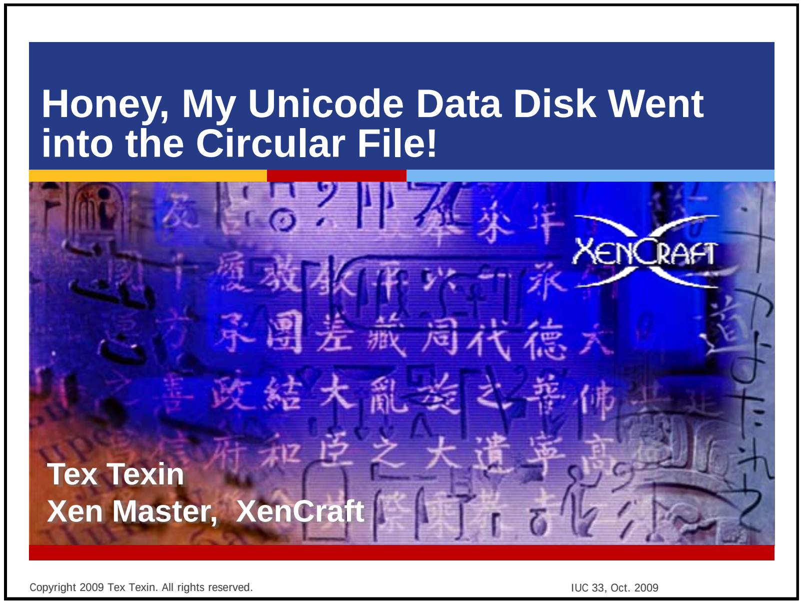
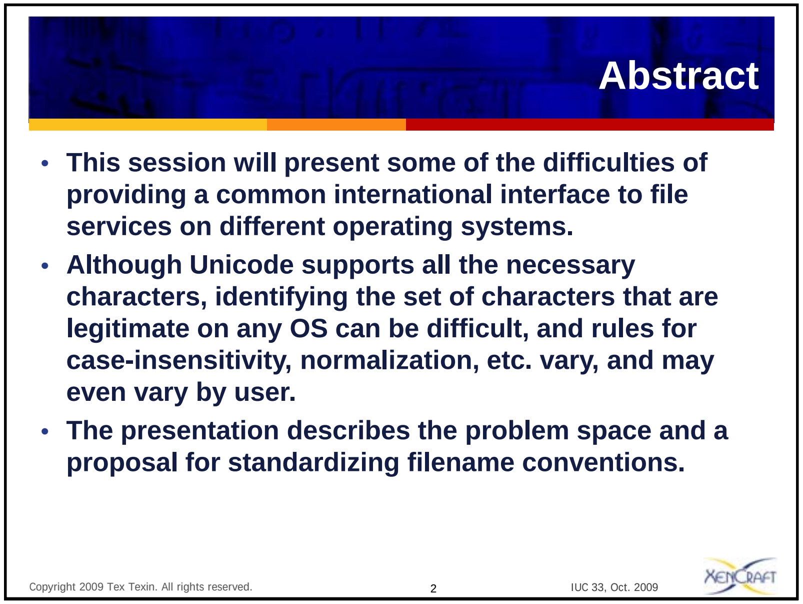


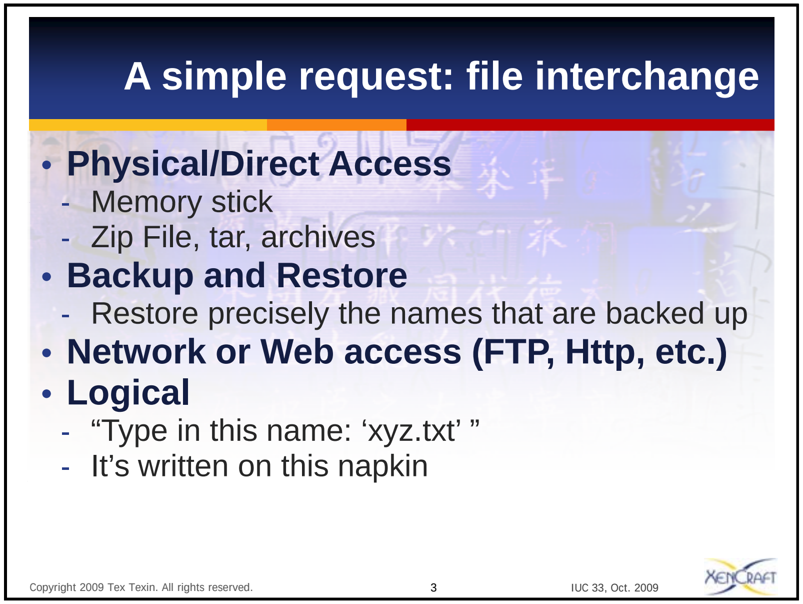
Honey, My Unicode Data Disk Went into the Circular File!

Tex Texin
Xen Master, XenCraft



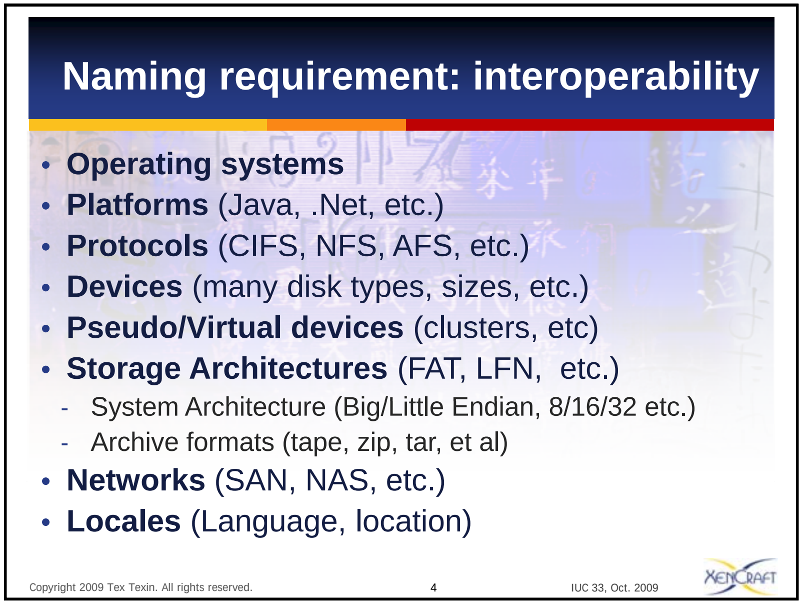
Abstract

- **This session will present some of the difficulties of providing a common international interface to file services on different operating systems.**
- **Although Unicode supports all the necessary characters, identifying the set of characters that are legitimate on any OS can be difficult, and rules for case-insensitivity, normalization, etc. vary, and may even vary by user.**
- **The presentation describes the problem space and a proposal for standardizing filename conventions.**



