Data Cleaning

J. McCaffrey

iDigBio Biodiversity Informatics Manager

Entomology Digitization Workshop, Chicago, Field Museum

April 24 & 25, 2013

This material is based upon work supported by the National Science Foundation under Cooperative Agreement EF-1115210. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Data Cleaning - - -

It's ugly,
no one wants to do it,
it has to get done,
and
it is never ending.

Definition

- Also referred to as *data cleansing or scrubbing*, the act of detecting and removing and/or correcting a database's dirty data (i.e., data that is incorrect, out-of-date, redundant, incomplete, or formatted incorrectly).
- The goal of data cleaning is not just to clean up the data in a database but also to bring consistency to different sets of data that have been merged from separate databases.
- Sophisticated software applications are available to clean a database's data using algorithms, rules and look-up tables, a task that was once done manually and therefore still subject to human error. [Wikipedia]

Watch words:

- Consistency -> data types
- Elimination of duplication & redundancy -> normalization
- Making a plan, writing standards

More on these as we go.

What follows are guidelines, cautions, advice for cleaning data.

GOAL: making data fit for use

Different Kinds of Cleaning

Occasions for cleaning in order of complexity:

1. **Once in a lifetime** – from one technology to a new one

   You've bought some new technology and your data has to be re-fitted to the new schema. Certain to involve normalization and de-duping.

2. **Episodic** – preparing new data for a new home

   Someone sends you a new dataset to import. Clean it from their standards and map to your schema, and may involve some normalization and de-duping.

3. **Periodic, on-going, forever** – with full awareness of your dataset, you have a list of areas that need sprucing up – dates, localities, taxa, people names, georeferencing.

Different Kinds of Cleaning (1)

Once in a lifetime – going from a one room studio to a 10 room house – data cleaning may involve normalizing and de-duplicating.

- GOOD– no data there yet, so you don’t have to merge, it’s a blank slate.
- BAD - the complexity of the new schema.

Different Kinds of Cleaning (2)

Episodic

- A little like the once in a lifetime example, but on a small scale,
- Involves meshing new data into existing data

Different Kinds of Cleaning (3)

Periodic, on-going, forever

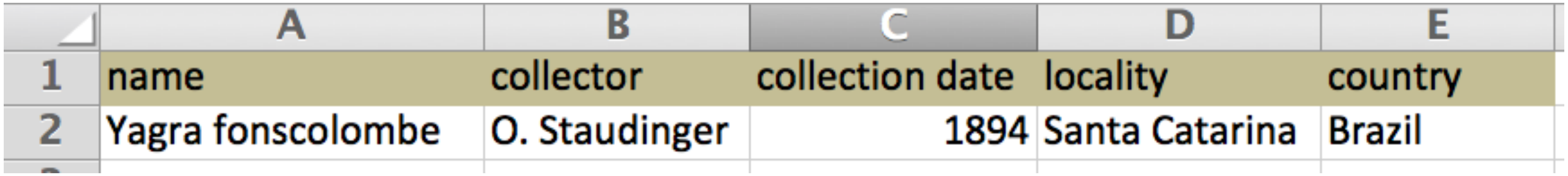
- How did it get like that?
  - got into the DB on import
  - users are entering incorrectly (violating published data standards).

  E.g., people names, taxa improvements, dates
- Data improvement campaigns
  - Georeferencing
    - Use a centroid
  - Taxa improvement (adding higher taxonomy to the record, authorities, dates, biblio).

