Plague in Iquique, 1903: Transcribing non tabular data using DataScribe

by Hernán Adasme

Introduction

This case study examines how to transcribe non-tabular or semi-structured data using DataScribe. Historical sources come in all shapes and forms; DataScribe is a tool that enables researchers not only to transcribe data, but to apply an external tabular shape to a semi or non-structured historical source. The case study starts with an historical overview of the project *Plague in Iquique 1903*, followed by description of the sources used. Then, the case study details the process of organizing and translating the sources into separate Omeka items that can be used to create a DataScribe project in the DataScribe module. Next, the case study walks the reader through the process creating a transcription form that captures information that suits the historical questions posed over the sources. Finally, the case study delves into how the data is being recorded in DataScribe, and some possible research paths that a DataScribe dataset allows historians to walk.

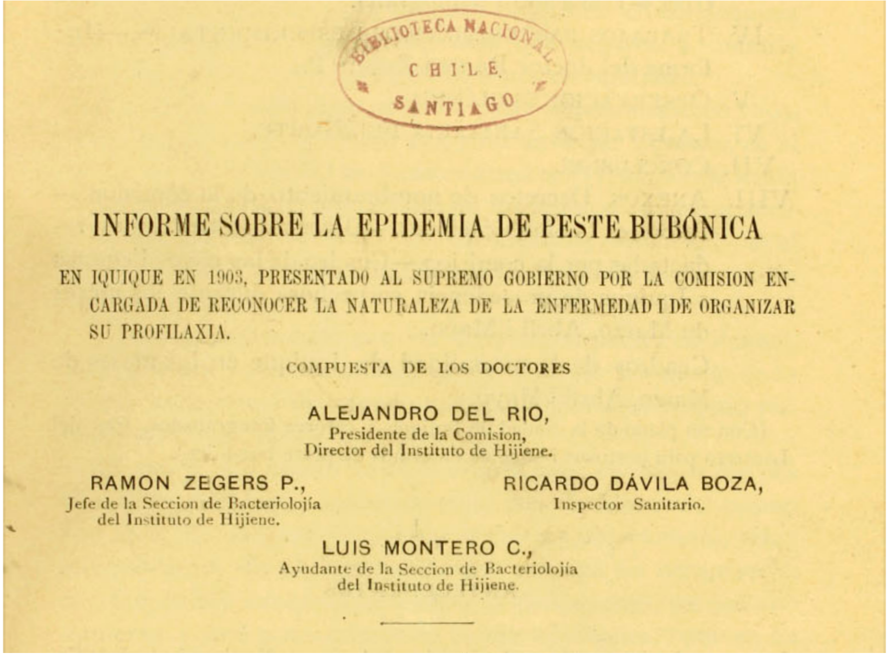
Historical Context

In May 1903 the bubonic plague arrived in Iquique, a port city located in the north of Chile. The arrival of the plague in Iquique was part of a wave of outbreaks that occurred in the Atlantic and Pacific coasts of South America between the late nineteenth-century and the mid twentieth-century. The plague showed up in several ports of Chile between 1903 and 1928, in successive outbreaks that also threatened the port of Valparaiso and northern city of Arica.

To halt the spread of the plague further south, the Chilean government put together a medical commission to study its causes and set containment measures. The commission, formed by a group of physicians and bacteriologists led by Doctor Alejandro del Rio, arrived in Iquique on June 1st 1903 and started working immediately to keep the city safe. Iquique was of crucial importance to the nitrate industry and to the government itself, since nitrate exports supplied 25% of the yearly fiscal revenues of the Chilean Treasury Department[1].

The commission visited and treated the sick, established a bacteriological laboratory, and kept detailed records of each case. Doctor Alejandro del Río and the local authorities arranged a *lazaretto* to take care of those infected with the plague. After three months of hard work, the commission and the local authorities had almost fully controlled the plague.

[1] See Sergio Gonzalez Miranda, “El ciclo de expansión del salitre” en *Camanchacha. Salitre: reencuentro, añoranza, realidad*, Iquique, Taller de Estudios Regionales, 1987. P.12



INFORME SOBRE LA EPIDEMIA DE PESTE BUBÓNICA

EN IQUIQUE EN 1903, PRESENTADO AL SUPREMO GOBIERNO POR LA COMISION ENCARGADA DE RECONOCER LA NATURALEZA DE LA ENFERMEDAD I DE ORGANIZAR SU PROFILAXIA.

COMPUESTA DE LOS DOCTORES

ALEJANDRO DEL RIO,
Presidente de la Comision,
Director del Instituto de Hijiene.

RAMON ZEGERS P.,
Jefe de la Seccion de Bacteriolojía
del Instituto de Hijiene.

RICARDO DÁVILA BOZA,
Inspector Sanitario.

LUIS MONTERO C.,
Ayudante de la Seccion de Bacteriolojía
del Instituto de Hijiene.

