Data Standardization
&
Normalization

INFX 551
Winter 2018

Course Outline

Week 3
(final week in this module)

Data	Data Systems	Policy, Privacy, & Ethics
Types & Roles	Repositories	Policy
Documentation	Preservation	Privacy
Standardization	Cost Models	Ethics

Topics

- Data Quality
- Data Normalization
- File Naming Conventions

Data Quality

Fitness for use …

ISO 8000 & ISO 9000

*"…the degree to which a set of **characteristics** of data fulfills **stated requirements**."*

*Examples of **characteristics** are:*
completeness, validity, accuracy, consistency, availability and timeliness

ISO
International Organization for Standardization

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

- Data Integrity | Validity & Verification | Quality
    - All of these are contextual
- Normalization
    - Literally - making data conform to a normal schema
    - Figuratively - transforming data structures, organizing variables (columns) and attributes (rows), and editing values so that they are consistent, interpretable, and match best practices in a field.

Database Normalization
(structure)

#	Customer	Order	Item	Delivery Address
1	Linda Porch	01366	Cosa-1	520 Alpha St.
2	Elliott Roof	01377	Ding-1	205 Beta Dr.
3	Kevin Chair	01334	Coisa-1	052 Theta Circle.
4	Todd Window	01355	Veshch-2	502 Gamma Avenue.
5	Diane Door	01353	Koto-2	250 Delta Rd.

Customer	Delivery Address
Linda Porch	520 Alpha St.
Elliott Roof	205 Beta Dr.
Kevin Chair	052 Theta Circle.
Todd Window	502 Gamma Avenue.
Diane Door	250 Delta Rd.

Customer	Order
Linda Porch	01366
Elliott Roof	01377
Kevin Chair	01334
Todd Window	01355
Diane Door	01353

Data Normalization
(values)

#	Customer	Order	Item	Delivery Address
1	Linda A. Porch	01366	Cosa 1	520 Alpha St.
2	Elliott Roof	1377	Ding-1	205 Beta Drive

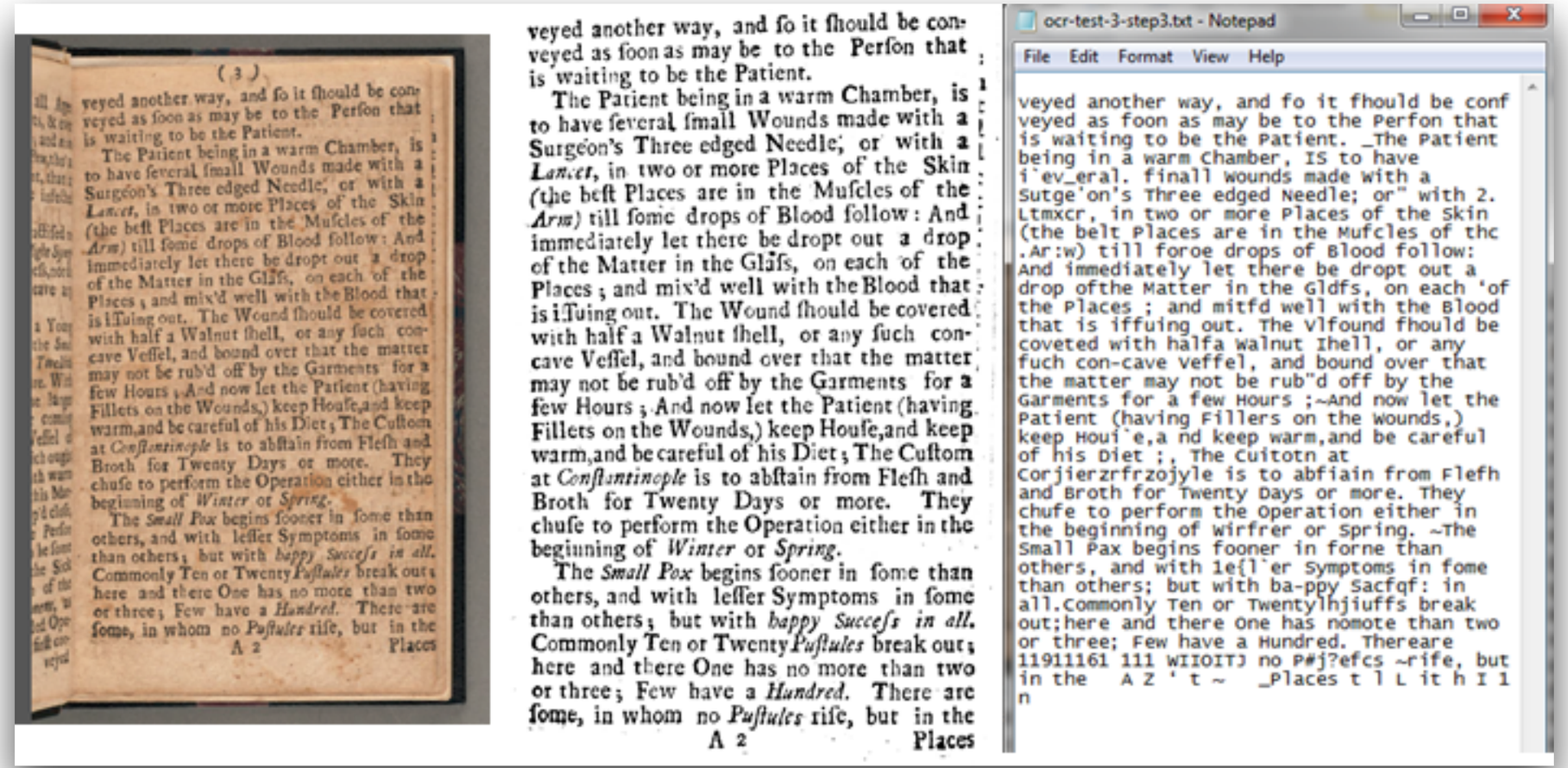
Parsing and standardization — aligning rows and columns, formatting of values into consistent layouts and convention of local standards (for example, postal authority standards for address data). Think of this as the Data Dictionary has to match the Data Value.