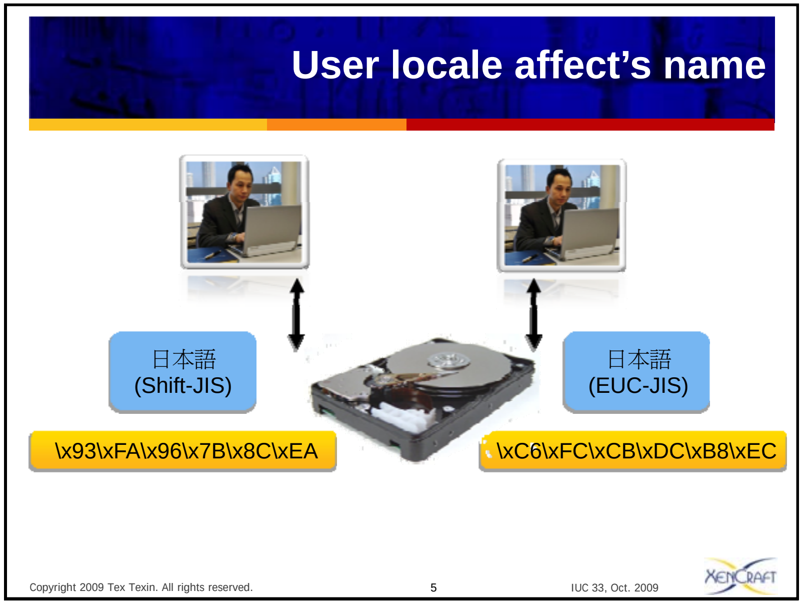
A simple request: file interchange

- **Physical/Direct Access**
  - Memory stick
  - Zip File, tar, archives
- **Backup and Restore**
  - Restore precisely the names that are backed up
- **Network or Web access (FTP, Http, etc.)**
- **Logical**
  - “Type in this name: ‘xyz.txt’ ”
  - It’s written on this napkin



Naming requirement: interoperability

- **Operating systems**
- **Platforms** (Java, .Net, etc.)
- **Protocols** (CIFS, NFS, AFS, etc.)
- **Devices** (many disk types, sizes, etc.)
- **Pseudo/Virtual devices** (clusters, etc)
- **Storage Architectures** (FAT, LFN, etc.)
  - System Architecture (Big/Little Endian, 8/16/32 etc.)
  - Archive formats (tape, zip, tar, et al)
- **Networks** (SAN, NAS, etc.)
- **Locales** (Language, location)



User locale affect’s name

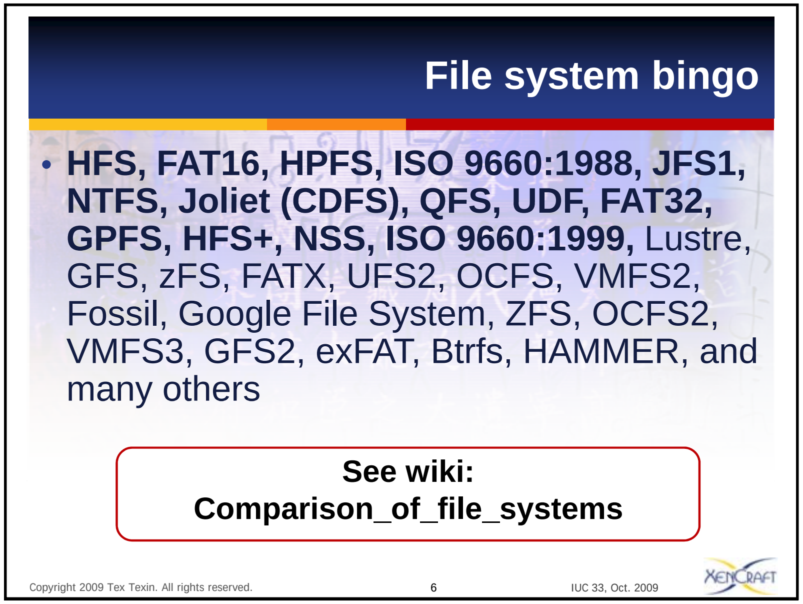
日本語
(Shift-JIS)

\x93\xFA\x96\x7B\x8C\xEA

日本語
(EUC-JIS)

\xC6\xFC\xCB\xDC\xB8\xEC

Copyright 2009 Tex Texin. All rights reserved. 5 IUC 33, Oct. 2009

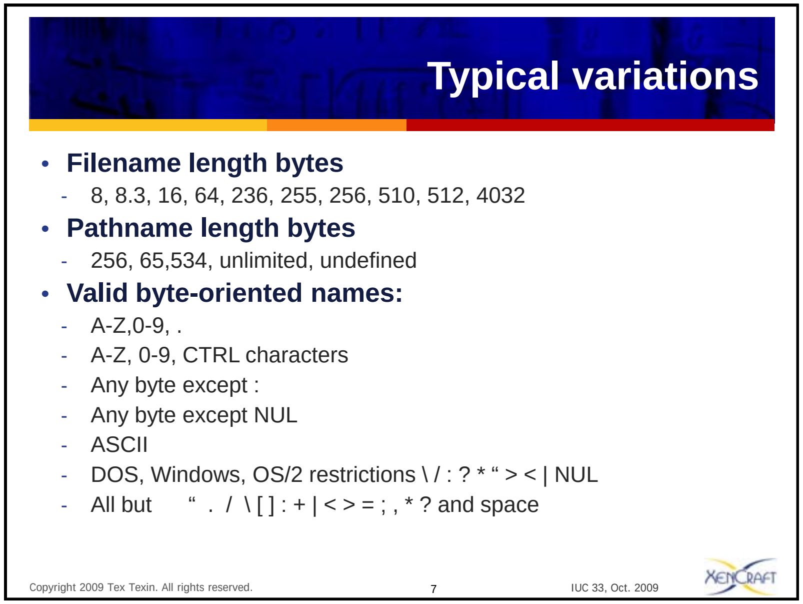


File system bingo

- **HFS, FAT16, HPFS, ISO 9660:1988, JFS1, NTFS, Joliet (CDFS), QFS, UDF, FAT32, GPFS, HFS+, NSS, ISO 9660:1999,** Lustre, GFS, zFS, FATX, UFS2, OCFS, VMFS2, Fossil, Google File System, ZFS, OCFS2, VMFS3, GFS2, exFAT, Btrfs, HAMMER, and many others

