format in PAX that traverses a file tree in depth-first order.

NDMP server: An instance of one or more distinct NDMP services controlled by a single NDMP control connection.

NDMP services: The state machine on the NDMP host accessed with the Internet protocol and controlled using the NDMP protocol. There are three types of NDMP Services: Data Service, Tape Service, and SCSI Service.

16-bit Unicode characters: Also called USC-2 by the Unicode Consortium. It is used to describe the 16-bit Unicode or "wide" characters used by Microsoft Windows NT, Windows 95, Windows 98, Windows XP and Windows 2000, Windows Server 2003. This encoding represents the first 65536 Unicode characters.

character set: The set of characters used, without regard to the font (as seen on a display) or encoding (as located in the storage or protocol message).

encoding: The mapping of individual characters in a character set to the specific bits used to represent the data (used in protocol messages and the way they are stored on disk).

Unicode: The family of universal character encoding standards used for representation of text for computer processing.

Unicode Consortium: The organization responsible for defining the behavior and relationships among Unicode characters, and for providing technical information to implementers (www.unicode.org).

UTF-8: Unicode (or UCS) Transformation Format, multi-byte encoding form. UTF-8 uses an algorithmic mapping scheme to convert every Unicode value to a unique 1-to 4-byte sequence, with no embedded null characters.

Internationalization: (a.k.a I18N) is a way of designing and producing products that can be easily adapted to different languages/countries/regions. This requires extracting all language, country/regional and culturally dependent elements from a product, and making them configurable.

Localization: (a.k.a L10N) is the process of creating or adapting a product to a specific target language/country/region, i.e., to the language, cultural context, conventions and market requirements

Locale: the subset of a user's environment that depends on language and cultural conventions. These include not just character sets, but also time/date, money, etc. formats.

Multi-byte Character Set (MBCS): a character set that uses a variable number of bytes per character. Each locale has its own MBCS definition, with no regard for other locales. A code point can represent different characters in different MBCS's. You must know the MBCS (locale) to interpret the code point as a character.

NMC: NetWorker Management Console

MBT: MBT encoding is a NetWorker-specific UTF-8 compatible encoding used to preserve multi-byte data.

NDMP I18N/L10N Backup and Recovery Workflow

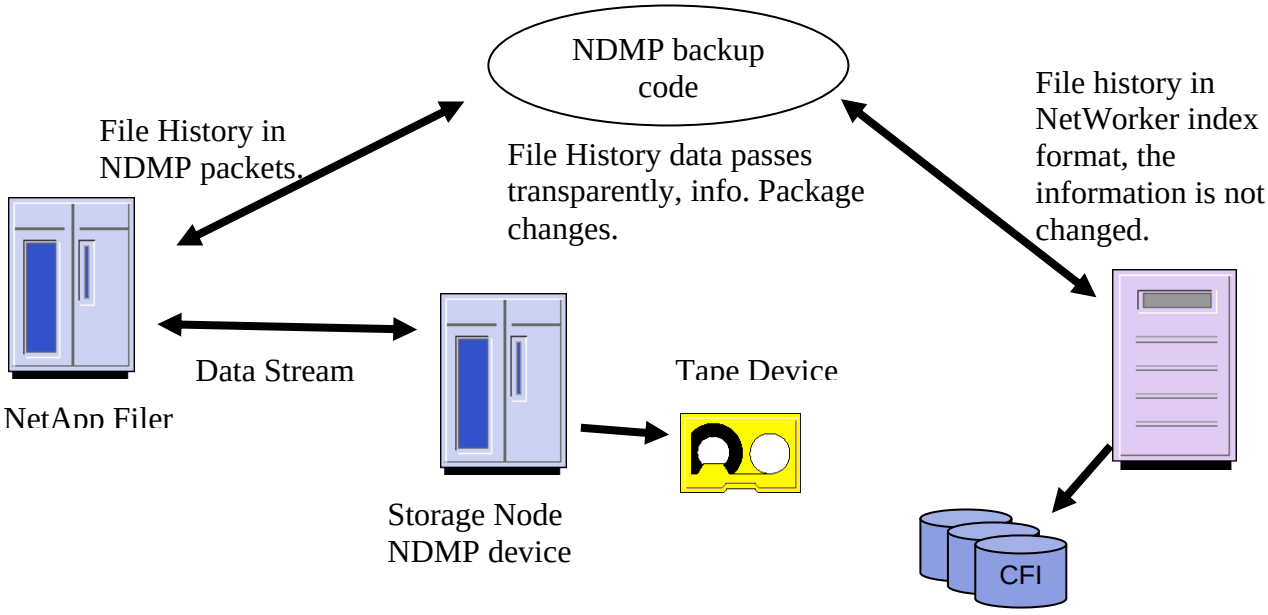
The 118N feature was introduced in NetWorker during the 5.0 release to handle non-ASCII data sets. Since then, it has been enhanced with each release. The NetWorker 7.4 release included major I18N enhancements with multilocale datazone support. In line with these enhancements on the core NetWorker side, there have been also been enhancements of the I18N feature for NDMP backup/recovery. NDMP support for non-ASCII characters for NDMP backup and recover began with NetWorker 6.1.3. This was done to overcome the pitfalls faced during NDMP backup of non-ASCII data of NAS vendors with pre-6.1.3 release.