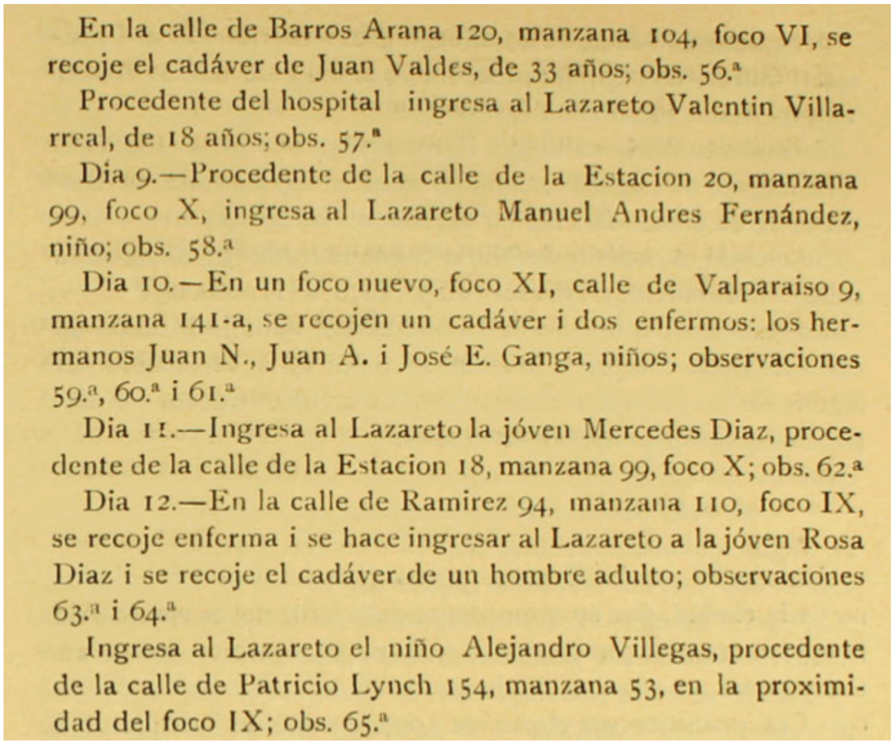
Figure 1. Cover of the Report of the Bubonic Plague in Iquique

The commission produced several documents that contained rich information about the living conditions in Iquique, the people's attitudes about the plague and the death, the clinical history of the sick, the treatments administered and the hygienic policies implemented by the authorities. This case study uses the *Informe Sobre la Epidemia de Peste Bubónica en Iquique in 1903* (Report about the Bubonic Plague in Iquique in 1903) presented to the government of Chile by the medical commission in 1904. The medical report offers textual information about the first 167 plague cases treated by the medical commission between May and September 1903. For most of the cases, the list provides location, date of the symptoms' onset and case's diagnosis, and demographic data of the patient. We'll analyze the data in the About the Data section of this case study.

The *Plague in Iquique 1903* project aims to analyze the advancement of the plague in the port of Iquique by applying spatial and time-series methodologies. DataScribe will make possible the transcription of information from the report to a tabular form, structured into rows and columns. The organization into rows and columns will allow us to export the data in a format suitable for computational analysis.

About the Data

This case study translates a list containing the first 167 cases treated by the commission included in the medical report using DataScribe. Each case in the list includes *at least* four crucial pieces of data: the case number of the sick person, the status of the case (whether the person was found dead or alive), the patient's address, and the date of the diagnosis. Most of the cases also show the gender, age, name, the date in which the person entered the public lazaretto (which usually matches with the diagnosis) and the outbreak phase of each individual. A minor subset of cases offers information about links between cases, their symptoms and the date in which they erupted, and the death or recovery date of the sick.



En la calle de Barros Arana 120, manzana 104, foco VI, se recoje el cadáver de Juan Valdes, de 33 años; obs. 56.ª

Procedente del hospital ingresa al Lazareto Valentin Villarreal, de 18 años; obs. 57.ª

Dia 9.—Procedente de la calle de la Estacion 20, manzana 99, foco X, ingresa al Lazareto Manuel Andres Fernández, niño; obs. 58.ª

Dia 10.—En un foco nuevo, foco XI, calle de Valparaiso 9, manzana 141-a, se recojen un cadáver i dos enfermos: los hermanos Juan N., Juan A. i José E. Ganga, niños; observaciones 59.ª, 60.ª i 61.ª

Dia 11.—Ingresa al Lazareto la jóven Mercedes Diaz, procedente de la calle de la Estacion 18, manzana 99, foco X; obs. 62.ª

Dia 12.—En la calle de Ramirez 94, manzana 110, foco IX, se recoje enferma i se hace ingresar al Lazareto a la jóven Rosa Diaz i se recoje el cadáver de un hombre adulto; observaciones 63.ª i 64.ª

Ingresa al Lazareto el niño Alejandro Villegas, procedente de la calle de Patricio Lynch 154, manzana 53, en la proximidad del foco IX; obs. 65.ª