**See wiki:
Comparison_of_file_systems**

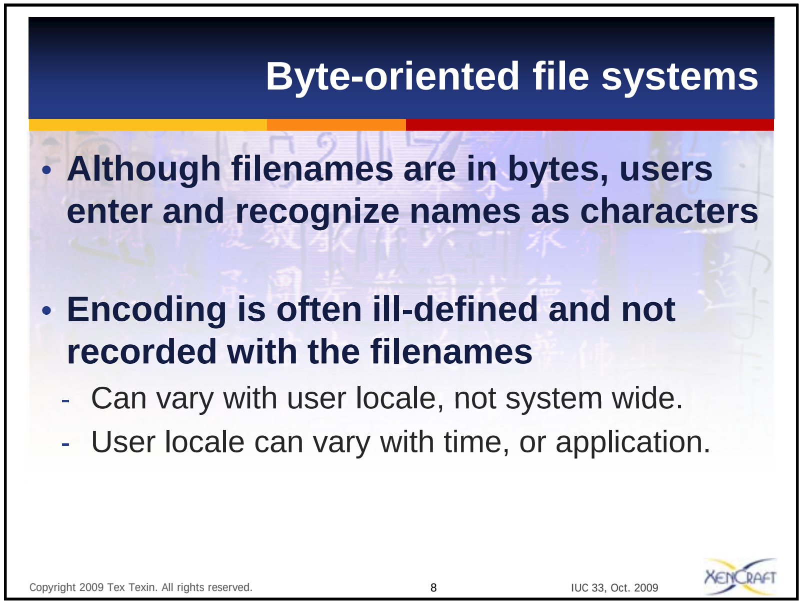
Copyright 2009 Tex Texin. All rights reserved. 6 IUC 33, Oct. 2009



Typical variations

- **Filename length bytes**
  - 8, 8.3, 16, 64, 236, 255, 256, 510, 512, 4032
- **Pathname length bytes**
  - 256, 65,534, unlimited, undefined
- **Valid byte-oriented names:**
  - A-Z,0-9, .
  - A-Z, 0-9, CTRL characters
  - Any byte except :
  - Any byte except NUL
  - ASCII
  - DOS, Windows, OS/2 restrictions \ / : ? * “ > < | NUL
  - All but “ . / \ [ ] : + | < > = ; , * ? and space

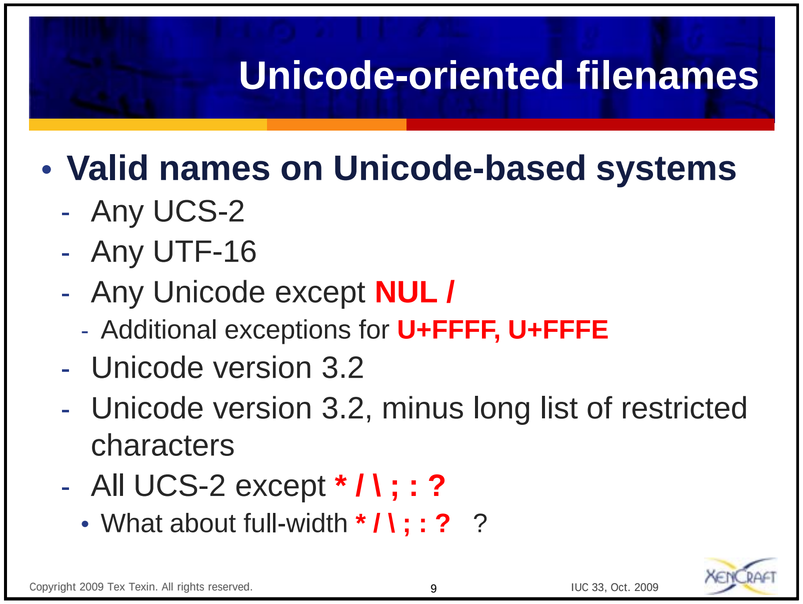
Copyright 2009 Tex Texin. All rights reserved. 7 IUC 33, Oct. 2009



Byte-oriented file systems

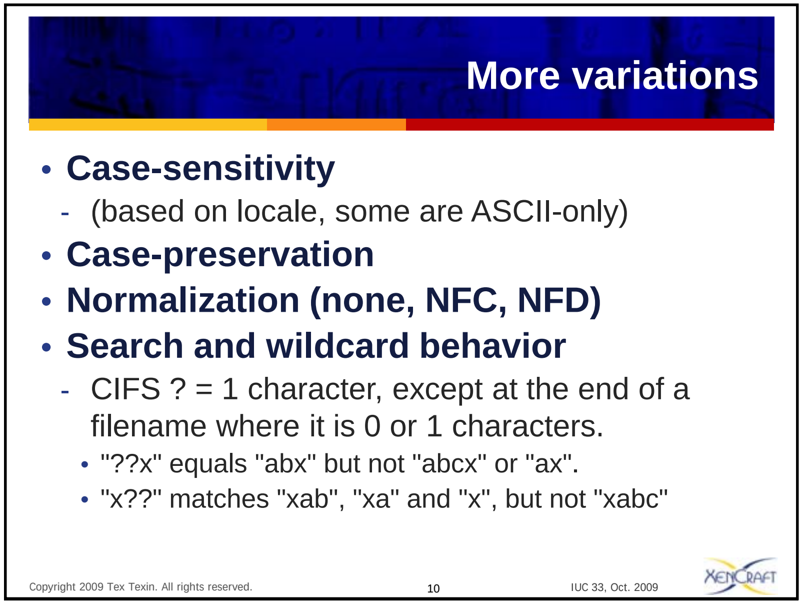
- **Although filenames are in bytes, users enter and recognize names as characters**
- **Encoding is often ill-defined and not recorded with the filenames**
  - Can vary with user locale, not system wide.
  - User locale can vary with time, or application.

Copyright 2009 Tex Texin. All rights reserved. 8 IUC 33, Oct. 2009



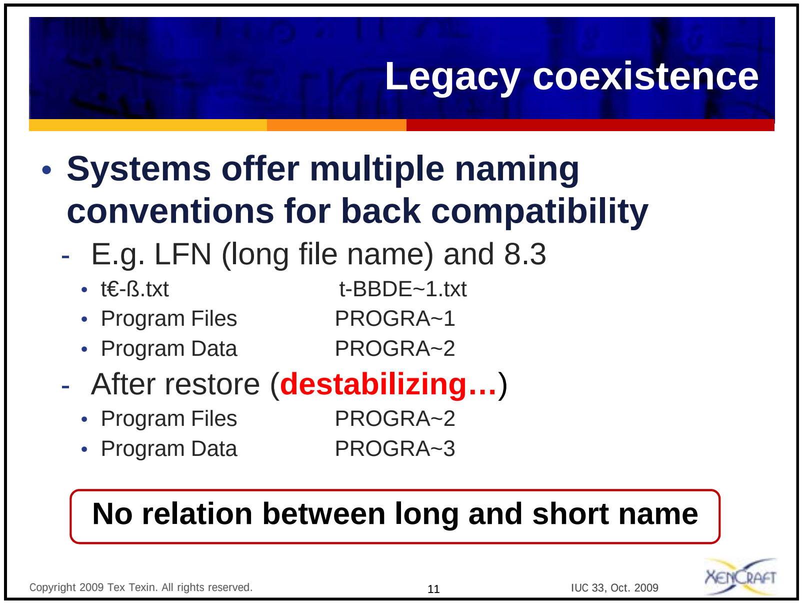
Unicode-oriented filenames

- **Valid names on Unicode-based systems**
  - Any UCS-2
  - Any UTF-16
  - Any Unicode except **NUL /**
    - Additional exceptions for **U+FFFF, U+FFFE**
  - Unicode version 3.2
  - Unicode version 3.2, minus long list of restricted characters
  - All UCS-2 except *** / \ ; : ?**
    - What about full-width *** / \ ; : ?**   ?