Cleaning (aka wrangling / scrubbing)— Modification of data values to meet domain restrictions (e.g. all blank values are to be titled NA), making sure that the **value constraints** of a data dictionary are met in the dataset.

Matching — Identification, linking or merging related entries within or across sets of data (e.g. Item numbers across datasets)

Profiling — Analysis of data to capture statistics that provide insight into the quality of the data and aid in the identification of data quality issues (e.g. We will use a “clustering” feature in Open Refine to group different values together.

Enrichment — Enhancing the value of internally held data. Oftentimes this means adding information (e.g. geographic coordinates; zip codes; etc.)

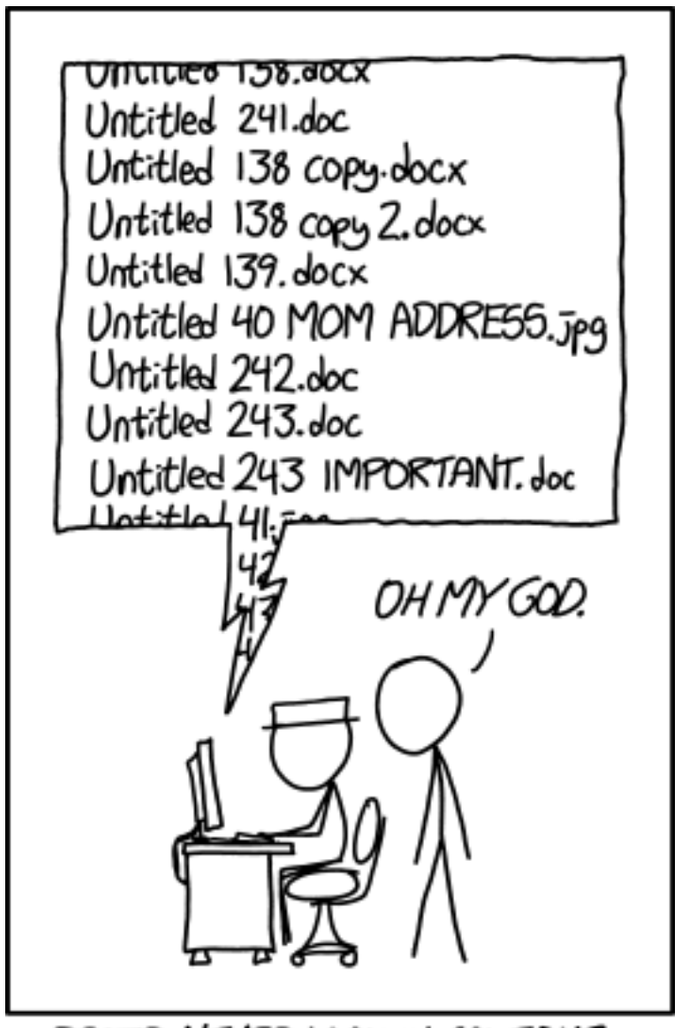


https://smallpoxinoculation.wordpress.com/ocr/

Text Normalization (unstructured data)

- Spelling (e.g. theatre or theater; organise vs organize)
- Vocabulary (e.g
- Punctuation (e.g. on-line vs online)
- Chunking (pargraphs, sentences, stanzas, acts, scenes, etc.)
- Markup (what schema did our XML use to encode a text?)