Figure 2. Textual organization of the List in the Report

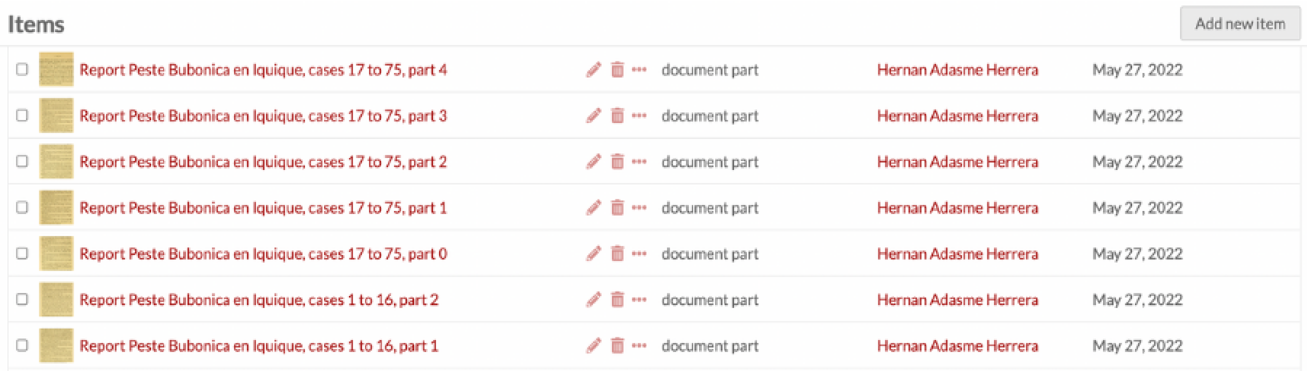
It is important to note that the list of 167 cases does not appear in one single section of the medical report. Instead, the whole list is divided in three parts, which can be found in different sections of the report. The list follows a consistent format throughout the sections; the case numbers are organized consecutively regardless of the divisions, and the basic features of each case (ID, status, address, and diagnosis date) remain in place throughout all the partitions. For example, cases 1-16 are located in pages 10-15 of the original report, and cases 17-75 can be found in pages 20-24.

Setting up the DataScribe Project part 1: items in the Omeka S Install

Adding the items into Omeka S

Before setting up the DataScribe project, all the items getting transcribed should be stored in an Omeka S installation. Keep in mind that DataScribe is a module of Omeka S; all the items must be kept in Omeka S in order to be available to the DataScribe tool. Omeka S is the vault from which DataScribe retrieves items to build datasets in a project. As a data management system, Omeka S allows the user to manage the data carefully, not only in terms of data allocation, but also at the metadata level.

It is important to think carefully how your historical sources would translate into being items in the Omeka S installation. In this case, each page containing sections of the plague cases list becomes an item in the Omeka S installation. This will give the flexibility to group the items in several different forms, depending on the types of questions we are posing over the data, and the ways in which we want to set up the transcription workflow. Each item will have a media file, which corresponds to a screenshot in png format of every page of the list with the 167 plague cases.



Items

□ Report Peste Bubonica en Iquique, cases 17 to 75, part 4	document part	Hernan Adasme Herrera	May 27, 2022
□ Report Peste Bubonica en Iquique, cases 17 to 75, part 3	document part	Hernan Adasme Herrera	May 27, 2022
□ Report Peste Bubonica en Iquique, cases 17 to 75, part 2	document part	Hernan Adasme Herrera	May 27, 2022
□ Report Peste Bubonica en Iquique, cases 17 to 75, part 1	document part	Hernan Adasme Herrera	May 27, 2022
□ Report Peste Bubonica en Iquique, cases 17 to 75, part 0	document part	Hernan Adasme Herrera	May 27, 2022
□ Report Peste Bubonica en Iquique, cases 1 to 16, part 2	document part	Hernan Adasme Herrera	May 27, 2022
□ Report Peste Bubonica en Iquique, cases 1 to 16, part 1	document part	Hernan Adasme Herrera	May 27, 2022

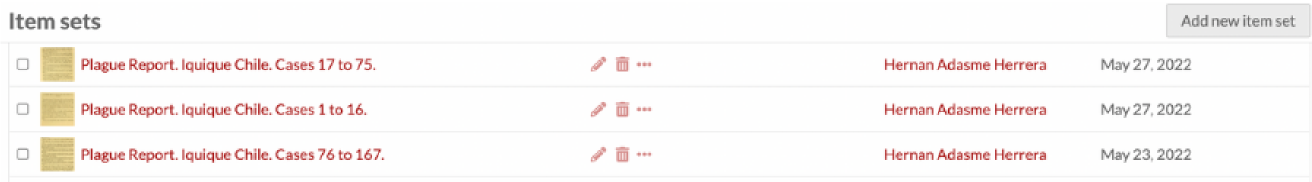
Figure 3. Items in the Omeka S Install

Since the whole project only contains 18 items, I managed to create each item in the Omeka Install with its corresponding metadata manually. In the Omeka S installation, I created a metadata template called *Plague in Iquique* and I selected the class *document part* --which is available in the class dropdown menu-- to build the metadata for all the items. For larger projects we recommend using the *Omeka CSV Import* plugin, which is the best way to import a large number of items and their accompanying metadata into Omeka S.

Creating Item Sets

Once all your items are uploaded into the Omeka S installation, the next step consists in creating Omeka S item sets. DataScribe uses Omeka S item sets as the basis for DataScribe datasets. For this case study, the Omeka S item sets have been created following the original partitions of the list with plague cases. I decided to group the items in correspondence to the pages in the report in order to keep the consistency between the items-sets and the list partitions of the original document. Had the multipage partitions become a multi-image single item, the DataScribe project would have drawn only from one item set to create datasets. That decision would have severely reduced the alternatives to arrange the transcription workflow. In addition, since the amount of information across the list partitions varies, having independent item-sets in congruence with the list partitions captures more adequately the nature of the source. This addresses the fact that there are differences between the details included in each of the partitions of the report, being the most data rich the portion that include the cases 17-75. Creating item sets and DataScribe datasets in correspondence with the page partitions in the report aims to preserve the features of the original document.

