

Analysis of Dataset

SHRASHTI SINGHAL

Data Curation- IS 531

Department of Information Science

The University of Illinois at Urbana Champaign

Abstract

The dataset contains details of Shiba Clan of Japan which existed in 13th to 17th century from 1240 to 1720 AD. The data is provided in an Excel data file.

The job is to provide a natural language account of the data file. We can optionally describe the dataset with a help of a conceptual schema.



Find the dataset here:

I. Description

The dataset is in Excel format. The file is arranged in a hierarchical order to provide the detail hierarchical details about the progenitor and their children.

- **Family Name:** It is same for all the records, which is ODA.
- **Name:** It is the name of the ruler
- **Birth Date:** When exact birthdate is unknown, AFTER keyword is used. Example: Oda Sadanobu (after 1512 - 1580s?)
- **Death Date:** When the exact Death Date is unknown, “?” symbol is used. Example: Oda Sadanobu (after 1512 - 1580s?)
- **District:** Optional District name is given, in case the ruler children move to a different area in Japan. District name is also the name Clan Branch. Example: Oda Hirotsue (1380s - 1450s?) {Haguri Branch} <Haguri-gun, Owari-kuni>. The format of mentioning district and state is < District, State>.
- **Area:** Indicates the Area/State of the district. Shiba Clan was majorly found in 2 areas of Japan, namely Owari-kuni & Echizen-kuni. Area details is optional in each record. It is specifically mentioned along with ruler's name, when they have changed the branch/District.
- **Clan Branch:** Clan Branch is name of the district in the area of Japan, where the ruler children's move to. This detail is mentioned in curly braces. Example: Oda Hirotsue (1380s - 1450s?) {Haguri Branch} <Haguri-gun, Owari-kuni>. Clan Branch name is optional, specially mentioned when the clan branch/district is changed by the ruler.



II. Arrangement

First 4 lines are arranged in 1st columns. It is one time description of the name of the Clan, area and progenitor of the clan. These details remain same throughout the clan. Though there can be expansion of the area of the clan in years to come, but their origin remains the same.

In remaining records, children are arranged in the column to its right and 1 row below the parent. Siblings from the same father are written in the same column, one below the other. The details are organized in a tree structure, to give a visual impression of the hierarchical structure. No 2 names are written in the same row in entire dataset.

III. Observations

- **Number Of Hierarchy:** It shows that number of generations of Shiba Clan is arranged in A to U in excel sheet. This makes it total 21 generations of Shiba Clan.
- **Number of Rulers:** There are total 272 Rulers in Shiba Clan. (It is obtained by : Selecting all the text → Home → Find & Select → Special go to → Blank → Tick → Ok → Right Click → Delete → Shift cells left)
- **Adopted:** Optional adopted detail is mentioned in case, the ruler is not the blood born child of his progenitor.
- **Branch Change:** It is noted that, in few cases of multiple sons, the elder son continues with the same district and area (Same Branch), but the younger sons have to find a different district in the same area.
In case of branch change, the font is bold for the ruler, progenitor of that branch.
- **No. of Sons:** At most 2 rulers have 9 sons, which are Oda Nobuhide (1510 - 1549) and Oda Nobunaga (1534 - 1582)
- **Major Changes:** After its first ruler Oda Chikamoto, his one of the 2 sons continued in Area/State Owari-kuni, while the other son, moved to Echizen-kuni. Therefore Shiba clan was settled in these 2 states. Later on the legacy of each of these sons continued to exist in these 2 states.

Owari –Kuni	Echizen-kuni
There have been multiple incidents, when the next ruler is adopted	There are no incidents of the adopted ruler/son
Rulers kept establishing their reign within different regions/Districts of Owari Kuni	Rulers just stayed in Nyuu-gun region within Echizen-kuni
Death dates of few rulers are unknown	Death dates of all the rulers are unknown
A ruler could have multiple sons. Details of these multiple sons/rulers are given. Not a normal form.	Details of only 1 son/ruler of the previous ruler is given. 1 Normal Form.
In few cases of multiple sons, a different branch is formed within Shiba Clan/Owari Kuni	No cases of branch change. No cases of multiple sons.
Birth Dates of few rulers are unknown.	Birth Dates are known for each ruler.
	