Pre-6.1.3 NDMP Backup/Recovery workflow

After introducing 18N support, NetWorker began storing each and every character string in UTF-8 format including NetWorker's CFI database. While browsing the database for a file by file / index recover, the 'nwrecover'/recover binary started reading the filenames in the CFI database as UTF-8 strings. However, the strings in the CFI database were not always in UTF-8 encoding for NDMP backups. The format of file index content was dependent on the locale value set at the filer level. If

the encoding was not set to UTF-8 format in Filer, it resulted in incorrect parsing of the file names by the NetWorker index browser.

Figure 1: NDMP Backup/recovery workflow with pre 6.1.3

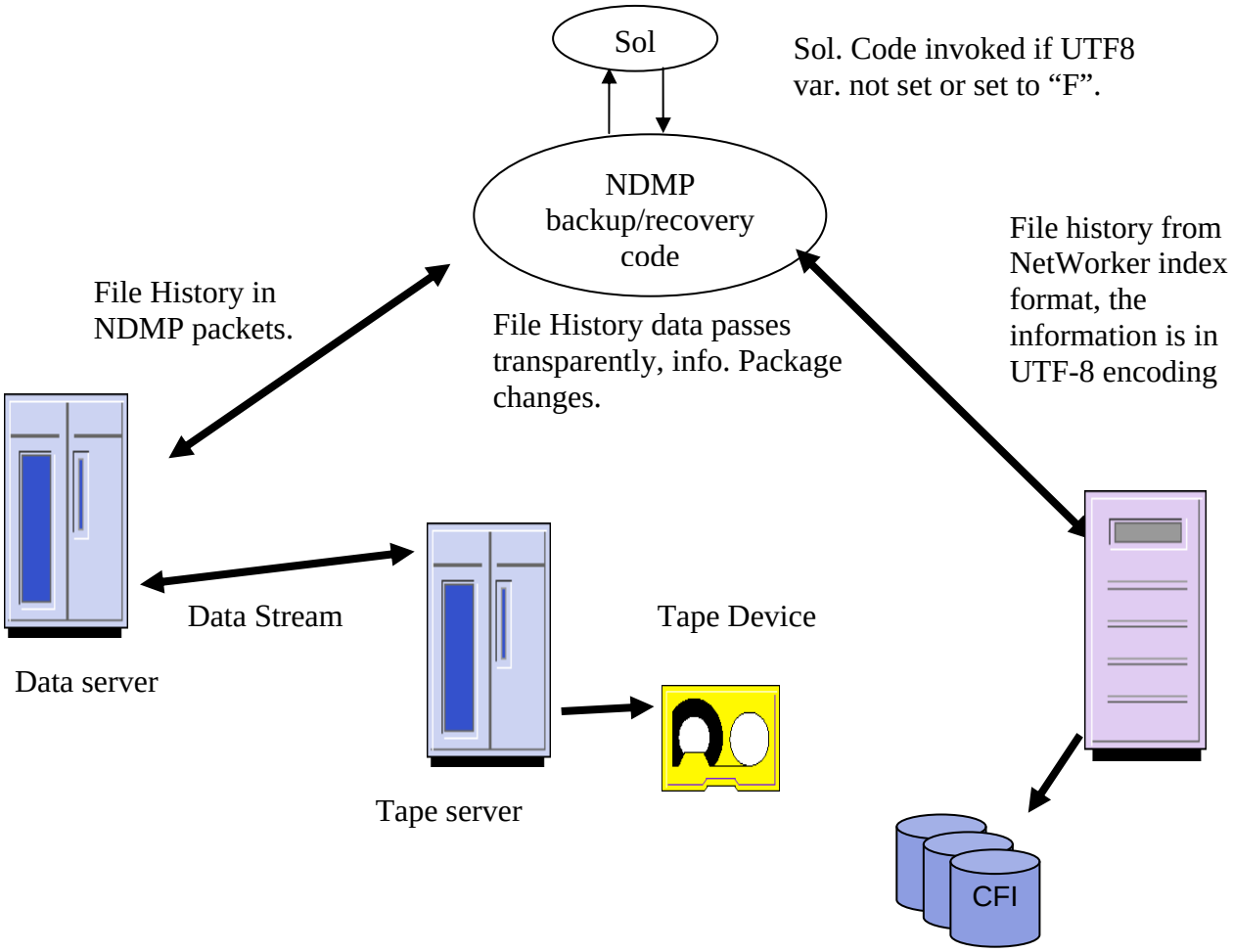


The NDMP gets the file history information from the NetApp Filer in NDMP packets. This information is handled by the NDMP, which transforms the packaging of the information (NDMP Packet structure) to NetWorker compliant formatting (NetWorker save record entries). The important point is that the NDMP Bridge does not attempt to interpret the values in the NDMP packets. The information given by the NetApp Filer is passed to the NetWorker Client Indexes without changes.

The problem stems from the fact that the file names stored in the CFI database, for a NDMP backup, are not compliant with the way NetWorker stores the names for non-NDMP backups. File names generated from the NDMP backups are stored in the native format sent by the filer, and may or may not be in UTF-8 format. In comparison, non-NDMP backups always result in the filenames being stored in UTF-8 format.

The solution was to convert the non UTF-8 coding sent by the NetApp filer to a UTF-8 format before committing to the CFI. As described in the Figure 2, the "NDMP" would also transform the incoming character strings to UTF-8 coding.

Figure 2 illustrates the workflow of NDMP backup/recovery during NetWorker 6.1.3