The first Omeka S item set titled *Plague Report. Iquique Chile. Cases 1 to 16*, contains the four items corresponding to the pages 10-15 of the Medical Report. In total, three different item sets have been created with the items containing the items 10-16, 17-75, 76-167, respectively. with the block of pages in the original document



Item sets

Add new item set

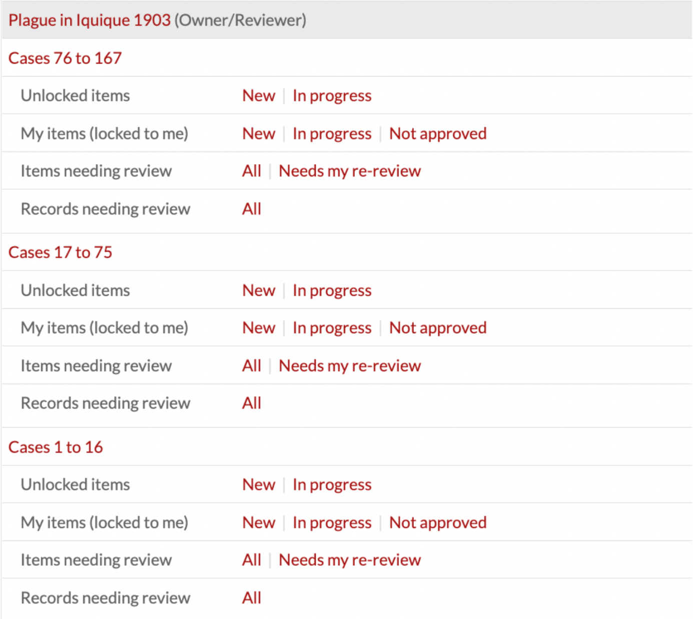
□	Plague Report. Iquique Chile. Cases 17 to 75.	···	Hernan Adasme Herrera	May 27, 2022
□	Plague Report. Iquique Chile. Cases 1 to 16.	···	Hernan Adasme Herrera	May 27, 2022
□	Plague Report. Iquique Chile. Cases 76 to 167.	···	Hernan Adasme Herrera	May 23, 2022

Figure 4. Item Sets in the Omeka S Install

Setting up the DataScribe Project part 2: creating a DataScribe project

Creating a DataScribe project

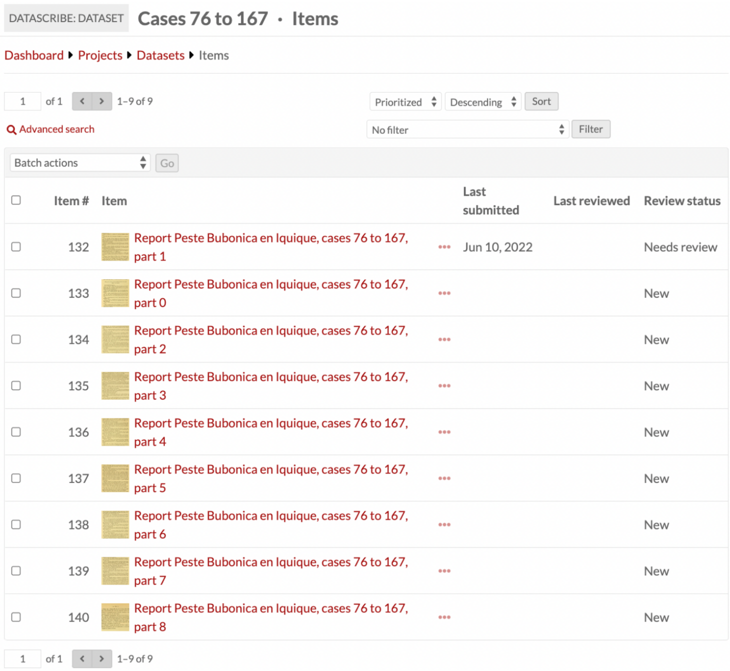
Once all items have been created in the Omeka S install and grouped into Item sets, all the elements required to create *Plague in Iquique 1903* DataScribe project are available. For a detailed description on how to create a DataScribe project check the [DataScribe documentation](). The *Plague in Iquique 1903* project includes three datasets; each of those three datasets corresponds to one of the item sets in the Omeka S install. Figure 5 shows three datasets: Cases 1 to 16, Cases 17 to 75, and Cases 76 to 167.



Plague in Iquique 1903 (Owner/Reviewer)	
Cases 76 to 167	
Unlocked items	New | In progress
My items (locked to me)	New | In progress | Not approved
Items needing review	All | Needs my re-review
Records needing review	All
Cases 17 to 75	
Unlocked items	New | In progress
My items (locked to me)	New | In progress | Not approved
Items needing review	All | Needs my re-review
Records needing review	All
Cases 1 to 16	
Unlocked items	New | In progress
My items (locked to me)	New | In progress | Not approved
Items needing review	All | Needs my re-review
Records needing review	All

Figure 5. Datasets in the Plague in Iquique:1903 DataScribe project

Each dataset contains as many items as pages in the report were dedicated to each partition of the list. For instance, the dataset Cases 76 to 167 includes 9 different items: *Report Peste Bubonica in Iquique, cases 76 to 167, part 0,* all the way up to part 8. Figure 6 showcases the items contained in the dataset.