Different Kinds of Cleaning - examples

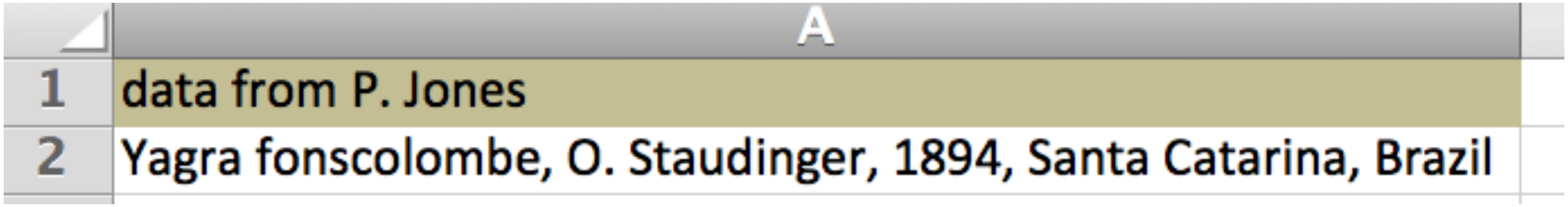
Someone sends you some data to import

- It looks like this



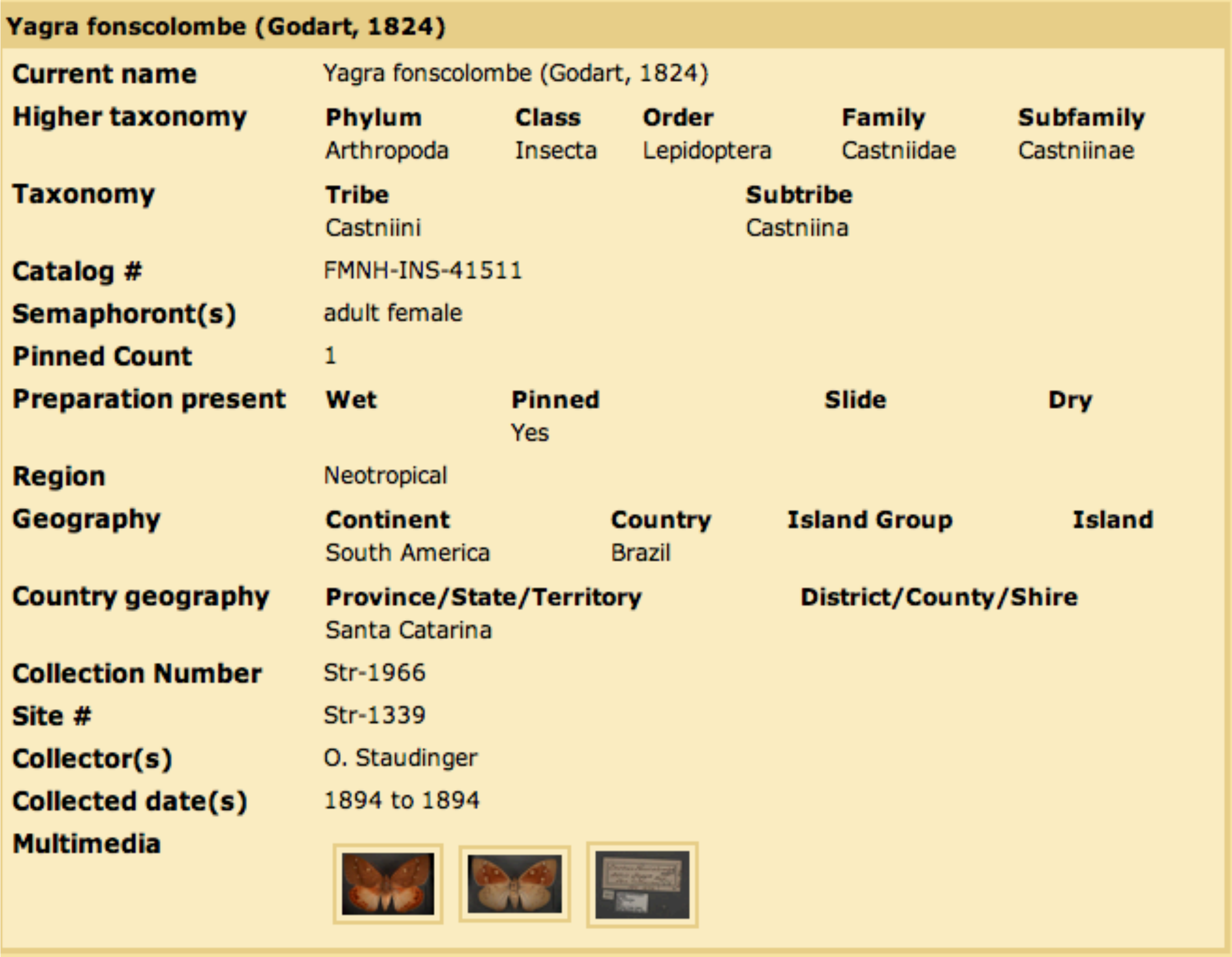
	A	B	C	D	E
1	name	collector	collection date	locality	country
2	Yagra fonscolombe	O. Staudinger	1894	Santa Catarina	Brazil