File Naming Conventions



PROTIP: NEVER LOOK IN SOMEONE ELSE'S DOCUMENTS FOLDER.

Naming Conventions

Every file, every table, every folder should follow a **naming convention** that is human readable, and follows a standard scheme…

Good	Not so good	The worst…
The_Wire_s01e01.mp4	tw0101.mp4	0178.mp4
The_Wire_s01e02.mp4	tw0102.mp4	012709.mp4
The_Wire_s01e03.mp4	tw0103.mp4	01293.mp4
The_Wire_s01e04.mp4	tw0104.mp4	098279.mp4
The_Wire_s01e05.mp4	tw0105.mp4	019283.mp4

Basic Rules

No spaces - use underscore (The_Wire), dash (The-Wire), or capitalization.

Follow a common practice for capitalization:

Camel Case: The first letter of an identifier is lowercase and the first letter of each subsequent concatenated word is capitalized

theWire_s01e01.mp4
runLolaRun_1998.mp4

Pascal Case: The first letter in the identifier and the first letter of each subsequent concatenated word are capitalized.

TheWire_s01e01.mp4
TheLandBeforeTime_1988.mp4

Version number (DOEproject_StaticArray_**V1**.tar)

File State (e.g. MyAwardWinningResearchPaper_**draft1**_01202018.md)

File Creator / Editor (DOEproject_StaticArray_**nmw**.tar)

Some more rules:

- Think about a structure to your file directory: **Start your FNC with the more general components and move to the more specific ones later on**. You should be able to understand the TREE of the folder structure by looking at an individual file (e.g. SPD_ParkingViolations_01202018.CSV)

- **Use meaningful abbreviations**. File names that contain too many characters can be unwieldy and cause problems in transferring files.

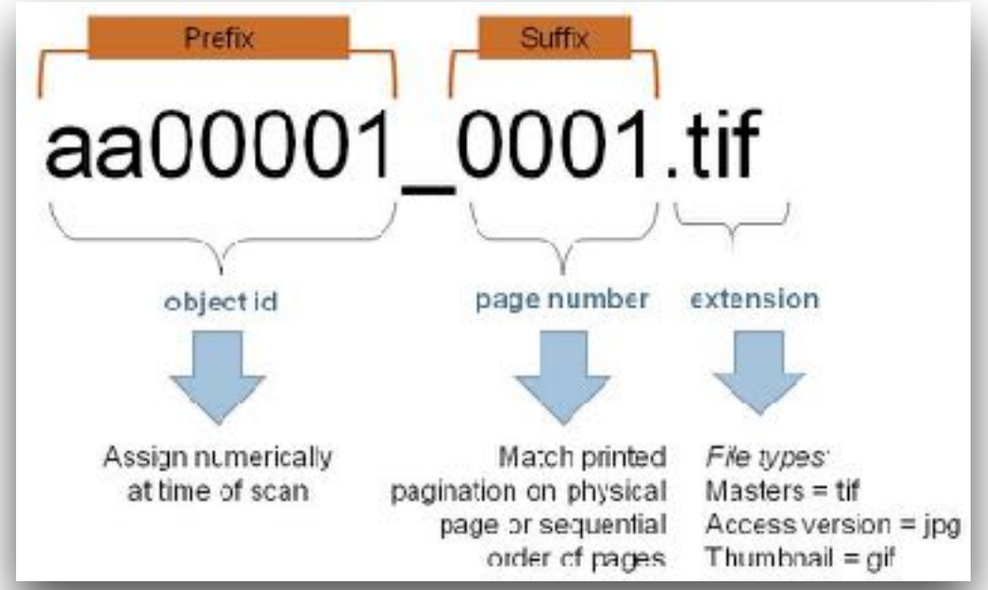
- **Document your decisions** including: what components you will use (the "project name" for example), what are the appropriate entries ("DOEProject"), what acronyms mean (DOE stands for the Department of Energy), etc. Dates should always follow a covention (e.g. yyyy-mm-dd) to organize files chronologically.