IV. Hierarchy/ Tree Map

The below hierarchy map is to denote how the expansion happened in case of multiple sons of the ruler.

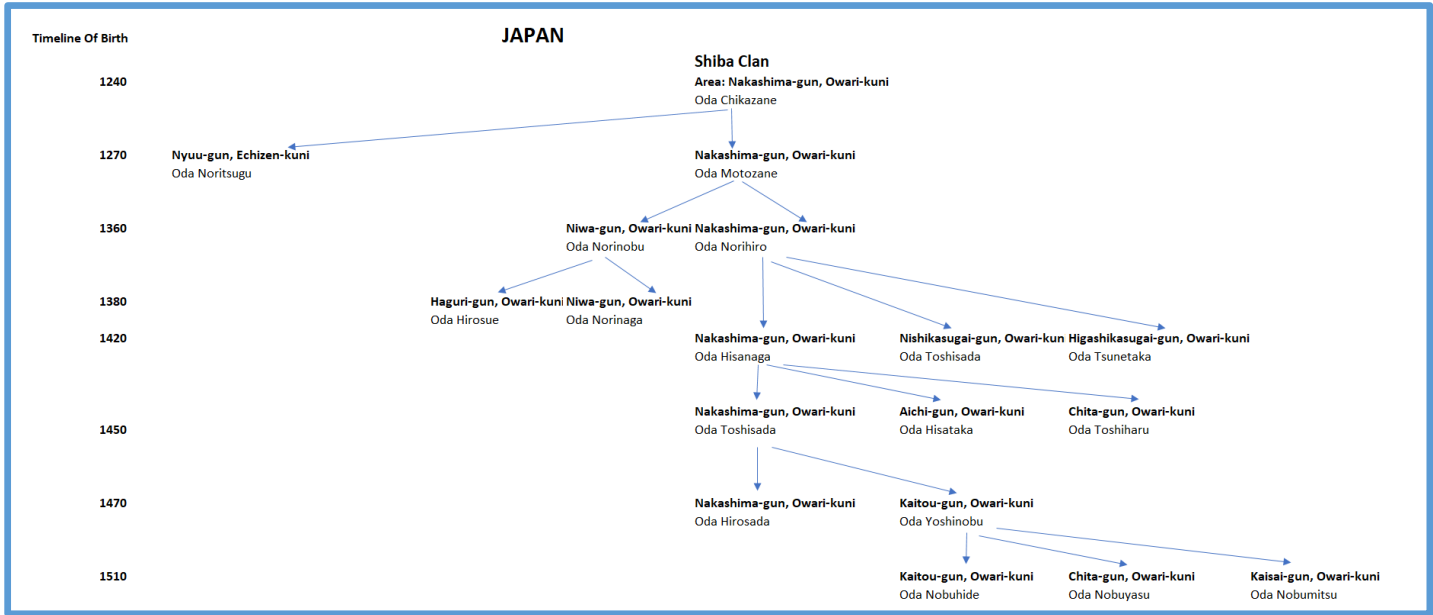


Image 1: The tree map shows the expanse of Shiba Clan, starting from Nakashima Gun to various other districts of Owari-Kuni to different States.

V. Representation

The data presented in the dataset is **not in 1F normal form**. If we organize the details in a relation, everything can be organised singularly in 1 column except the records of the sons.

We can transform the data into 1NF by mentioning at max 9 columns for the sons.

Order: is the hierarchical order of the clan starting from 0 to 21.

Confirm Birth Year/Confirm Death Year: are the Boolean variable with values YES or NO to denote if the birth and death year is exact.

Birth Status: is a 2 value field variable, values can be either blood or adopted to denote if the person is blood borne to his father or the adopted one.