- Or worse



	A
1	data from P. Jones
2	Yagra fonscolombe, O. Staudinger, 1894, Santa Catarina, Brazil

And this is where you want to put it:



Yagra fonscolombe (Godart, 1824)

Current name	Yagra fonscolombe (Godart, 1824)				
Higher taxonomy	**Phylum**	**Class**	**Order**	**Family**	**Subfamily**
	Arthropoda	Insecta	Lepidoptera	Castniidae	Castniinae
Taxonomy	**Tribe**	**Subtribe**			
	Castniini	Castniina			
Catalog #	FMNH-INS-41511				
Semaphoront(s)	adult female				
Pinned Count	1				
Preparation present	**Wet**	**Pinned**	**Slide**	**Dry**	
		Yes			
Region	Neotropical				
Geography	**Continent**	**Country**	**Island Group**	**Island**	
	South America	Brazil			
Country geography	**Province/State/Territory**	**District/County/Shire**			
	Santa Catarina				
Collection Number	Str-1966				
Site #	Str-1339				
Collector(s)	O. Staudinger				
Collected date(s)	1894 to 1894				
Multimedia					

Parameters of the Cleaning Task

Several dimensions:

- Data Syntax – data format, defining data dictionaries (support from normalized tables, defining reserved vocabulary)
- Table Design – content informs design (case 1)
- Source <-> Destination – possibility of mis-match
- Convenience / Expediency – freezing schema and data (technical side, social side)

Parameters – Data Syntax

- Make the same things the same, e.g., names
- Dates - textual dates (Spring 1910), dates (10/4/2006)
- Other measurements, e.g., lat, lon, depth, width (units of measure (feet vs. meters)
- Data types e.g., text, number
- Authority lists, restricted vocabulary, e.g., taxonomy

De-duplication: Name Example

- Names have a way of propagating – 15 variations of the same name

Collector	Collector	Collector
L. A. de Escobar	Linda Albert	Linda E.
Katherine Albert de Escobar	L. C. A. de Escobar	Katherine de Albert
L. Albert de Escobar	Linda C. Albert	Linda Catherine Albert
L. de Escobar	L. A. Escobar	K. de Escobar
L. Escobar	LAE	L. K. A. de Escobar

First	Middle	Last	Brief
Linda	Katherine Albert de	Escobar	L. K. A. de Escobar

De-duplication: Locality Example

Analyzing this set of records revealed to the field biologist that they were all the SAME!

- Santa Rosa National Park, Sector Murcielago
- Area de Conservacion Guanacaste, Sector Santa Rosa National Park, Murcielago
- Guanacaste National Park, Santa Rosa Section, at Murcielago
- Guanecaste Conservation Area, Murcielago
- Guanecaste, Parque Nacional Santa Rosa Section, Sector Murcielago