More variations

- **Case-sensitivity**
  - (based on locale, some are ASCII-only)
- **Case-preservation**
- **Normalization (none, NFC, NFD)**
- **Search and wildcard behavior**
  - CIFS ? = 1 character, except at the end of a filename where it is 0 or 1 characters.
    - "??x" equals "abx" but not "abcx" or "ax".
    - "x??" matches "xab", "xa" and "x", but not "xabc"



Legacy coexistence

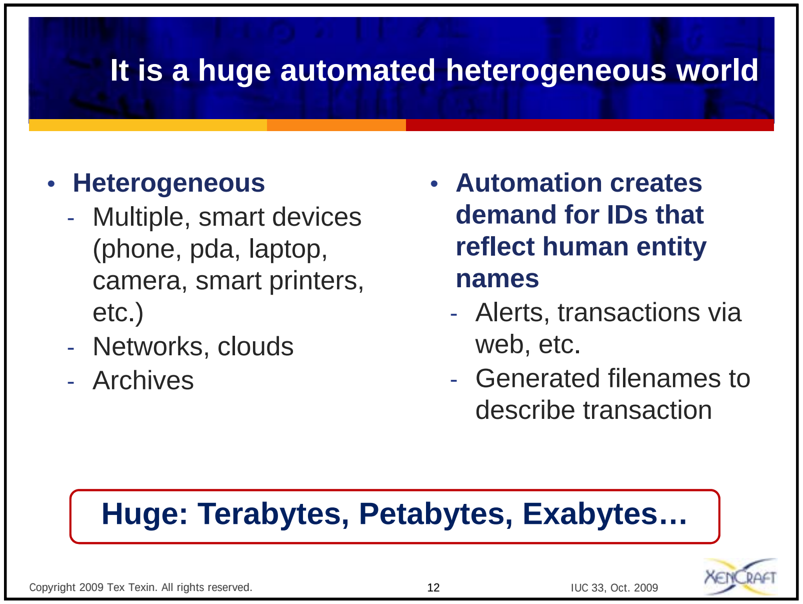
- **Systems offer multiple naming conventions for back compatibility**
  - E.g. LFN (long file name) and 8.3

t€-ß.txt	t-BBDE~1.txt
Program Files	PROGRA~1
Program Data	PROGRA~2

  - After restore (**destabilizing…**)

Program Files	PROGRA~2
Program Data	PROGRA~3

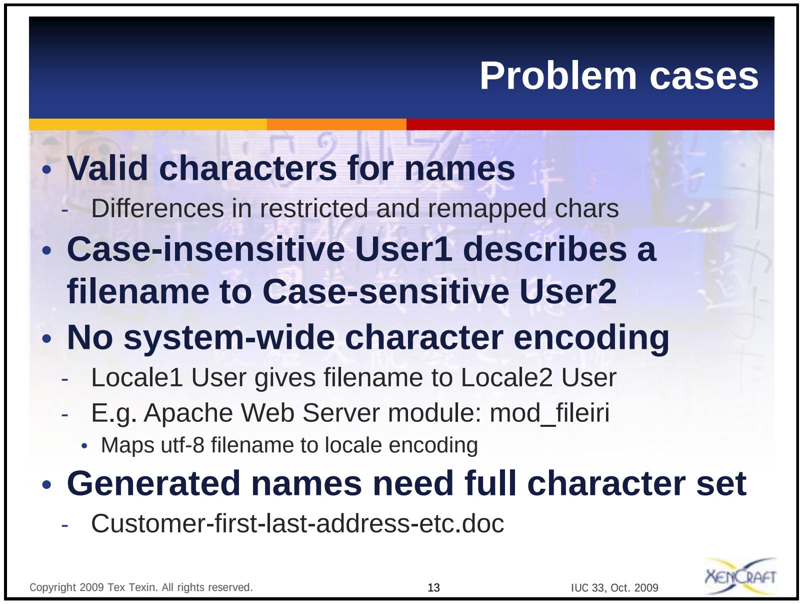
No relation between long and short name



It is a huge automated heterogeneous world

- **Heterogeneous**
  - Multiple, smart devices (phone, pda, laptop, camera, smart printers, etc.)
  - Networks, clouds
  - Archives
- **Automation creates demand for IDs that reflect human entity names**
  - Alerts, transactions via web, etc.
  - Generated filenames to describe transaction

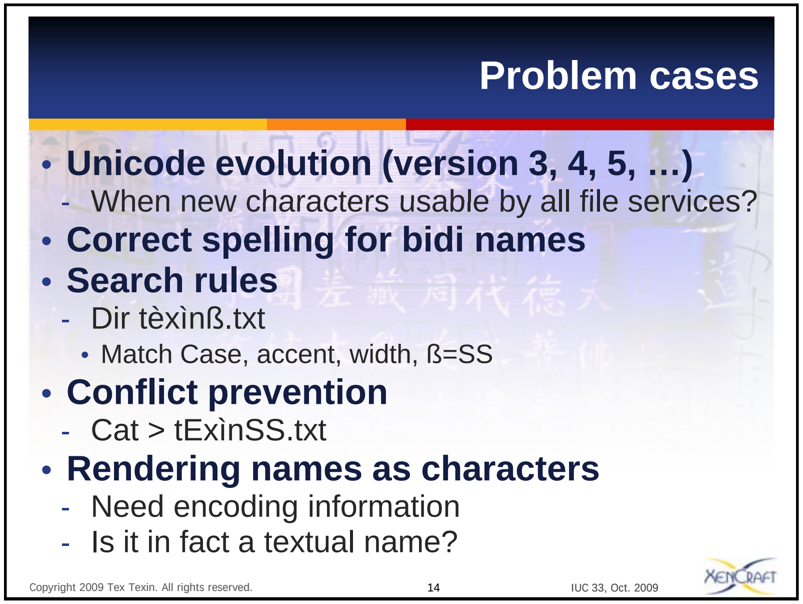
Huge: Terabytes, Petabytes, Exabytes…



Problem cases

- Valid characters for names
  - Differences in restricted and remapped chars
- Case-insensitive User1 describes a filename to Case-sensitive User2
- No system-wide character encoding
  - Locale1 User gives filename to Locale2 User
  - E.g. Apache Web Server module: mod_fileiri
    - Maps utf-8 filename to locale encoding
- Generated names need full character set
  - Customer-first-last-address-etc.doc