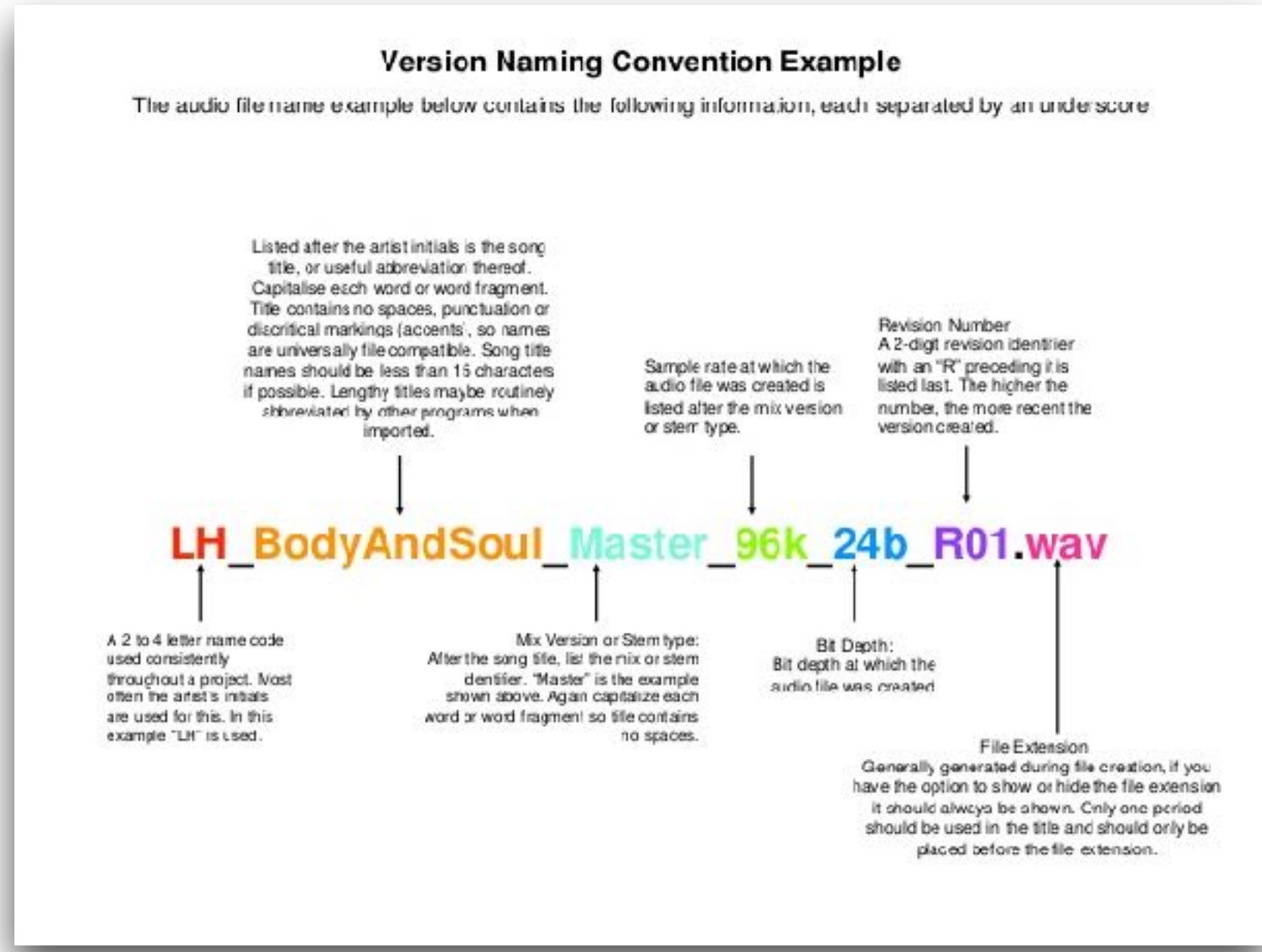
File Naming