DATASCRIBE: DATASET Cases 76 to 167 · Items

Dashboard ▸ Projects ▸ Datasets ▸ Items

1 of 1 ‹ › 1–9 of 9

Prioritized Descending Sort

Advanced search

No filter Filter

Batch actions Go

□	Item #	Item		Last submitted	Last reviewed	Review status
□	132	Report Peste Bubonica en Iquique, cases 76 to 167, part 1	•••	Jun 10, 2022		Needs review
□	133	Report Peste Bubonica en Iquique, cases 76 to 167, part 0	•••			New
□	134	Report Peste Bubonica en Iquique, cases 76 to 167, part 2	•••			New
□	135	Report Peste Bubonica en Iquique, cases 76 to 167, part 3	•••			New
□	136	Report Peste Bubonica en Iquique, cases 76 to 167, part 4	•••			New
□	137	Report Peste Bubonica en Iquique, cases 76 to 167, part 5	•••			New
□	138	Report Peste Bubonica en Iquique, cases 76 to 167, part 6	•••			New
□	139	Report Peste Bubonica en Iquique, cases 76 to 167, part 7	•••			New
□	140	Report Peste Bubonica en Iquique, cases 76 to 167, part 8	•••			New

1 of 1 ‹ › 1–9 of 9

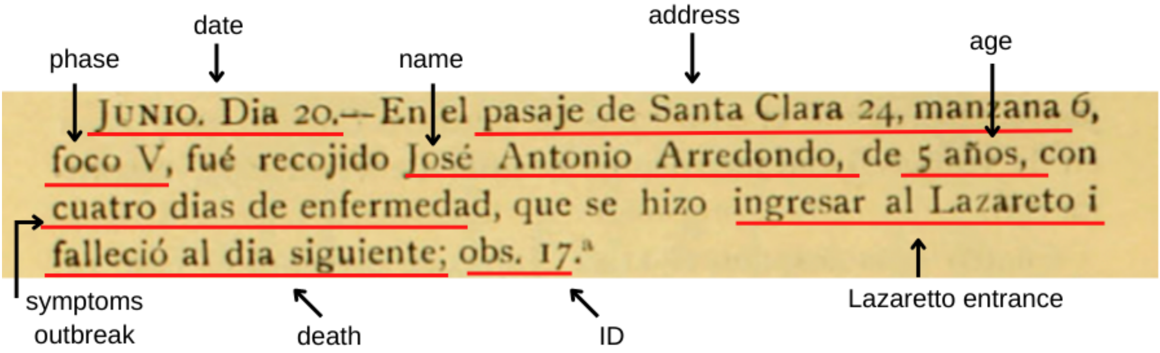
Figure 6. Items in the dataset

Building the Transcription Form: giving structure to unstructured data.

Building the transcription form is a crucial part of setting up a DataScribe project. For *Plague in Iquique 1903* the transcription form provides the basis for transforming semi structured data into a set of data structured in a tabular form. The transcription form is the tool that makes the transcription posible. DataScribe allows the user to exploit both the qualitative and quantitative information available in the commission's Medical Report, by framing the textual non-structured data into a transcription form that translates into rows and columns. Although a spreadsheet would serve a similar purpose --that is, organizing and storing the information into rows and columns-- DataScribe guarantees a consistent and standardized transcription workflow especially when more than one transcriber participates in the process.

The process of transcribing data from the plague list starts by visually assessing the sources, in order to establish what fields and datatypes are suitable for accurately capturing the information available. Figure number 7 presents a map of the source, which highlights the pieces of data available in the textual listing of plague cases. Keep in mind that you can build multiple datasets and multiple transcription forms using the same Omeka S item set; this means that DataScribe allows you to capture different takes of information from the same source.

Figure 7 underscores the explicit pieces of data available for the vast majority of cases listed, which are able to be transcribed in the DataScribe project.



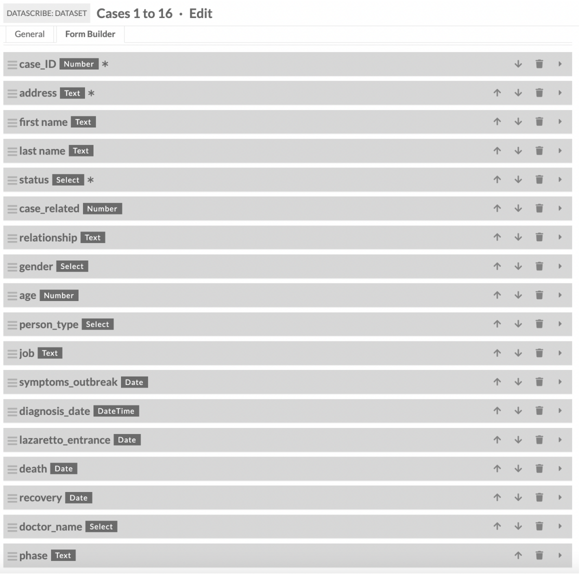
JUNIO. Dia 20.—En el pasaje de Santa Clara 24, manzana 6, foco V, fué recojido José Antonio Arredondo, de 5 años, con cuatro dias de enfermedad, que se hizo ingresar al Lazareto i falleció al dia siguiente; obs. 17.ª

date
address
age
phase
name
symptoms
outbreak
death
ID
Lazaretto entrance

Figure 7. Map of the data pieces included in the case list.