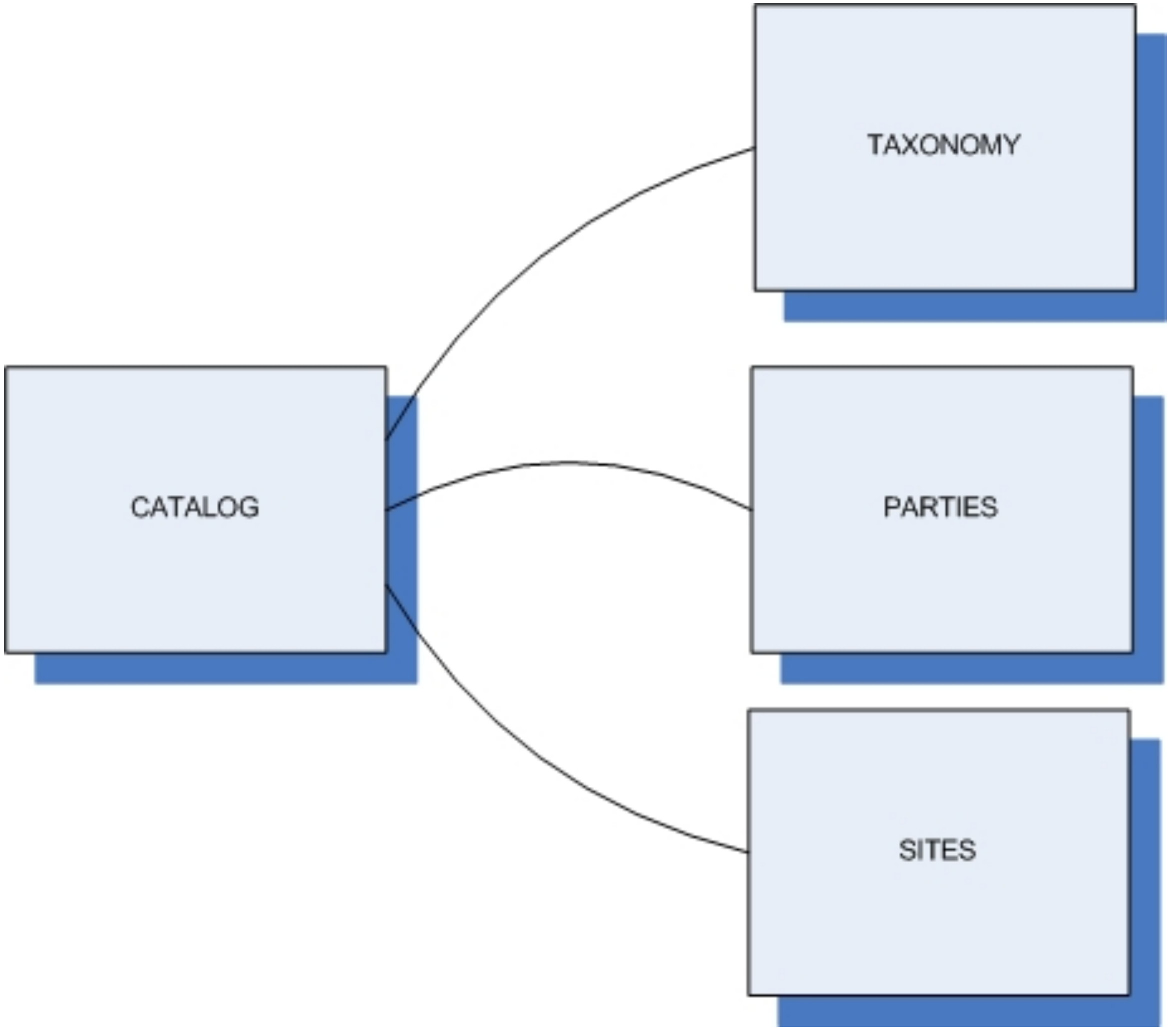
Preferred form is: Santa Rosa National Park (Guanacaste Conservation Area), Sector Murciélago

Parameters - Data Mapping Mis-Match

The source and the destination databases may not match well, due to different purposes, not to mention the obvious differences of different designs/designers.

- Part of scrubbing is mapping the source to the destination, and being sure not to leave anything behind (legacy fields are great).

Normalization – the ultimate goal?



Parameters - Design

For case (1) -

- Data cleaning is a means to design your catalogue. It is an opportunity to become familiar with your data.
- When cleaning data you'll recognize patterns of problems, you should note these and design your catalog to minimize the occurrence of these problems.

Planning - The Plan

- Understand all the parameters before starting - one person in charge
- Close interaction with all members of the team, especially if the work is across multiple databases, and the scrubbing is dispersed – it's a team exercise, shared goals, and tools

Planning - Which Standards?

However long and hard you plan and estimate, cleaning will take longer – count on it. Set standards, and decide on deviation leeway: how much backtracking you are willing to do to make things in sync?