Father order: can be found out by subtracting 1 from the rulers order, but for quick discovery of data, Fathers name and order can be used as a composite key to find the details of the father.

Sons order: can be found out by adding 1 to the order of the ruler, but for quick discovery of data, sons name and order can be used as a composite key to find the details of the son. But for unique identification of 1 data record, fathers name and his order are marked as composite foreign key.

Son's details can be up to 18 columns, 9 for son's name, and 9 columns for sons order.

No. of Siblings/name of Siblings can also be added. For now, these details has been skipped in the schema, without losing any details.

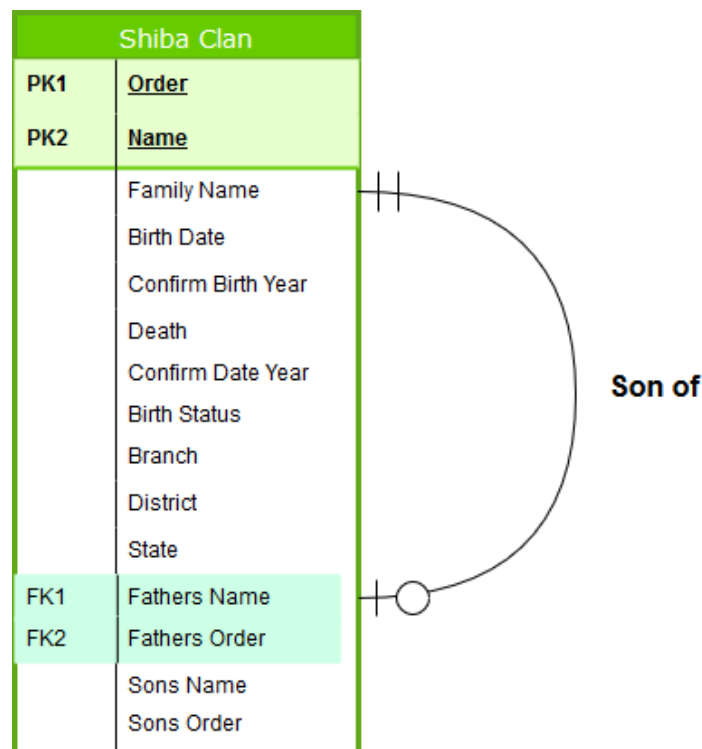


Table 1: Relational Representation of the Dataset- Crows Foot notation

Order	Family		Birth date	Confirm Birth		Death Date	Confirm Death		Birth Status	Branch	District	State	Father Last Name	Father Order	1 Son/s		2 Son/s		3 Son/s	
	Name	Name		Year	Year		Year	Year							Name	Order	Name	Order	Name	Order
0	Oda	Chikazane	1240	YES	1310	NO	Blood	Nakashima	Nakashima-gun	Owari-kuni					Chikamoto	0				
1	Oda	Chikamoto	1250	YES	1320	NO	Adopted	Nakashima	Nakashima-gun	Owari-kuni	Chikazane	0	Motozane	1	Noritsugu	1				
2	Oda	Motozane	1270	YES	1340	NO	Blood	Nakashima	Nakashima-gun	Owari-kuni	Chikamoto	1	Sanemasa	3	Tsunemasa	3			Masayuki	3
2	Oda	Noritsugu	1270	YES	1340	NO	Blood	Echizen	Nyuu-gun	Echizen-kun	Chikamoto	1	Hiromitsu	3						
3	Oda	Sanemasa	1290	YES	1360	NO	Blood	Nakashima	Nakashima-gun	Owari-kuni	Motozane	2	Takamasa	4	Sanetou	4				
3	Oda	Tsunemasa	1296	YES	1364	YES	Blood	Nakashima	Nakashima-gun	Owari-kuni	Motozane	2	Tsunekatsu	4	Nobumasa	4				
3	Oda	Masayuki	1297	NO	1370	NO	Blood	Nakashima	Nakashima-gun	Owari-kuni	Motozane	2	Tomonaga	4	Shigeharu	4				
3	Oda	Hiromitsu	1290	YES	1360	NO	Blood	Echizen	Nyuu-gun	Echizen-kun	Noritsugu	2	Yasunori	4						

Table 2: Relational Representation of the Dataset- table can go upto 9 son's details

VI. SHORTCOMINGS:

As the dataset is very well organised. It follows consistent symbols throughout the document. Though there are several problems with the data.