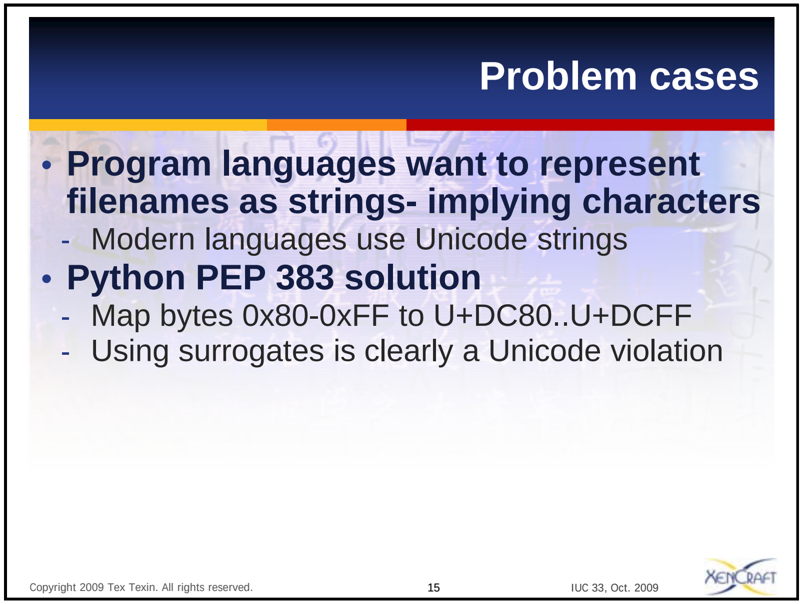
Copyright 2009 Tex Texin. All rights reserved. 13 IUC 33, Oct. 2009



Problem cases

- Unicode evolution (version 3, 4, 5, …)
  - When new characters usable by all file services?
- Correct spelling for bidi names
- Search rules
  - Dir tèxìnß.txt
    - Match Case, accent, width, ß=SS
- Conflict prevention
  - Cat > tExìnSS.txt
- Rendering names as characters
  - Need encoding information
  - Is it in fact a textual name?

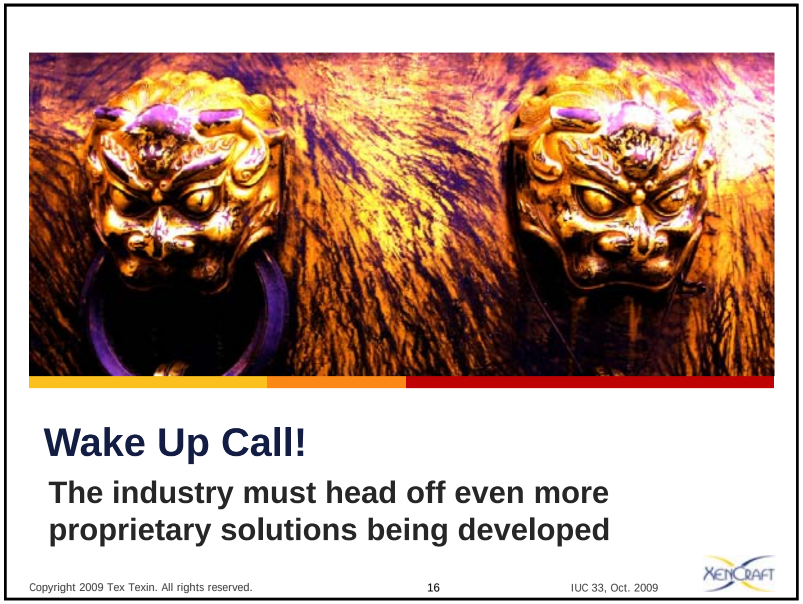
Copyright 2009 Tex Texin. All rights reserved. 14 IUC 33, Oct. 2009



Problem cases

- Program languages want to represent filenames as strings- implying characters
  - Modern languages use Unicode strings
- Python PEP 383 solution
  - Map bytes 0x80-0xFF to U+DC80..U+DCFF
  - Using surrogates is clearly a Unicode violation

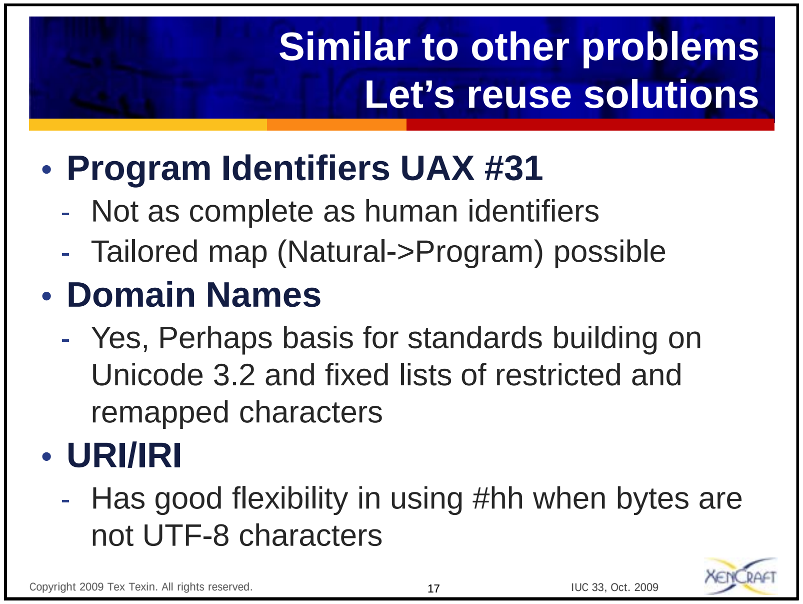
Copyright 2009 Tex Texin. All rights reserved. 15 IUC 33, Oct. 2009



Wake Up Call!

The industry must head off even more proprietary solutions being developed

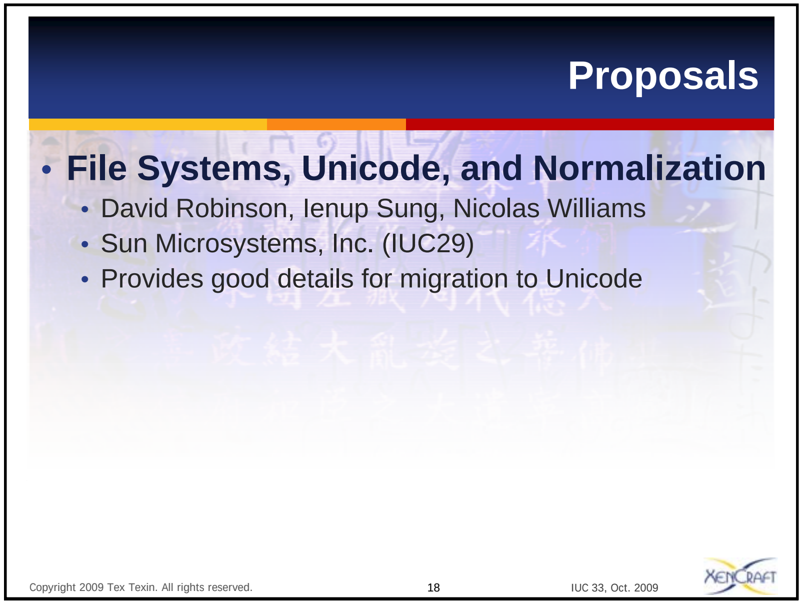
Copyright 2009 Tex Texin. All rights reserved. 16 IUC 33, Oct. 2009



