Presentation Goals

- Identify the potential public and staff impact of moving to Unicode™
- Suggest ways to search for diacritics and special characters in collections
- Present diacritic and special character cleanup strategies
Presentation Pathway

- Preparing for Unicode™
- Identifying Potential Records with Diacritics
- After Conversion to Voyager® with Unicode™: Cleanup Strategies
- Preview a Voyager® with Unicode™ Implementation
- Wrap Up

Obligatory Disclaimer & Caveat

- The record errors and display problems we describe are due to the peculiar history of our data, upgrades, and migrations
- Most of the problems we experienced have been resolved during Beta and Early Release testing
About Pacific Lutheran University

- Comprehensive Private University
  - Liberal Arts & Professional Schools
  - 3,400 Students
- Languages Taught:
  - American Sign Language, Chinese, Classics, French, German, Norwegian/Scandinavian Studies, Spanish
- Wang Center for International Studies
- Our Library Collection
  - 274,000 BIBs currently
  - 104 languages
- Voyager® Customer: almost 3 years
- Voyager® Clients: 28

What is Unicode™?

- “Unicode™ provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language”
- 40,000+ characters for all major scripts and technical symbols
- See Support Web, “Road to Unicode™”
- Impact of Upgrade: Display of diacritics and special characters (and any encoding errors) in browsers and clients
Beta Testing Voyager® with Unicode™

Motivation
- University initiative for internationalization
- Database cleanup & quality control
- On Voyager® 2000.1.3; skip intermediate upgrades
- Gain new functionality
- Bulk Imports and index regeneration
- Ability to display diacritics and special characters in cataloging and WebVoyáge

Beta Testing Voyager® with Unicode™
- Became Beta Partner
- PLU database conversion (August 20, 2003) - 1 day
  - SupportWeb estimates:
    - < 300,000 bibs = 1 day
    - > 300,000 bibs but < 1,000,000 bibs = 2 days
    - > 1,000,000 bibs = 3 days
- Test mission-critical functions; then day-to-day workflow
- Report problems; track resolutions; communicate with staff
- Review conversion log; cleanup data
- Upgrade clients/server 4 times + server patches
- Currently Early Release Partner
What Happens During Unicode™ Conversion?

- Oracle 8.05 upgraded to 9i
- MARC records (BIB, AUTH, & MFHD) converted from Voyager® Encoding (VRLIN) to UTF-8 (UCS Transformation Format-8), which is the MARC standard
- New clients & ODBC drivers
- New functionality for importing, editing, and displaying diacritics
- Mapping in cataloging—cataloging preferences depend upon where you get your data
- WebVoyage and clients display diacritics and special characters

Communicating & Managing Expectations

- **Public**
  - Downtime
  - Browsers & settings

- **Staff**
  - Downtime
  - Public and staff workstations
    - OS requirements
    - Browser requirements
  - Reporting & resolving issues
  - Cleanup process
  - Cleanup pace

How painful, disruptive, [you supply word(s)] will it be?

What will it mean to me?
Browser/Client Configurations

- Install newer browsers/OS
  - See SupportWeb page details
- Set font to Arial Unicode MS
- Semi-graceful degradation for earlier browsers
  - Potential Display Problems:
    - Boxes rather than diacritics
    - Punctuation: HTML entities
    - Line breaks in some fields may be lost
    - Access to patron record

<table>
<thead>
<tr>
<th>Browser/OS</th>
<th>Voyager Release 201.1</th>
<th>Voyager Release 201.2</th>
<th>Voyager Release Unicode</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet Explorer</td>
<td>5.6.0</td>
<td>5.6.0</td>
<td>Supported</td>
</tr>
<tr>
<td>Netscape</td>
<td>4.7.6.2</td>
<td>4.7.6.2</td>
<td>6.2</td>
</tr>
</tbody>
</table>

PC Operating System (CLIENT):

<table>
<thead>
<tr>
<th></th>
<th>Supported</th>
<th>Not Supported</th>
<th>Not Supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>Windows 98 Standard Edition</td>
<td>✔️</td>
<td>✔️</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Windows 98 Professional</td>
<td>✔️</td>
<td>✔️</td>
<td>Not Supported</td>
</tr>
<tr>
<td>Windows 2000 Professional</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
<tr>
<td>Windows XP</td>
<td>✔️</td>
<td>✔️</td>
<td>✔️</td>
</tr>
</tbody>
</table>

http://support.endinfosys.com/cust/voy/cust_supp/tech/clientplatform.html
Captured 4-13-2004

Client Configurations

- Arial Unicode MS Font
- Set cataloging session preferences for uploading records
- ODBC Access Reports
  - Install Oracle 9 drivers
  - Relink tables
  - Import custom queries
  - Adjust font
  - Implement UTF8to16 conversion function in queries
WebVoyáge Display

Cataloging Client Display
**Unicode™ Transition May…**

- display skeletons in database
- affect searching
- require problem identification & cleanup strategies

**The Skeletons**

- If your diacritics or special characters were not correctly coded and converted, in WebVoyage you now see…
  - character error icons
And in the Cataloging Client You See...