- database
- specimen label, in the record book, or invoices

Consider : Is this an opportunity to take inventory?

Planning - Which Standards?

No matter what line you draw, you'll cross it. Give yourself latitude to correct large mistakes with your data.

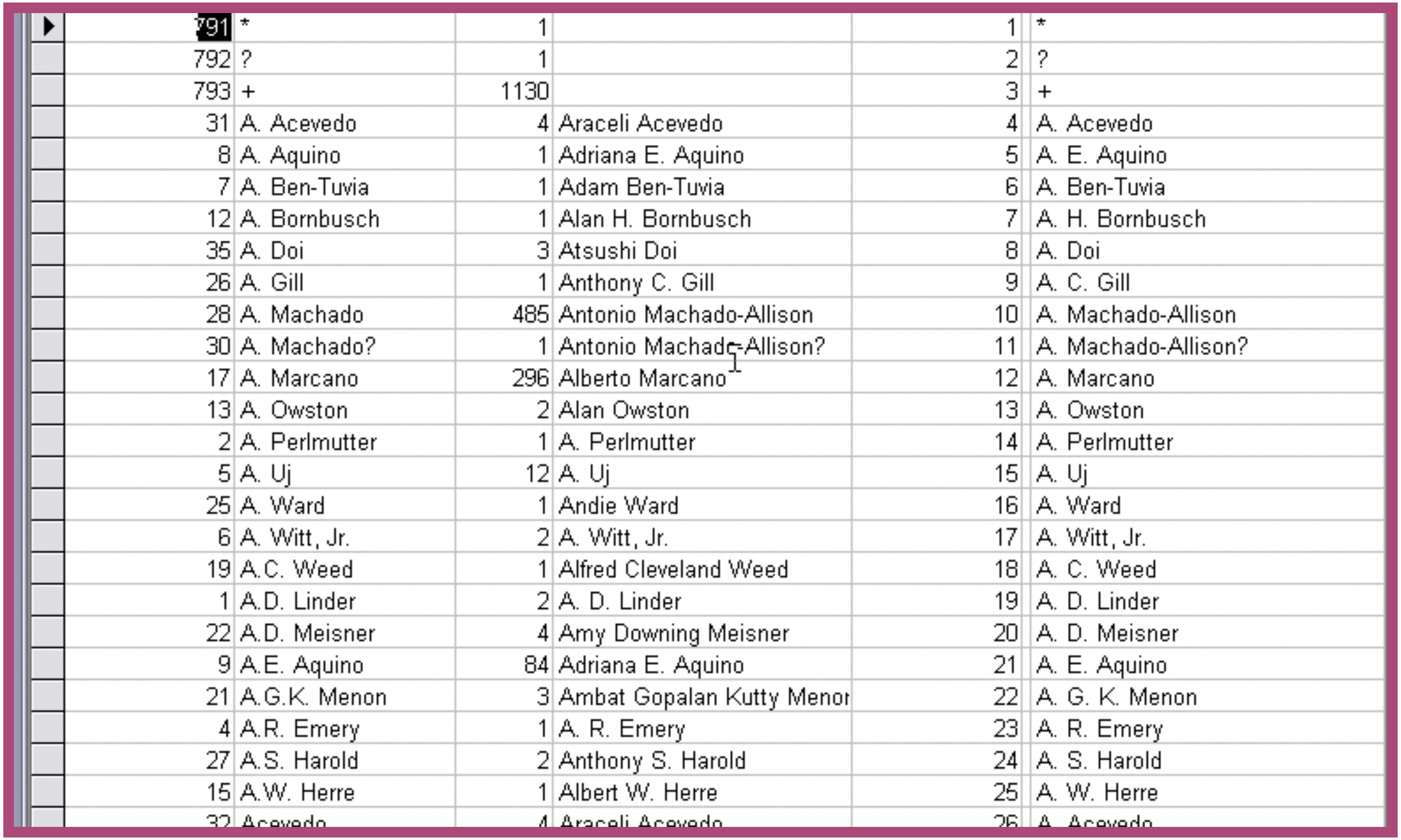
- e.g, if your catalog does not allow duplicate catalog numbers and in your your cleaning you find duplicates, you'll need to re-catalog at least one specimen and determine why it has the wrong number .