Similar to other problems
Let’s reuse solutions

- **Program Identifiers UAX #31**
  - Not as complete as human identifiers
  - Tailored map (Natural->Program) possible
- **Domain Names**
  - Yes, Perhaps basis for standards building on Unicode 3.2 and fixed lists of restricted and remapped characters
- **URI/IRI**
  - Has good flexibility in using #hh when bytes are not UTF-8 characters

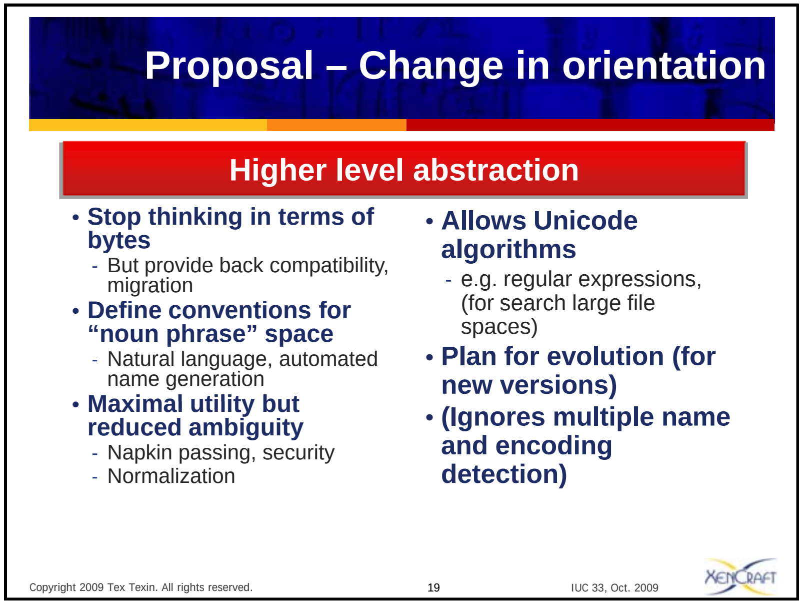
Copyright 2009 Tex Texin. All rights reserved. 17 IUC 33, Oct. 2009



Proposals

- **File Systems, Unicode, and Normalization**
  - David Robinson, Ienup Sung, Nicolas Williams
  - Sun Microsystems, Inc. (IUC29)
  - Provides good details for migration to Unicode

Copyright 2009 Tex Texin. All rights reserved. 18 IUC 33, Oct. 2009

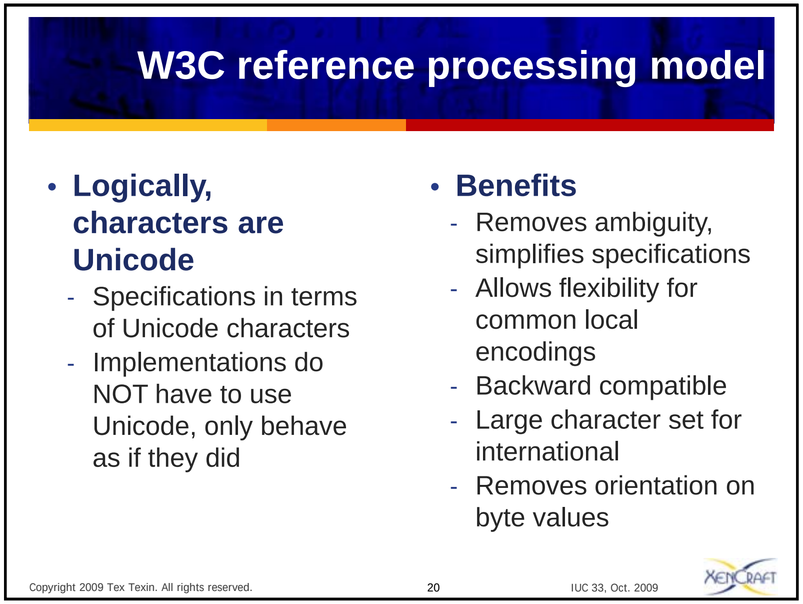


Proposal – Change in orientation

Higher level abstraction

- **Stop thinking in terms of bytes**
  - But provide back compatibility, migration
- **Define conventions for “noun phrase” space**
  - Natural language, automated name generation
- **Maximal utility but reduced ambiguity**
  - Napkin passing, security
  - Normalization
- **Allows Unicode algorithms**
  - e.g. regular expressions, (for search large file spaces)
- **Plan for evolution (for new versions)**
- **(Ignores multiple name and encoding detection)**

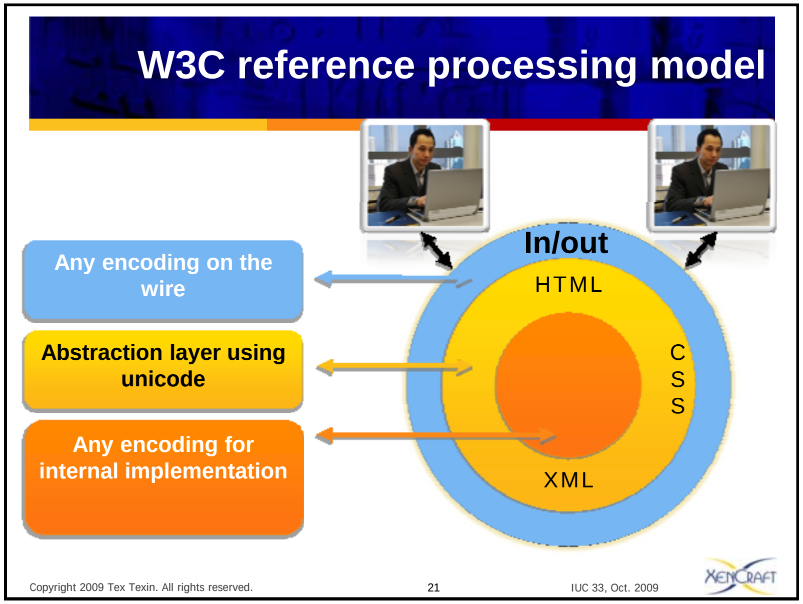
Copyright 2009 Tex Texin. All rights reserved. 19 IUC 33, Oct. 2009



W3C reference processing model

- **Logically, characters are Unicode**
  - Specifications in terms of Unicode characters
  - Implementations do NOT have to use Unicode, only behave as if they did
- **Benefits**
  - Removes ambiguity, simplifies specifications
  - Allows flexibility for common local encodings
  - Backward compatible
  - Large character set for international
  - Removes orientation on byte values