- More errors
- New tools

Effects on Searching

- Searching for terms with encoding errors may fail to retrieve expected results
- e.g. keyword anywhere search for “espanola”
Assessing the Impact

- How many records might have diacritics or special characters?
- How many records actually contain diacritics or special characters?
- In what tags are they most likely to lurk?
- What percentage of the collection is involved?

Estimating Foreign Language Record Numbers

- Count BIB_TEXT.LANGUAGE
- <> “eng” or <> “”

SELECT BIB_TEXT.LANGUAGE,
       Count(BIB_TEXT.LANGUAGE) AS CountOfLANGUAGE1
FROM BIB_TEXT
GROUP BY BIB_TEXT.LANGUAGE
HAVING (((BIB_TEXT.LANGUAGE)<>"eng") OR
       (((BIB_TEXT.LANGUAGE)<>""))
ORDER BY Count(BIB_TEXT.LANGUAGE) DESC;
**Language Frequency**

- nor 3,188
- ger 2,766
- fre 2,113
- spa 1,428
- swe 1,111
- dan 675
- lat 547
- ita 346
- chi 189
- grc 112

**Foreign Language Record Estimates**

**Assumptions**

- BIB_TEXT.LANGUAGE is a good field for estimates (not necessarily valid)
- Diacritics & special characters live primarily in non-English coded records (not necessarily valid)
- Not all non-English coded records are likely to contain diacritics & special characters

**Raw Numbers**

- eng: 20,000
- <> eng: 234,170

**Percent of Collection**

- No diacritics?: 7.3
- Possible diacritics: 92.7
EndUser 2004: Session 45, April 16, 2004  
Transitioning To Unicode: Strategies for Tidying Your Data  
Budde, Lane Rasmus, & Nordgren

### Records with Diacritics & Special Characters

**Subsample Records in Query**

- **First 100 with diacritics:** 8%
- **Last 100 with diacritics:** 13%; most in Swedish or Norwegian

**RECORDS WITH DIACRITICS (high) 13%**

13% of 20,000 = ~ 2,600 records  
Percentage = ~ 0.95% of total records

**Another View: Bulk Importing**

- Records Sent to OCLC = 235,518  
- Total Parse Errors on Bulk Import = 2,614  
- Percentage: 1.11% of Bulk Imported records

### Record Cleanup Strategies

- Review Conversion Log Reports
- Develop Access Reports
- Consider Record Replacements
- Make Global Headings Changes
- Utilize Bulk Import and WebAdmin
Conversion Logs

- In our logs:
  - BIB Log (267,000)
    - 866 undefined character errors
    - 3,432 loose character warnings
  - AUTH Log (420,000)
    - 1 undefined character error
    - 5 loose character warnings
  - MFHD Log (268,000)
    - 16 undefined character errors
    - 400 loose character warnings

Working with the Conversion Log

- Helpful to import log into a spreadsheet
- Sort by type of entry
  - undefined character errors
  - loose character warnings
- Once you can visually see the data, patterns start to emerge
Undefined Character Errors

- Out of 866 errors, we discovered
  - 750 in 650 tag (86%) “—” used instead of delimiter (‡)
  - 70 in 590 tag
  - 25 in 440 tag
  - 21 in all other tags combined
- Nearly all errors were from non-MARC records input in our previous ILS

Undefined Character Errors

- The non-MARC records could be grouped into the following types:
  - 400 videos
  - 321 special collections
  - 92 media types (cassettes, CDs, slides)
  - 53 miscellaneous items
- Special cleanup projects were already planned for the non-MARC records
- *Ultimately, there were no diacritic or special character errors in our converted data!*
Loose Character Warnings

• What is a loose character?
  – Examples are carriage returns, line feeds, backspaces, and MARC-8 superscript and subscript numbers
• All but a handful of our 3,432 loose characters were in non-MARC records
• Primary problem caused by improper Backspace key emulation settings on terminals in previous ILS

What the Conversion Log Missed

• Improper input of ASCII characters as diacritics
  – ~ (tilde)
  – ` (grave)
  – ^ (circumflex)
• By happenstance, after conversion we used Bulk Import on files not converted to Unicode™
  – We noticed the err.imp files listed these as ASCII characters as undefined characters or parse errors
  – These 2,614 errors were not among the 866 listed on the BIB conversion log
• Now we knew where the real cleanup focus needed to be in the PLU database
Finding ASCII Characters Used as Diacritics

- Characters
  - ~ (tilde)
  - ` (grave)
  - ^ (circumflex)
- Used Access/SQL to identify the characters
- First tried the BIBBlob to search the entire bib record, but it failed