As shown in the Figure 2, the incoming stream is a stream of non-ASCII data which is not in UTF8 format (For example, Japanese characters in native format). This would then be converted to its UTF-8 equivalent. This behavior is the default behavior of NetWorker during NDMP backups. On the other hand, a file history stream that already is in UTF-8 encoding would be properly handled only if an environment variable "UTF8" is set in the "Application Info" of the client resource. So if the filer is sending out the file history information in UTF8 format, NetWorker can be forced to skip the conversion attempt of the file history stream by setting the "UTF8" environment variable to "T."

Absence of the environment variable or forcing the environment variable to “F” would force “NDMP” to carry out conversion for every character string in the file history information. Therefore, this part of the solution to handle UTF-8 compliant streams from the filer would require a manual override specifying the environment variable for the save set.

Algorithm

nsrndmp_save starts nsrndmp_2fh
nsrndmp_save talks to the Filer and starts the backup
Sends the save set name that needs to be backed up
nsrndmp_save passes on the file entries to nsrndmp_2fh for index processing.
nsrndmp_2fh checks the value of UTF-8 client attribute
If UTF-8=Y, nsrndmp_2fh does not perform conversion on the file entries.
If UTF-8=N, nsrndmp_2fh performs conversion on the file entries.
nsrndmp_2fh stores the entries in a temporary file.
nsrndmp_save starts nsrdmpix to commit entries to NetWorker.
nsrdmpix reads entries from the temporary file.
nsrdmpix stores the entries in NetWorker Index database

An enhancement to the above solution, during the 7.0 releases, handles the non-ASCII data more efficiently. The functionality of UTF8 variable is extended to dictate the behavior of save set names being sent to NDMP server at the time of backups with the original functionality of dictating the format of file history information. If UTF8=Y, saveset path names would be sent to the NDMP server in UTF8 format. Otherwise, they would be sent in native format. The value of UTF8 would signify whether the data to be backed up (residing on the NDMP server) is in UTF8 format. This data includes not only file entries but also the directories that contain those files.