Summary

- Definition and Goals
- Parameters of the cleaning task are informed by the context
  - syntax, data types, data mapping,
  - convenience/expediency, design
- Process: the plan, standards

Questions?

People Name Mapping



▶	791	*	1		1	*
	792	?	1		2	?
	793	+	1130		3	+
	31	A. Acevedo	4	Araceli Acevedo	4	A. Acevedo
	8	A. Aquino	1	Adriana E. Aquino	5	A. E. Aquino
	7	A. Ben-Tuvia	1	Adam Ben-Tuvia	6	A. Ben-Tuvia
	12	A. Bornbusch	1	Alan H. Bornbusch	7	A. H. Bornbusch
	35	A. Doi	3	Atsushi Doi	8	A. Doi
	26	A. Gill	1	Anthony C. Gill	9	A. C. Gill
	28	A. Machado	485	Antonio Machado-Allison	10	A. Machado-Allison
	30	A. Machado?	1	Antonio Machado-Allison?	11	A. Machado-Allison?
	17	A. Marcano	296	Alberto Marcano	12	A. Marcano
	13	A. Owston	2	Alan Owston	13	A. Owston
	2	A. Perlmutter	1	A. Perlmutter	14	A. Perlmutter
	5	A. Uj	12	A. Uj	15	A. Uj
	25	A. Ward	1	Andie Ward	16	A. Ward
	6	A. Witt, Jr.	2	A. Witt, Jr.	17	A. Witt, Jr.
	19	A.C. Weed	1	Alfred Cleveland Weed	18	A. C. Weed
	1	A.D. Linder	2	A. D. Linder	19	A. D. Linder
	22	A.D. Meisner	4	Amy Downing Meisner	20	A. D. Meisner
	9	A.E. Aquino	84	Adriana E. Aquino	21	A. E. Aquino
	21	A.G.K. Menon	3	Ambat Gopalan Kutty Menon	22	A. G. K. Menon
	4	A.R. Emery	1	A. R. Emery	23	A. R. Emery
	27	A.S. Harold	2	Anthony S. Harold	24	A. S. Harold
	15	A.W. Herre	1	Albert W. Herre	25	A. W. Herre
	32	Acevedo	4	Araceli Acevedo	26	A. Acevedo

Parameters - Tips

- Decide which can be done without freezing data and which need to be handled with other tools.
- After cleaning data in your live platform, you should decide if it is possible to alter your existing database to prevent bad data from being re-entered. (e.g., convert a text field to a look-up field with values built from the cleaned data.)

Techniques

- **Analysis first**
- **Tokenization & Packetization** : allows for massive cleaning of a target field, but requires analysis, part of mapping exercise
- **Mapping tables** : removes the problematic data from the live version and allows manipulation without affecting your data set. Once the work has been done on unique values, those changes can be applied at exportation. Allows you to clean data in a field, as well as parse apart data of a single field into many fields.
- **Encoding** : allows partial deferment of cleaning, or transformations at exportation
- **Mirror DB tables/modules** : helps shape direction towards the goal, saves money
- **Typos, language variations, abbreviations, diacritic marks** : prevents duplication of names
- **Site / Locality descriptions** :
- **Determination Qualifiers** :

Techniques

- **Tokenization & Packetization** : make a unique token out of each data parcel, like a name, and then re-format it for export. E.g., M. Thayer, J. Boone, and Beka Shuman becomes M. K. Thayer\J. Boone\R. Baquiran as a packet of Brief Names for linking into the People & Organizations table.
- **Mapping tables** : use to scrub tokens
- **Encoding** : field extensions, and notes fields, scrub now or later
- **Mirror your database tables/modules** : Catalogue, Parties, Taxonomy, Multimedia, Collection Events & Sites
- **Typos, language variations, abbreviations, diacritic marks** : prevents duplication of names
- **Qualifiers** : * and ? In taxa can be treated at cf. and aff. when applied to catalogue entries and in general need to be dealt with as far as possible – can't search for.