Copyright 2009 Tex Texin. All rights reserved. 20 IUC 33, Oct. 2009



W3C reference processing model

Any encoding on the wire

Abstraction layer using unicode

Any encoding for internal implementation

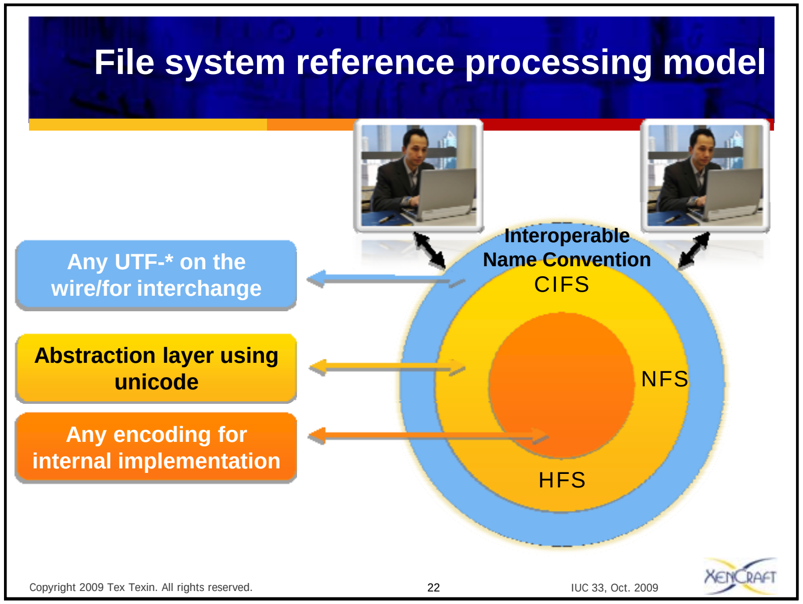
In/out

HTML

CSS

XML

Copyright 2009 Tex Texin. All rights reserved. 21 IUC 33, Oct. 2009



File system reference processing model

Any UTF-* on the wire/for interchange

Abstraction layer using unicode

Any encoding for internal implementation

Interoperable Name Convention

CIFS

NFS

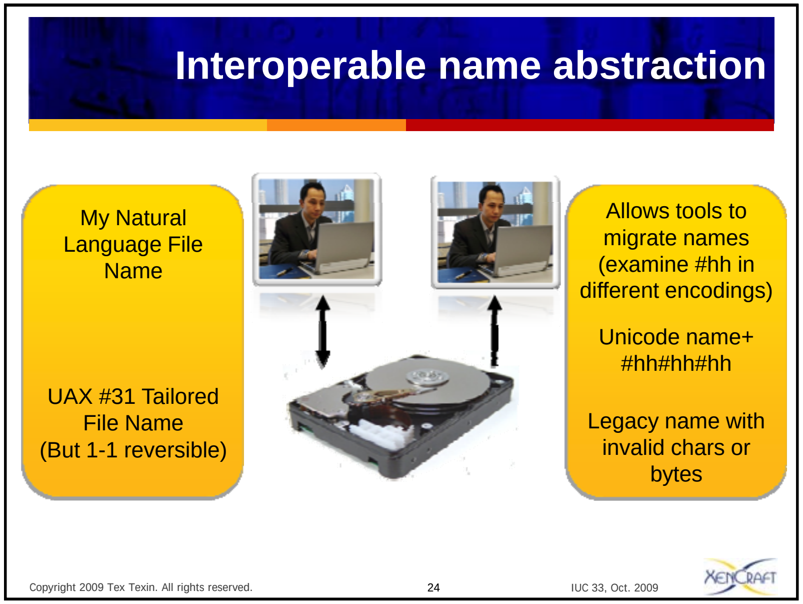
HFS

Copyright 2009 Tex Texin. All rights reserved. 22 IUC 33, Oct. 2009

Interoperable name abstraction

- **Recommend a Unicode (sub)set**
  - Pick a version (5.1)
  - Determine allowed, disallowed characters
    - Provide a property or a publicly available table
    - Tailoring UAX #31 program identifiers
  - Recommend bidi spelling, etc.
- **Disallowed characters can be accepted but then names are not interoperable**
  - Systems can provide warnings
- **Escapes, ala #hh, for invalid names**

Copyright 2009 Tex Texin. All rights reserved. 23 IUC 33, Oct. 2009



Interoperable name abstraction

My Natural Language File Name

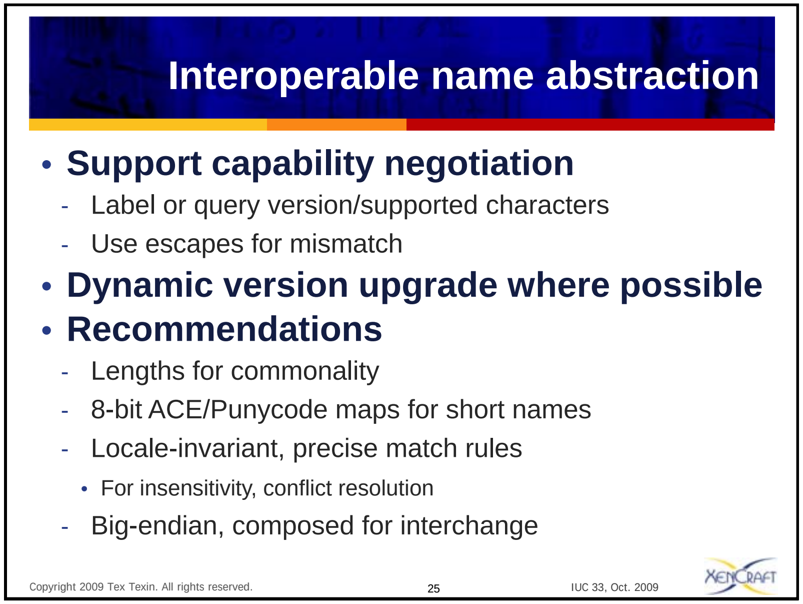
UAX #31 Tailored File Name (But 1-1 reversible)

Allows tools to migrate names (examine #hh in different encodings)