There is a limitation with the above solution. NetWorker would not handle file history streams if they happen to have both UTF8 and native character encoding. The solution was supported only for the file names in the file history stream with one character-encoding. With the I18N enhancement during the NetWorker 7.3 release, all UNIX filenames will be stored as UTF-8 MBT in the save stream and client file index. In this way, NetWorker can guarantee the reconstruction of the original filenames during recovery.

NetWorker continued to store filenames in UTF-8 for Windows filenames and UTF-8-MBT for UNIX filenames because:

- Encoding information of Unix filenames is preserved in UTF-8-MBT
- UTF-16 encoded Windows filenames can be represented in UTF-8 without any loss of information.
- Ease of cross platform index browsing
- Maximum number of platforms NetWorker can support with this solution

For NDMP backups on UNIX, if the filer is able to guarantee that the file system's filenames are in UTF-8 instead of MBS, save stream and index filenames from NDMP backups were stored in UTF-8. Otherwise, they were stored in UTF-8-MBT. Since this solution resulted in the failure of some backups later, the save streams and file index entries were stored only in MBT version. Also, the NetWorker Management Console, introduced during 7.3, was enabled to save the MBT version of the Saveset name.

Currently, the UTF8 variable is ignored since all the current NDMP filers that support non-ASCII have UNIX file systems. The conversion is always done to the MBT version for UNIX file names. This is NetWorker's behavior since the 7.3.3 release.

Configuration on Netapp filers

- Supports two names for every file entry on a volume: UNIX name and Windows name
- Can configure volumes using "vol lang" command from the login console. (need to reboot the filer after this step)
  For example: vol lang zh set the volume to the locale 'zh'
- The setting allows generation of "UNIX name" if "Windows name" is given or vice versa.
  - If files are created from Windows then the "UNIX name" would be generated using the "vol lang" setting.
  - If files are created from UNIX then the "Windows name" would be generated using the "vol lang" setting.

- Names are generated on first access of that name.
  - UNIX name would be generated only if accessed across a NFS.
  - Windows name would be generated only if accessed across a CIFS.
  - Generation is a one time affair, once generated the names are permanent.

- NetApp sends UNIX names during NDMP backups.
  - If files were created from Windows, UNIX names are generated during NDMP backup (if not generated earlier).
  - If files were created from UNIX, Windows names are not generated unless files are accessed using CIFS.

- NetApp does not guarantee that generated names are correct
  - If a Japanese volume has Chinese names (created from Windows) the UNIX names (generated) would be incorrect.

- NetApp does not check that the names created by the user are correct

  If a user creates Chinese filenames on a Japanese volume, NetApp would not alert.

