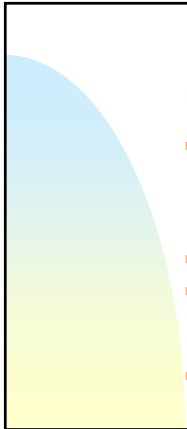


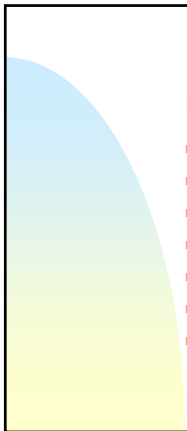
Jump Starting New Projects

David K. Ream
© Leverage Technologies, Inc.
www.LevTechInc.com



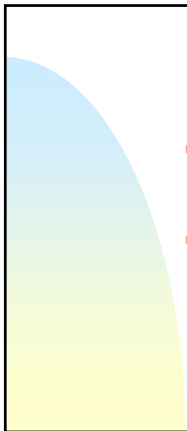
Goals

- Provide background on types of electronic files likely to be received from a client for use with new projects
- How to vet and perform triage
- Introduce the variety of possible scenarios for jump starting a project
- Tools & services for achieving conversions & data extractions



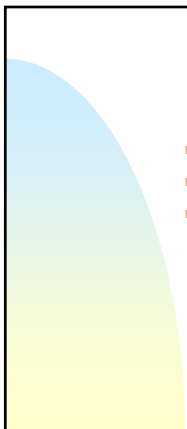
Overview

- Project Types
- File Formats
- Conversions
- Triage/Vetting
- Other Issues Along the Way
- Extraction Projects
- Cost Justification



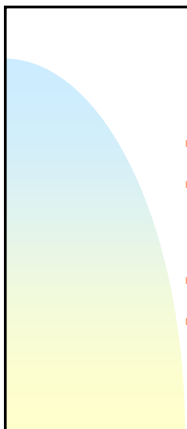
Project Types

- Revision of book (conversion)
 - ◆ Subject index
 - ◆ Tables (cases, citations)
- New project material (extraction)
 - ◆ Tables of contents
 - ◆ Bibliography for names index
 - ◆ Specialty indexes
 - ◆ Electronic data



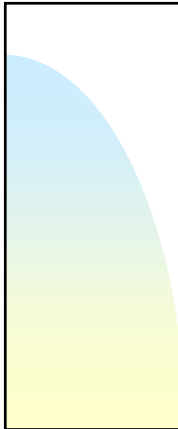
File Formats

- PDF (not a source file)
- Word (doc, docx, rtf)
- Tagged text (often txt)
 - ◆ QuarkXpress (xtg)
 - ◆ InDesign (also xml)
 - ◆ SGML (sgm, sgml), XML (xml)
 - ◆ Composition systems
 - ◆ Spreadsheets & Databases



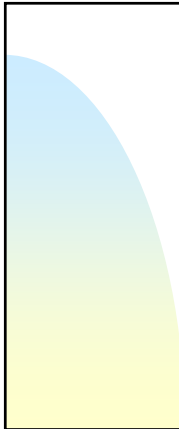
File Formats

- Perform triage or vet client files on receipt
- Problems with opening the file, corruption, the wrong version can be identified sooner rather than the 11th hour when you start to work on the project
- What format can or will the client give you?
- File Properties (example)
 - ◆ Unblock (WinZip, ...)
 - ◆ Read-only (copies CD files)



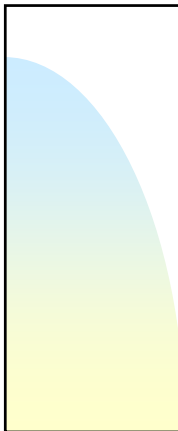
Triage: PDF

- PDFs are not all created equally: scanned or generated
- Try to copy text to determine if it is a scanned document
- Are there crop marks, a filename?
- Lowest level of structural info
- Ask the client for the file that produced the PDF document (not a Word export of the PDF)



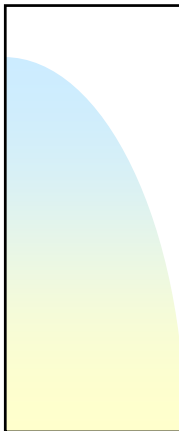
Triage: PDF

- PDFs are not all created equally: scanned or generated
- Try to copy text to determine if it is a scanned document
- Are there crop marks, a filename?
- Lowest level of structural info
- Ask the client for the file that produced the PDF document (not a Word export of the PDF)



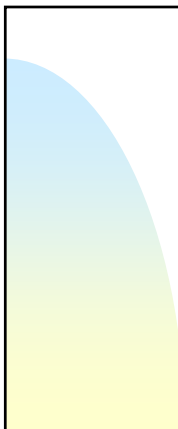
Triage: PDF

- Delete pages that won't be indexed
- Combine multiple files, such as chapters, into one big file (helpful when searching)
- For revisions, use previous and new PDFs and use Compare



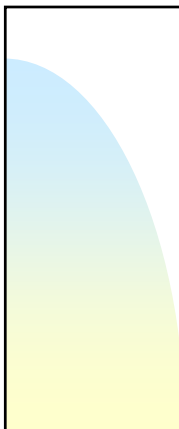
Triage: PDF

- Scanned image document: can be OCR'd but also results in data & structural issues
- Issues
 - ◆ Lost of italics & bold styles
 - ◆ Character problems
 - ◆ 'a' becomes '8', '1' becomes 'l'
 - ◆ ligatures ('fi', 'ff') become 'f.'
 - ◆ accents 'é' becomes 'e'
 - ◆ Multicolumns are concatenated
 - ◆ No page numbers exported
- Export to text, XML, HTML?