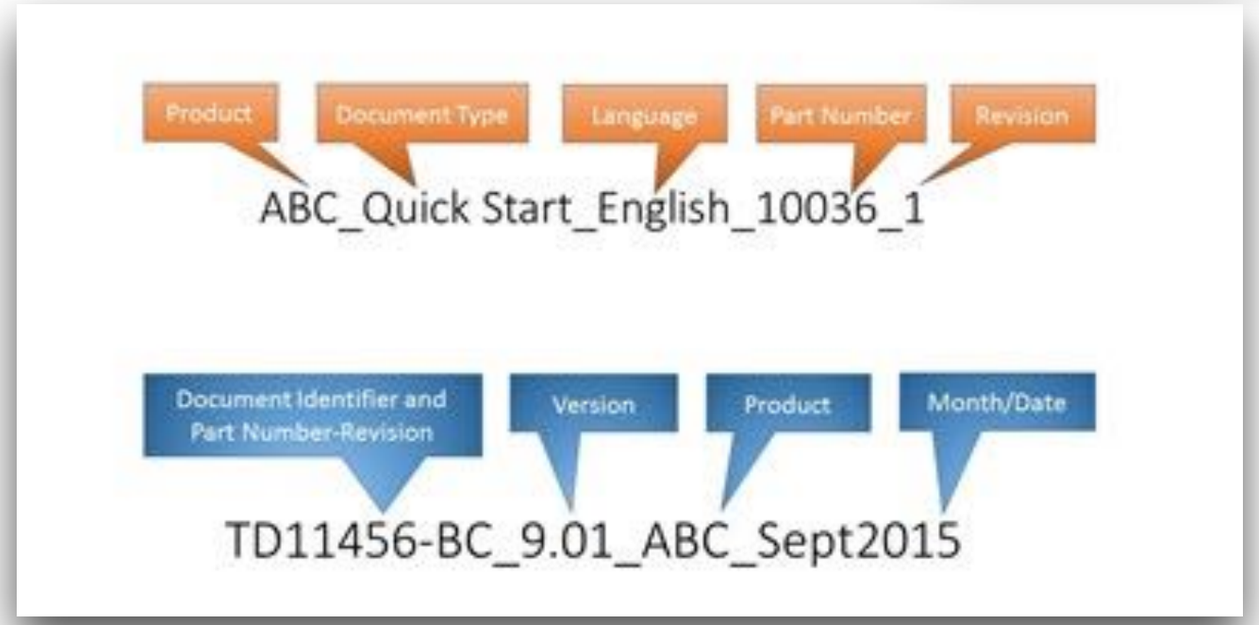
(versioning)