- There is no context description about the dataset. From the data one can judge that this is some hierarchy of some ruler/clan. But the region/continent/real place of the ruler is unknown. It is difficult to extract the father of the ruler, in case ruler has a sibling, and his kingdom details are listed before mentioning of the ruler.
- Data is organised in an excel sheet, but it is not in a relation format nor in a XML format. Such kind of data is difficult to decode or put to use through regular scripts and validating tools.

- Data is not quantified as father/son relationship. The analyser has to guess it from its organizational structure.
- Data is unknown for some death dates and birthdates.
- Structure of the data may be good for visualization, but it is not a standard way to denote such kind of data. XML would be more suited for it. If it can be normalized and properly referenced, it can be organised in relational format as well.
- Data is hard to automate. Manual intervention would be required, which makes data more prone to errors.

VII. METADATA:

Context	Hierarchical Details of Shiba Clan of Japan existed from 1240 to 1720
Storage	Microsoft Excel Sheet
No. Of Rows	275
No. Of Columns	21
Author	Unknown
Analysis By	Shrashti Singhal
Conceptual Schema	ER Diagram- Crows Foot Format
URL	https://github.com/ShrashtiSinghal/Data-Curation/blob/master/Assignment%205-%20Data%20Set%20Analysis/2%20Input%20dataset.xlsx
Origin of Dataset	http://www.geocities.jp/kawabemasatake/index.html

VIII. Questions

1. Speculate as to the content of the data set. What is it describing?

This dataset is a history table of Japanese Clan called Shiba. This clan family name is Oda. The clan existed from 1240-1720 AD. The document is organised in a family tree structure to denote parent and child relationship between various rulers.

2. Say at least two things about how the data is organized. (E.g., Is it in first normal form? Is there anything significant regarding how it is visually organized? Is it significant that the first two (or three) rows appear different than the subsequent rows? Is the text formatting significant? Etc.)

- The data is not in 1FN. The family tree has more than 1 branch each at a time. Therefore the domain son is not atomic. If organised like son1, son2 etc. the data can be converted to 1FN.
- The data is visually organised in a hierarchical tree structure to ensure it is understandable as a family tree.
- First 4 lines are arranged in 1st column. First 2 lines describes the Clan name and area of origin of the clan. Next 2 lines describes the progenitor of the clan and its 1st ruler. For remaining children record, it is organised in a column wise-hierarchical order, one row below the parent.
- When the clan expand to a different district in the same or different area, the name of the ruler is mentioned in bold.

3. How many different entities and/or attributes are apparent in the data? (You can illustrate this with a diagram if you wish.)

List of Entities and attributes: [CLICK HERE](#)

Entities shown in a schema: [CLICK HERE](#)

4. Describe the dataset's shortcomings with respect to data curation objectives.

Shortcoming of the dataset from curation point of view: [CLICK HERE](#)

5. Given that one of the sources of the data is the following webpage

<http://www.geocities.jp/kawabemasatake/index.html>, name at least one assumption that was made when the data file EC_A was created. (There are at least three assumptions that are relatively easy to pick out.) Why are the assumptions significant? What do they say about the quality of the data?

It looks like the dataset is entirely created from the webpage. The webpage and its subpages describes complete account of the ranks, life and other details of each ruler of the clan. In fact is also tells how this clan was linked to other clans. It mentions the relationships of the members of this clan with other clans. The webpage gives character details of most of the high rank rulers. It provides brief detail of how they formed their military.

This dataset looks correct and reliable to most extent, as it taken from the place where details more than just family tree is given, except for the birth dates and death dates, which are unknown for some of the rulers.