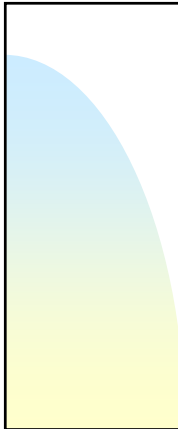
Triage: PDF

- Examples



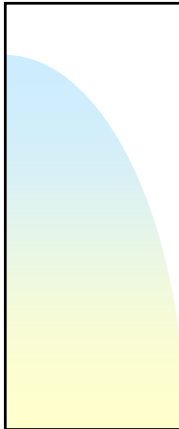
Triage: Word

- More structural context than PDF
- Large amount of variability and inconsistency
- Hierarchy of levels: spaces, tabs, indents, paragraph styles
- Bad line wraps: forced hard breaks
- Italics or bold across paragraphs
- Styles may not be retained
- Undo run-ins & suppressed prefixes



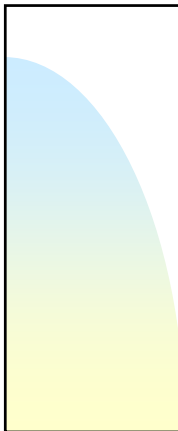
Triage: Word

- Examples



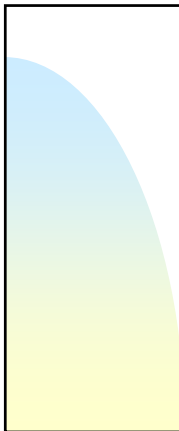
Triage: Tagged

- Source or export from desktop or commercial publishing systems
- Best potential structural and/or content information
- Hardest for non-programmers to utilize
- SGML or XML better than others



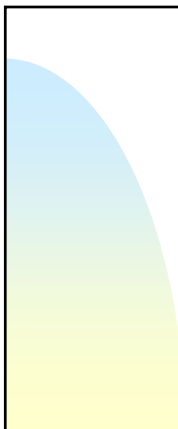
Triage: Tagged Text

- Examples



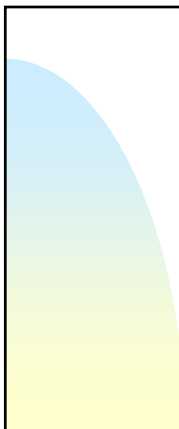
Conversions

- *normalized* file format for import
- flat indexes & tables
 - ◆ global sub a tab
- multilevel headings
 - ◆ no global sub to duplicate upper levels headings
- no global to unsuppress locator prefixes
- tools, macros, services



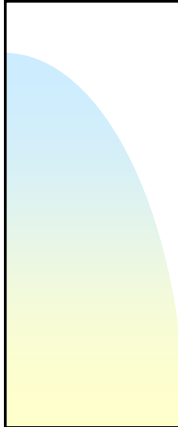
Extraction Projects

- Complex or hard-to-type data
- Large projects
- Large number of chapters/sections
- Frequent/on-going projects



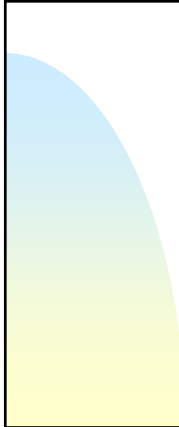
Extraction Projects

- Plants Encyclopedia
 - retrieve italicized genus/species
 - 20,000 entries of hard to type data
- World Handbooks Series
 - names by scanning for capitals;
 - insert page number flags;
 - create names and countries files



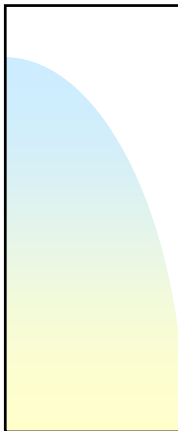
Extraction Projects

- Biblical citations
Word files; extract chapter/verse cites and 2 others
- Names index from bibliography
2400 references → 4700 names
- Electronic/search data
spreadsheet of terms, xrefs, ids;
import/export, cross checking
- Weekly reports in SGML



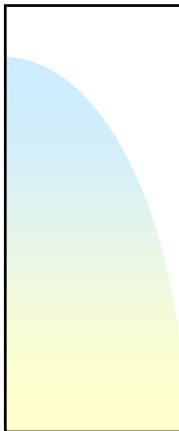
Tools

- IndexDeconstructor
www.editorium.com
- CodeWright (Borland)
Amazon, any software web site;
robust editor with patterns
- EntryExpander (LevTech)
for names indexes
- Conversions & custom software
LevTech & others
- The Mac question



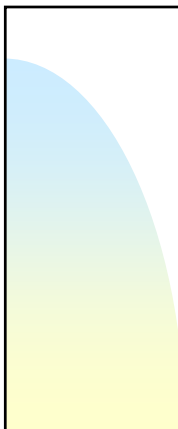
Costs

- Tools vary from \$35 to \$300
- Services vary
 - ◆ Simple conversions
 - ✦ \$45/file; \$.01/entry created
 - ◆ Complex conversion are quoted;
programming costs are one-time
 - ◆ Editing inconsistencies may be required;
vetting saves this cost element
 - ◆ Editing the output to "reformat" to look as
an index can be cheaper than retyping



Cost Justification

- Size of project (number of pages)
- % of reusable entries (revisions)
- Easier to delete unwanted entries
- Difficult manual entry causing loss
of time & accuracy
- Same file formats/data from client
for different projects
- Large clients may cover extra
costs for conversions/programming



Summary

- Save time by using available data
so focus is on indexing not
entering data and proofing it
- Slides posted on
www.LevTechInc.com under
Resources & Links
Presentations
- create a new income opportunities