Fields in the Transcription Form

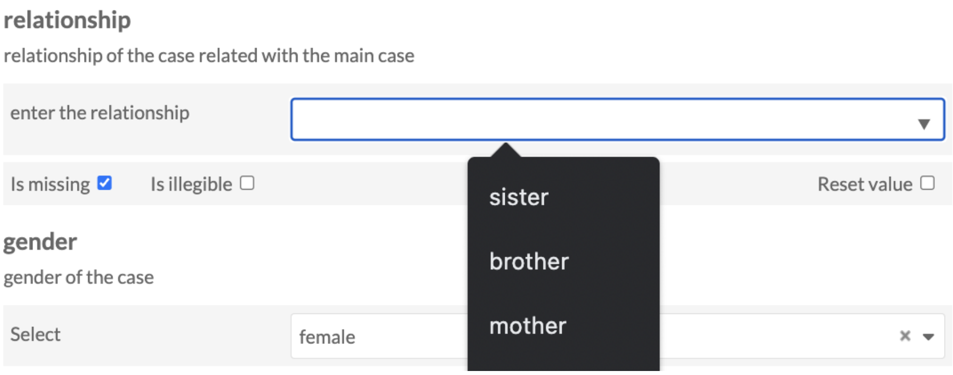
The transcription form contains 18 fields. Four fields in the transcription form are defined as required fields: *case_ID*, *address*, *diagnosis-date,* and *status*. The reason why *first_name* and *last_name* are not set as required fields is because the commission was not able to identify a small number of dead bodies, and thus their names are absent from the report. These four required fields lie at the core of the research project; they are the bare minimum necessary to identify the case, perform spatial analysis research methods, and examine the data from a time-series perspective.



case_ID Number *
address Text *
first name Text
last name Text
status Select *
case_related Number
relationship Text
gender Select
age Number
person_type Select
job Text
symptoms_outbreak Date
diagnosis_date DateTime
lazaretto_entrance Date
death Date
recovery Date
doctor_name Select
phase Text

Figure 8. Fields in the Transcription Form. The asterisk indicates a required field.

Six fields in the form are set as text datatype: *first_name*, *last_name*, *address*, *relationship*, *phase* and *job*. Within this category, both *phase* and *relationships* offer input labels to help the transcription process. Input labels are options displayed in the text box to help the transcriptor to populate the field. Even though the field *phase* conceptually refers to the consecutive number of plague outbreaks, the field was set as a text because the original report identifies the outbreak phases using Roman numerals. I have decided not to transcribe the Roman numerals into Arabic numerals to avoid errors in the transcription due to unfamiliarity with the Roman notation. Figure 9 shows the input labels that are suggested by the transcription form to avoid spelling errors during the transcription process for the case of the *relationship* field



relationship

relationship of the case related with the main case

enter the relationship

Is missing ☑ Is illegible ☐ Reset value ☐

sister

brother

mother

gender

gender of the case

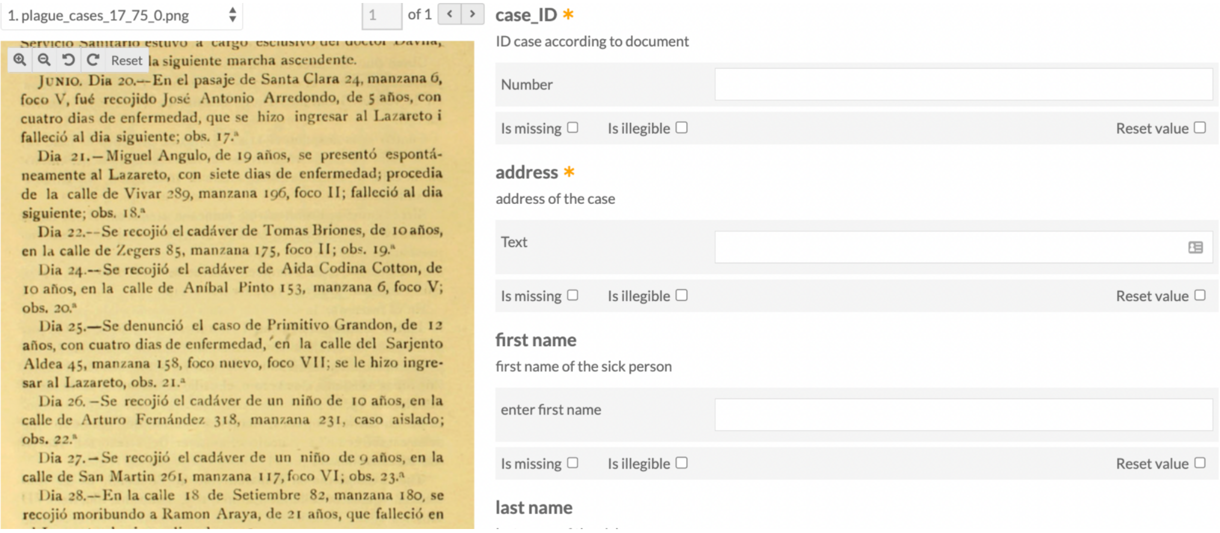
Select female

Figure 9. Input labels in the field that captures the *relationship* between cases

Three fields in the form are numeric: *case_ID*, *case_related*, and *age*. The field *case_ID* captures the original ID used by the commission in the report; this field has been set as the primary key, that is, the field by which the record will be identified. *Case_related* references the ID of another case in the list that is somehow related to the principal case being transcribed. It is conceptually designed to be a self-referencing or recursive foreign key, in the sense that it is a foreign key that references the same table.