**SQL:** Opted to search field-by-field for these characters. Example below is for the 100 tag

```
SELECT BIB_INDEX.INDEX_CODE, BIB_INDEX.BIB_ID, BIB_INDEX.DISPLAY_HEADING
FROM BIB_INDEX
WHERE (((BIB_INDEX.INDEX_CODE) Like "*100*")
AND ((BIB_INDEX.DISPLAY_HEADING) Like "*~*" Or
(BIB_INDEX.DISPLAY_HEADING) Like "*^*" Or
(BIB_INDEX.DISPLAY_HEADING) Like "*\*")));
```
**Record Replacement**

- When should you replace a record instead of correcting the diacritics and special characters?
  - When you do not have in-house language expertise
    - For PLU, we will replace our records containing Chinese still in Wade-Giles to Pin Yin
  - Staff Time & Cost
    - When a record has multiple errors, it may save time
    - OCLC search and export charges ~ $1.00 per record
  - Workflow Issues

---

**Global Headings Change (GHC)**

- GHC is a way to update diacritics in all name, title, name/title, or subject fields in bibliographic records by changing them in the authority record
- When a change is made to an authority record the Global Change Queue (in cataloging) displays the change as well as the associated records
Bulk Import and WebAdmin

- 035 cleanup happenstance
- Bulk export all bib and authority records before conversion
- After conversion “pseudo” bulk import to generate err log
- Generates parse error report without replacing records (does nothing to the database)
- Will not work with post-conversion data

Bulk Import Parse Error Report

Pbulkimport -fFILENAME -iPROFILE -oUID -b99999 -e99999

log.imp file

BIB ID

err.imp file
**Tools & Features for Diacritics**

- ALA Character Entry Table in OCLC is helpful to determine which diacritics and special characters go with which language
- Voyager’s® Special Character Entry lists diacritics and special characters by name
- PLU’s website with a listing of the most frequently found diacritics and special characters in the PLU Library catalog as represented in Voyager and OCLC: [http://www.plu.edu/~libr/EndUser2004/diacritics.html](http://www.plu.edu/~libr/EndUser2004/diacritics.html)
- ![Character error icon to indicate incorrect diacritic](http://library.plu.edu)
- UTF8toUTF16 function in Access reports (allows you to see the diacritics in Access reports)

**PLU Library Catalog**

[http://library.plu.edu](http://library.plu.edu)
Wrap Up

• At last we can view diacritics and special characters properly
• Converting to Unicode™ is not a difficult transition though you may need to do some cleanup
• Strategies exist to locate errors
• Developing, refining, and sharing cleanup strategies would be beneficial to others

Contact Info

Fran Budde
buddefc@plu.edu
253-535-7511

Francesca Lane Rasmus
lanerafr@plu.edu
253-535-7141

Layne Nordgren
nordgrle@plu.edu
253-535-7197
SESSION 45: Transitioning to Unicode: Strategies for Tidying Your Data
April 16, 2004 2:45 pm
Fran Budde, Francesca Lane Rasmus, Layne Nordgren
Pacific Lutheran University

Web Resources

Unicode4.0.1
http://www.unicode.org/versions/Unicode4.0.1/

What Is Unicode?

Unicode Code Charts
http://www.unicode.org/charts/

Support Web “The Road to Voyager with Unicode™”
http://support.endinfosys.com/cust/voy/upgrade/unicode/index.html

Support Web “Supported Browsers, PC Operating Systems & Recommended PC Hardware Requirements”
http://support.endinfosys.com/cust/voy/cust_supp/tech/clientplatform.html

PLU Library Catalog
http://library.plu.edu

PLU Library’s Coding Diacritics in Voyager website
http://www.plu.edu/~libr/EndUser2004/diacritics.html

Princeton’s Diacritics by Language website
http://infoshare1.princeton.edu/katmandu/catcopy/diatoc.html

SQL Queries in Access

Estimating Foreign Language Record Numbers

```sql
SELECT BIB_TEXT.LANGUAGE, Count(BIB_TEXT.LANGUAGE) AS CountOfLANGUAGE1
FROM BIB_TEXT
GROUP BY BIB_TEXT.LANGUAGE
HAVING (((BIB_TEXT.LANGUAGE)<>"eng") OR (((BIB_TEXT.LANGUAGE)<>""))
ORDER BY Count(BIB_TEXT.LANGUAGE) DESC;
```

Finding ASCII Characters Used as Diacritics

```sql
SELECT BIB_INDEX.INDEX_CODE, BIB_INDEX.BIB_ID, BIB_INDEX.DISPLAY_HEADING
FROM BIB_INDEX
WHERE (((BIB_INDEX.INDEX_CODE) Like "*100*") AND ((BIB_INDEX.DISPLAY_HEADING) Like "*~*") Or ((BIB_INDEX.DISPLAY_HEADING) Like "*^*") Or ((BIB_INDEX.DISPLAY_HEADING) Like "*`*"));
```