Unicode name+ #hh#hh#hh

Legacy name with invalid chars or bytes

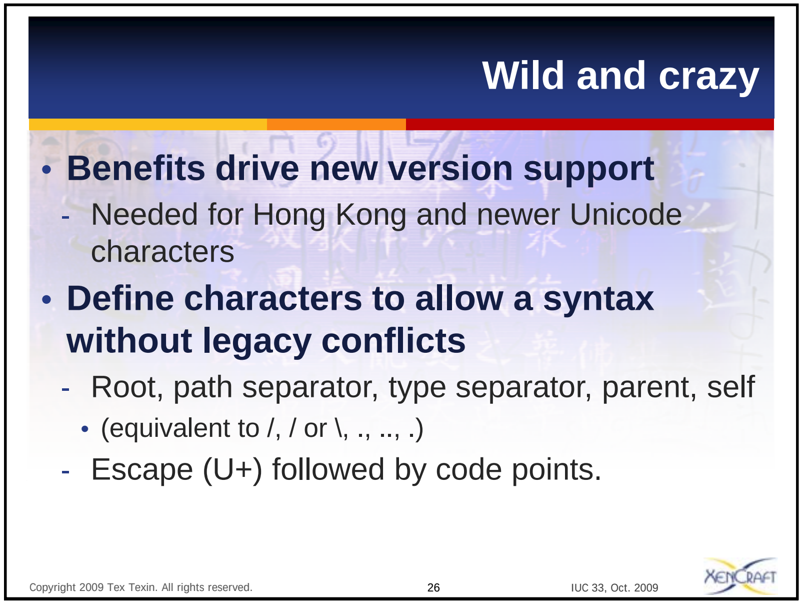
Copyright 2009 Tex Texin. All rights reserved. 24 IUC 33, Oct. 2009



Interoperable name abstraction

- **Support capability negotiation**
  - Label or query version/supported characters
  - Use escapes for mismatch
- **Dynamic version upgrade where possible**
- **Recommendations**
  - Lengths for commonality
  - 8-bit ACE/Punycode maps for short names
  - Locale-invariant, precise match rules
    - For insensitivity, conflict resolution
  - Big-endian, composed for interchange

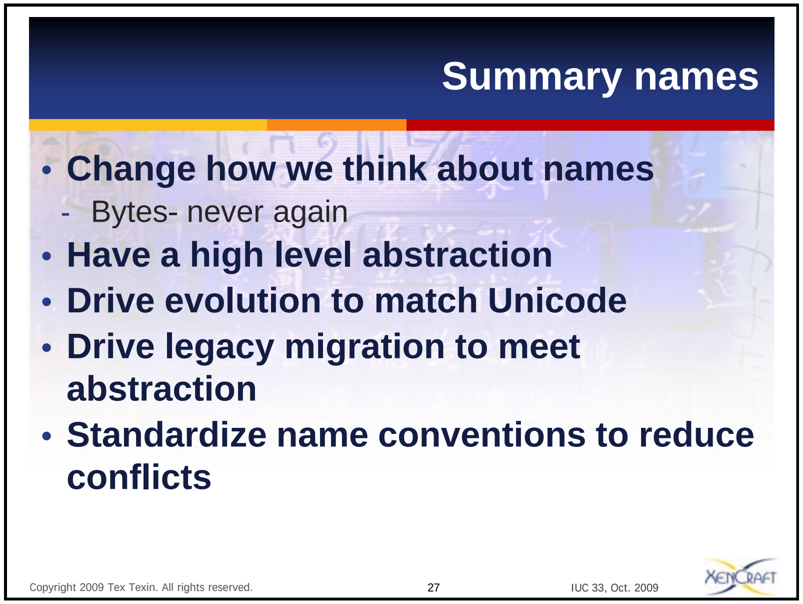
25 IUC 33, Oct. 2009



Wild and crazy

- **Benefits drive new version support**
  - Needed for Hong Kong and newer Unicode characters
- **Define characters to allow a syntax without legacy conflicts**
  - Root, path separator, type separator, parent, self
    - (equivalent to /, / or \, ., .., .)
  - Escape (U+) followed by code points.

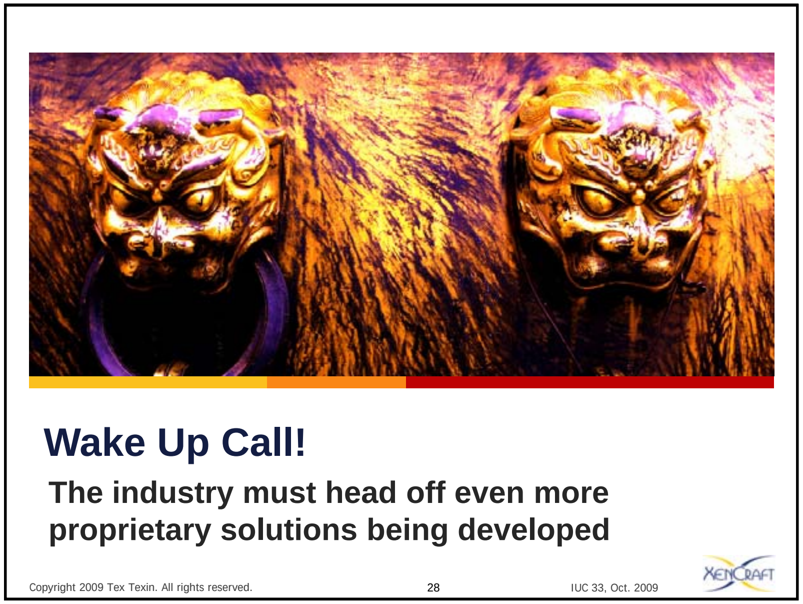
26 IUC 33, Oct. 2009



Summary names

- **Change how we think about names**
  - Bytes- never again
- **Have a high level abstraction**
- **Drive evolution to match Unicode**
- **Drive legacy migration to meet abstraction**
- **Standardize name conventions to reduce conflicts**

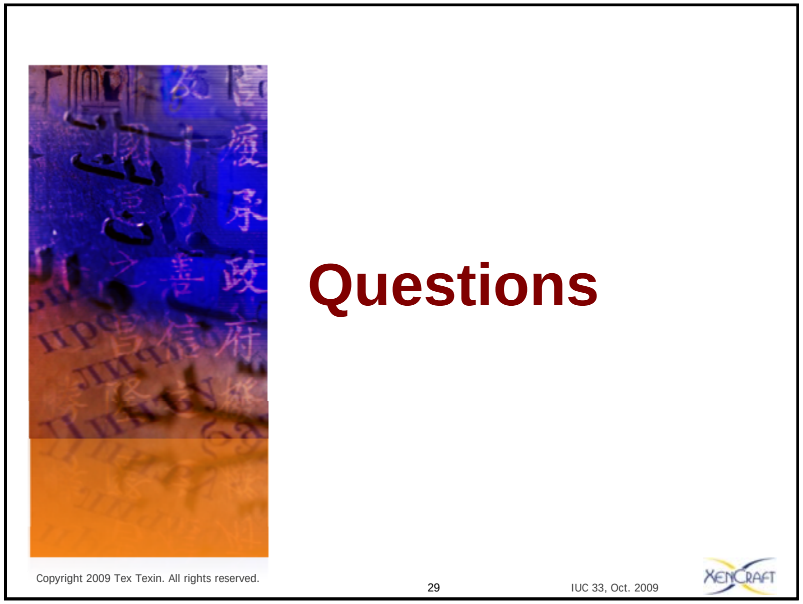
27 IUC 33, Oct. 2009



Wake Up Call!

The industry must head off even more proprietary solutions being developed

28 IUC 33, Oct. 2009



Questions