Configuration on Celerra Filers

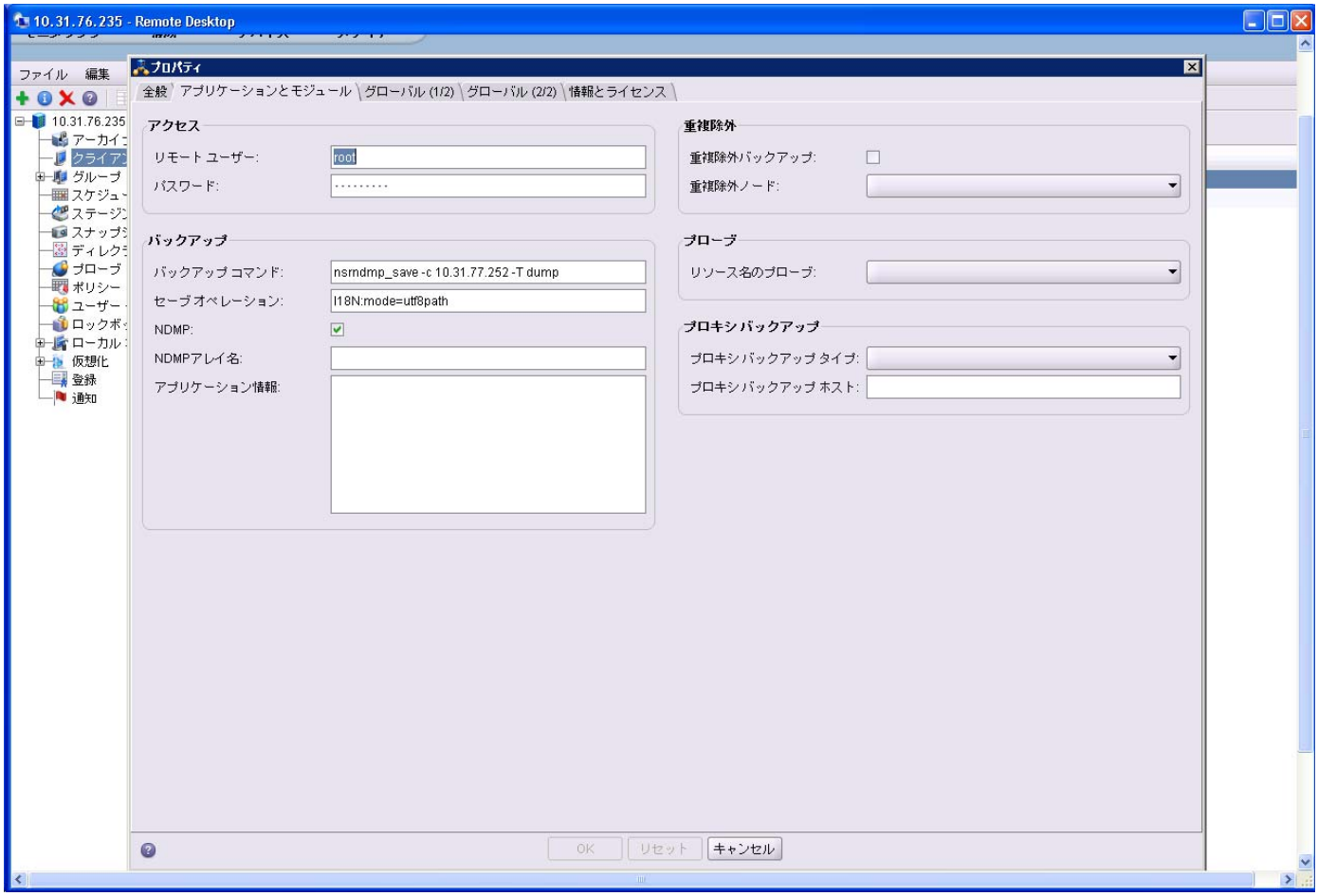
- Supports two names for every entry on a volume
  - UNIX Name and Windows Names.
- Can configure encoding for every client that accesses the system.
  - As opposed to the same encoding for the whole volume (irrespective of the client accessing it) by NetApp.
  - Only one encoding is recommended with one client.
- Control station does not play a role in I18N support. You need to configure individual Data Movers.
- Celerra can be Unicode enabled
  - Using “uc_config -on” command
  - Converts names created from Windows to UTF-8 for UNIX
  - Converts UTF-8 names created from UNIX to Windows compatible names
- If creating non-UTF8 names from UNIX (in addition to enabling Unicode for Celerra)

  - Configuration achieved by means of editing xlt.cfg file (located in /nas/site/locale). Consult Celerra manuals to edit the file.
    - Populate the settings to data movers using “/nas/sbin/uc_config” command.
    - Individual Data Movers can be updated using “/nas/sbin/uc_config” command.
  - If non-UTF8 name and setting in xlt.cfg do not match, names are stored in non-UTF8 format and Windows name is corrupt
- The setting allows generation of “UNIX” name” if “Windows name” is given or vice versa.
  - If files are created from windows then the “UNIX name” would be generated in UTF-8.
  - If files are created from UNIX then the “Windows name” would be generated using the setting for the client (as specified in xlt.cfg file). In addition, Celerra will store the UNIX name in UTF-8 format.
- Alternate names are generated the moment the entry is created.
  - UNIX name is generated as soon as Windows name is created.
  - Windows name is generated as soon as UNIX name is created.
  - Different from NetApp, where generation happens during the first access.
- NDMP Backups support additional parameters for I18N characters
  - /nas/site/slot_param for parameter across all data movers.
  - /nas/site/slot_xx/param for parameter for a specific data mover.
- Following parameters are supported
  - convDialect – Data mover specific, used during file by file recovery, converts file names sent by NetWorker to match the configuration for the data mover, default value English. Set when recovering non-UTF8 file names to a Unicode Data Mover.
  - Dialect – Data mover specific, used during backup, converts file names sent to NetWorker as per the value, default value UTF-8.
- Entries sent to NetWorker during NDMP backups are UNIX names.
  - If Celerra Unicode is enabled, entries are in UTF-8 (unless “dialect” parameter is defined).
  - If “dialect” parameter setting does not match the UNIX name on the volume, the entry is not sent.
  - If Celerra is not Unicode enabled, entries are in whatever format NFS clients stored them.