Five fields in the transcription form correspond to date and datetime data type. The transcription collects the *diagnosis_date* (which is also a required field), *symptoms_outbreak* date, *lazzaretto_entrance* date, *death* and *recovery* dates*.* The field *diagnosis_date* is a datetime datatype, which allows the user to enter the precise time, available for a few cases. Both date and datetime will be exported in ISO 8601 format using the Gregorian Calendar. These pieces of date and time data are crucial to perform computational time series analysis and trace the advancement of the plague outbreak.

Finally, four fields are of Select type, which create a dropdown with options to choose from: *status*, *gender*, *person_type*, and *doctor's_name*. Status offers two options to choose from: alive or dead. As explained above, this is a piece of information that should be inferred by the transcriber, because it is exclusively mentioned when the medical commission found and diagnosed a dead body. The field *gender* shows a dropdown with only two alternatives: male and female. For most of the cases, the gender of the sick person should be inferred from the name and the gendered use of the Spanish language. The field *person_type* is used when the age of the patient is not explicitly mentioned, but the source describes the sick person as either a baby (*bebé*), a kid (*niño* or *niña*) or an adult (*hombre adulto* or *mujer adulta*). Finally, the field *doctor_name* captures the physician who performed the diagnosis and first treatments, if mentioned by the source.



Servicio Sanitario estuvo a cargo esclusivo del doctor Davila,
la siguiente marcha ascendente.

JUNIO. Dia 20.—En el pasaje de Santa Clara 24, manzana 6, foco V, fué recojido José Antonio Arredondo, de 5 años, con cuatro dias de enfermedad, que se hizo ingresar al Lazareto i falleció al dia siguiente; obs. 17.ª

Dia 21.—Miguel Angulo, de 19 años, se presentó espontáneamente al Lazareto, con siete dias de enfermedad; procedia de la calle de Vivar 289, manzana 196, foco II; falleció al dia siguiente; obs. 18.ª

Dia 22.—Se recojió el cadáver de Tomas Briones, de 10 años, en la calle de Zegers 85, manzana 175, foco II; obs. 19.ª

Dia 24.—Se recojió el cadáver de Aida Codina Cotton, de 10 años, en la calle de Aníbal Pinto 153, manzana 6, foco V; obs. 20.ª

Dia 25.—Se denunció el caso de Primitivo Grandon, de 12 años, con cuatro dias de enfermedad, en la calle del Sarjento Aldea 45, manzana 158, foco nuevo, foco VII; se le hizo ingresar al Lazareto, obs. 21.ª

Dia 26.—Se recojió el cadáver de un niño de 10 años, en la calle de Arturo Fernández 318, manzana 231, caso aislado; obs. 22.ª

Dia 27.—Se recojió el cadáver de un niño de 9 años, en la calle de San Martin 261, manzana 117, foco VI; obs. 23.ª

Dia 28.—En la calle 18 de Setiembre 82, manzana 180, se recojió moribundo a Ramon Araya, de 21 años, que falleció en

case_ID *
ID case according to document
Number
Is missing □ Is illegible □ Reset value □

address *
address of the case
Text
Is missing □ Is illegible □ Reset value □

first name
first name of the sick person
enter first name
Is missing □ Is illegible □ Reset value □

last name

Figure 10. Transcription work area

The same fields already described can be also grouped according to three families of data: spatial data, time data, and relational data. The address of each case is recorded as a string that includes the address and the block number. The fields that capture multiple dates found in the historical source translate the information into a standardized data type format which in turn ensures the homogeneity among date formats. Finally, some cases provide information about how cases relate with one another. That is, what plague cases correspond to a cluster of cases among related people. The fields *case_related* and *relationship* have been designed to get the ID of the case as a foreing key, and the type of relationship between cases.

If mentioned, the type of relationship can be selected from a dropdown menu that prevents typing errors.

Implicit data: the challenge of transcribing semi-structured sources.