Semantic Versioning

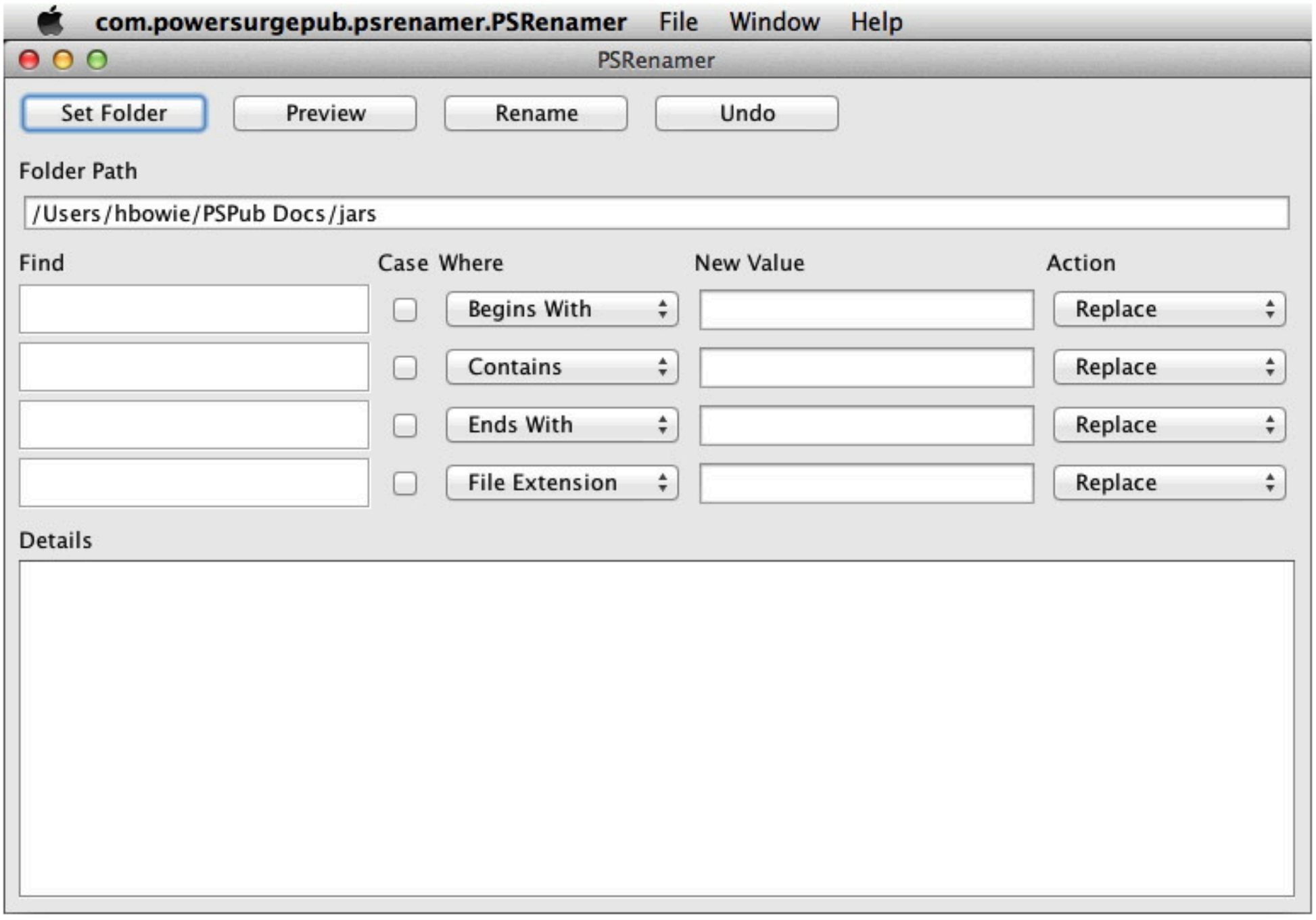
Stage	Rule	Example
Data collected and deposited in storage facility	Start with 1.0.0	Weber-SocCapitalSurvey-**1.0.0**.tab
Data normalized	Increment third digit	Weber-SocCapitalSurvey-**1.0.1**.tab
Data updated to correct errors, or add analysis	Increment second digit	Weber-SocCapitalSurvey-**1.1.0**.tab
New data collected or produced	Increment whole first digit	Weber-SocCapitalSurvey-**2.0.0.**tab


Prefix
Suffix
aa00001_0001.tif
object id
page number
extension
Assign numerically at time of scan
Match printed pagination on physical page or sequential order of pages
File types:
Masters = tif
Access version = jpg
Thumbnail = gif


Version Naming Convention Example
The audio file name example below contains the following information, each separated by an underscore
Listed after the artist initials is the song title, or useful abbreviation thereof. Capitalise each word or word fragment. Title contains no spaces, punctuation or diacritical markings (accents), so names are universally file compatible. Song title names should be less than 15 characters if possible. Lengthy titles may be routinely abbreviated by other programs when imported.
Sample rate at which the audio file was created is listed after the mix version or stem type.
Revision Number
A 2-digit revision identifier with an "R" preceding it is listed last. The higher the number, the more recent the version created.
LH_BodyAndSoul_Master_96k_24b_R01.wav
A 2 to 4 letter name code used consistently throughout a project. Most often the artist's initials are used for this. In this example "LH" is used.
Mix Version or Stem type:
After the song title, list the mix or stem identifier. "Master" is the example shown above. Again capitalize each word or word fragment so file contains no spaces.
Bit Depth:
Bit depth at which the audio file was created
File Extension
Generally generated during file creation, if you have the option to show or hide the file extension it should always be shown. Only one period should be used in the title and should only be placed before the file extension.


Product
Document Type
Language
Part Number
Revision
ABC_Quick Start_English_10036_1
Document Identifier and Part Number-Revision
Version
Product
Month/Date
TD11456-BC_9.01_ABC_Sept2015

PSRenamer



http://www.powersurgepub.com/products/psrenamer/index.html

- Data Quality: …the degree to which a set of characteristics of data fulfills stated requirements.”

- Data Normalization: Applying data quality standards to our data…

- 5 activities: standardization, cleaning, matching, profiling, enrichment

- File Naming Conventions: micro-metadata for files. Document your decisions, and use semantic versioning.