NetWorker Configuration

For Windows clients, all non-ASCII saveset names are supported, where the “save operation” attribute in the client resource must be specified as “I18N:mode=utf8path”.

Figure 3: NDMP client configuration from windows NetWorker client.

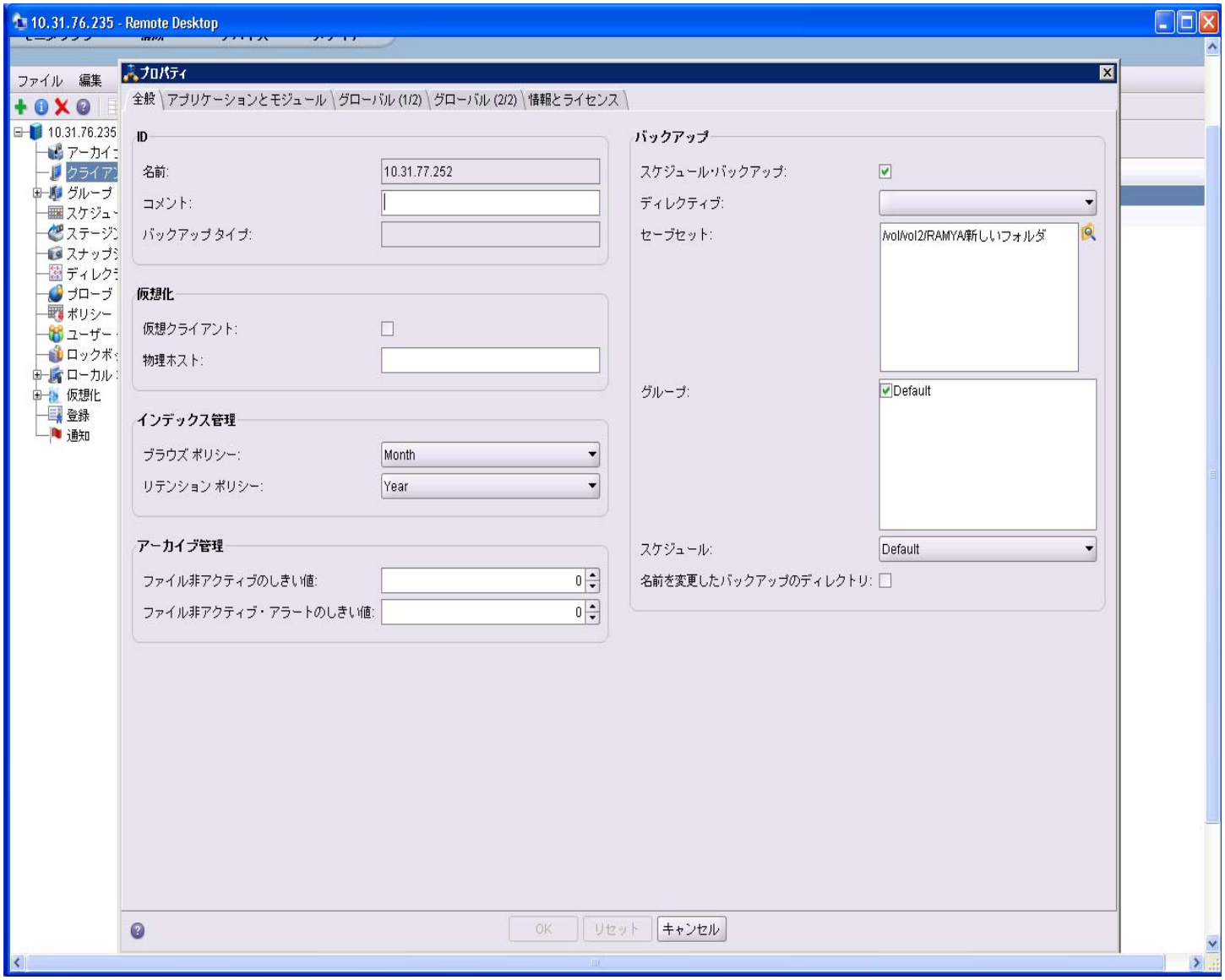


For UNIX clients, the following limitations apply:

- When specifying a non-ASCII saveset value for a UNIX client, the “save operation” attribute in the client resource must be filled with “I18N:mode=nativepath” for 7.4 or later NetWorker clients.
- The save operation value should be “I18N: mode=utf8path” for pre-NetWorker 7.4 clients.

The user should ensure that the saveset path in the NMC is displayed properly for the specific locale. (As described in the Figure 4)

Figure 4: specifying the save set name of the non-ASCII characters using NMC.



Best Practices

NDMP configuration issues may be one reason for the failure of NDMP backup or recovery in your environment. The following list provides reasons and possible solutions should you experience configuration issues.

- NDMP clients are treated as UNIX clients for all practical purposes. So for all UNIX flavored locale settings on the filer (including UTF-8), the NMC client needs to be run on a UNIX client set to the exact same locale setting as the filer.

  For Windows flavored code sets like SJIS, the NMC can be run anywhere (as UNIX allows SJIS to be set as locale) and the filer locale needs to be set to SJIS (ja_JP.PCK*).

  Use these workarounds if the filer setting is only in UTF-8: Run NMC client on an UNIX host set to UTF-8 locale. If a UNIX host is not available, use the NMC client on Windows to configure the client saveset. Specify the ASCII paths in this mode. Changing anything else will break the 7.4 I18N design assumption (do not rely on Customer inputs for determining locale, avoid all unnecessary conversions and preserve the bit patterns through MBT).

  Backup and recovery operations can be run on any locale, but if you try to browse on a locale that is different from the original locale the filenames will show up as random characters.

- Users must be careful while setting the configuration file on Celerra. Data is backed up onto tape using the tar, dump, or vbb NDMP backup type. If a translation configuration file is chosen, the filename is converted to the client encoding and sent to the client as the file history. During an NDMP tar, dump, or vbb backup, if a file was created using a character not appearing in the code page's character set, or a file is found that the translation configuration file cannot translate, the file information (a file name with a random inode number appended) is sent to the backup client to provide a file history. However, this filename cannot be used to restore the file. An error log is created, and the remaining NDMP backup continues normally. A full restore of the directory is required to recover files with catalog information unable to be backed up.

- Enable Unicode on a newly installed Celerra Network Server before creating user files and directories. User files and directories that were created before the system is configured are considered existing files. If there are any non-ASCII file or directory names, they must be converted to Unicode by performing the upgrade procedure.

- While configuring the translation file on Celerra, specify only the number of lines required by your operating environment; performance could be impacted in systems specifying large numbers of translation strings. Also, be sure to remove translation strings that are no longer being used.

- While performing index recovery of NDMP backups, make sure that you change to the target sub-directory to add the files to be recovered. In some scenarios adding the file from the root directory may cause issues.

- NDMP backup of non-ASCII data for SnapImage and CBRM is not supported.

- From version 7.3.3 onwards, the user does not need not to use the UTF8=Y variable in the Appinfo column during NetWorker client configuration for NDMP clients. If the customer upgrades to this version from a previous version, and if the client is configured with this variable, remove it from the client configuration for successful recovery and backup operations.

- Avoid backing up data of other locales from specific locale settings. (For example, if the locale is set to 'jag' make sure that only Japanese data set is given for backup. Avoid having the Chinese saveset within the same backup path).

Conclusion

With I18N support, using EMC NetWorker software and the NDMP service that is running in the NAS filers, users can backup/recover their non-ASCII data. Also, the best practices tips presented in this article can help you to resolve the most common issues encountered in Localization environments.

References

- NDMP website: http://www.ndmp.org
- EMC NetWorker homepage: http://software.emc.com/products/software_az/networker.htm
- EMC Powerlink®: http://Powerlink.EMC.com
- http://ddtsres.legato.com (NetWorker's Bug tracking tool)
- http://neon.legato.com/legato/repos/ (NetWorker's document repository)

Biography