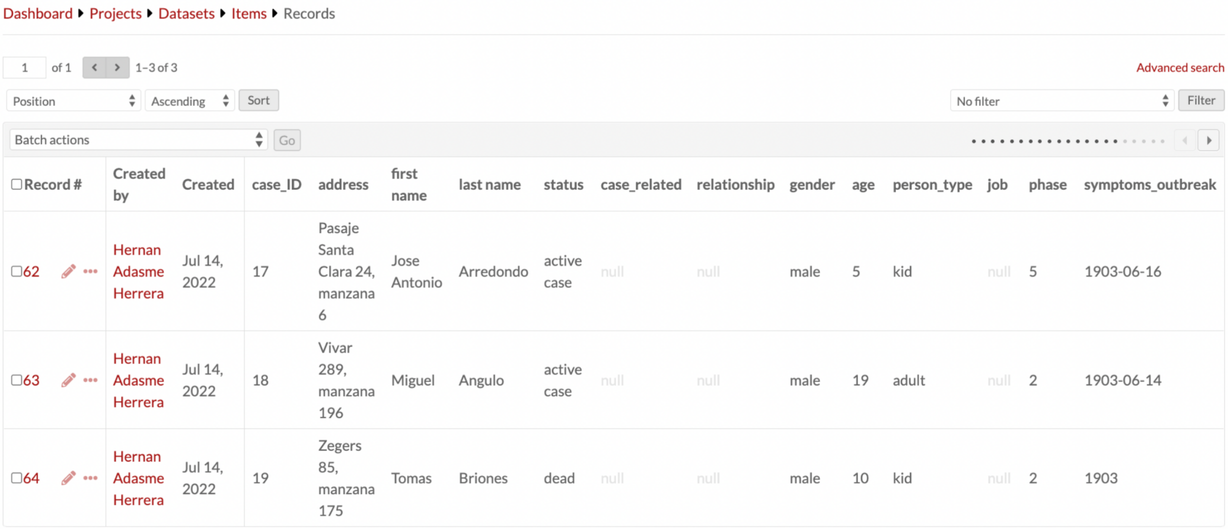
It is important to note that some bits of information are not explicit and should be inferred from the source. For instance, the field status, which refers to the condition of case at the moment of the diagnosis --whether the sick person was found dead or alive-- it is not explicitly mentioned in the listing; only the cases that had passed away were described as “found dead” or “corpse found at”. Thus, although not plainly mentioned, we can infer that the ones not labeled as dead were found alive. Consequently, the field status in the transcription forms allows the transcriber to select the option “alive” or “dead” from a dropdown menu.

The same occurs with gender; as mentioned before, for most of the cases the gender of the should be inferred from the name of the patient and the gendered usage of Spanish grammar. Finally, both the fields *diagnosis_date* and *lazarreto_entrance* correspond to the same date -patients were sent to the lazaretto the same day of the diagnosis. This date also matches with the general date that organizes how the cases are presented in the report. For example; if an active case was recorded on June 20th in the commission’s report, both the diagnosis date and the lazzaretto entrance dates would *always* be June 20th. It is important to keep in mind that the list is organized by date; the precise date in which the cases were diagnosed by the doctors is the main criteria by which the commission sorted the cases and determined the ID of each observation. Put differently, is the diagnosis date which mattered to arrange the list, and not the reported onset of the symptoms by the patients.

In order to help the transcription of these bits of information that are not easily translatable, it is important to add guidelines to every dataset in the Plague in Iquique: 1903. Transcribing semi-structured sources poses challenges because the textual information is not always outlined following a strict format: the structure of the text changes from case to case. The guidelines allow adding specific instructions and useful information to ensure the consistency of the transcription work throughout the transcription process.

Records

Figure 11 shows a screenshot of the records in its final form. The record number 62 in the image corresponds to the same piece of data showcased in figure number 7. Image number eleven showcases how the fields *case_id*, *address*, *first_name*, *last_name*, *gender*, *age*, *person_type*, *phase* and *symptoms_outbreak* have been translated from textual semi-structured data to information stored in rows and columns.



□Record #	Created by	Created	case_ID	address	first name	last name	status	case_related	relationship	gender	age	person_type	job	phase	symptoms_outbreak
□62 ✎ •••	Hernan Adasme Herrera	Jul 14, 2022	17	Pasaje Santa Clara 24, manzana 6	Jose Antonio	Arredondo	active case	null	null	male	5	kid	null	5	1903-06-16
□63 ✎ •••	Hernan Adasme Herrera	Jul 14, 2022	18	Vivar 289, manzana 196	Miguel	Angulo	active case	null	null	male	19	adult	null	2	1903-06-14
□64 ✎ •••	Hernan Adasme Herrera	Jul 14, 2022	19	Zegers 85, manzana 175	Tomas	Briones	dead	null	null	male	10	kid	null	2	1903

Figure 11. Records in the dataset

Final Reflections

Although DataScribe has been designed to help the transcription of structured historical data, the software is flexible enough to make possible the transcription of textual and semi-structured historical sources into a tabular form. DataScribe supplies the tools to create the architecture that re-frames the pieces of information into rows and columns. For this project, we used DataScribe to build an external structure that dissects the historical source, permits capturing information of interest, and re-organizes it in a way that allows performing computational methods of analysis. DataScribe minimizes transcription errors by making the data collection more consistent and highly standardized, especially when the workflow includes more than one transcriber.

Transcribing non structured data comes with its own challenges. Since we are translating text into tables, the transcription work requires crystal clear guidelines to help the transcribers to agree and comply with the transcription criteria. The workflow would also require constant review, because the style and the details contained in the written information varies between records. In the case of the sources used for *Plague in Iquique 1903* the amount of information that is not explicit and should be inferred makes the process of transcription difficult and error prone. Not all the fields can be filled by simply copying the information from the corresponding box in the historical source. Transcribing those particular pieces of information requires a fair amount of close reading when creating a transcription record. DataScribe's transcription forms can be constantly updated in order to keep improving the quality and functionality of the transcription work. Multiple datasets can be created to apply multiple transcription forms to the same historical source. All of them will capture some material, while leaving other material behind. DataScribe hopes to come to the rescue of the data going missing.